Variational Inference with Coverage Guarantees in Simulation-Based Inference

Yash Patel [1] **Declan McNamara** [1] **Jackson Loper** [1] **Jeffrey Regier** [*1] **Ambuj Tewari** [*1]

Abstract

Amortized variational inference is an often employed framework in simulation-based inference that produces a posterior approximation that can be rapidly computed given any new observation. Unfortunately, there are few guarantees about the quality of these approximate posteriors. We propose Conformalized Amortized Neural Variational Inference (CANVI), a procedure that is scalable, easily implemented, and provides guaranteed marginal coverage. Given a collection of candidate amortized posterior approximators, CANVI constructs conformalized predictors based on each candidate, compares the predictors using a metric known as predictive efficiency, and returns the most efficient predictor. CANVI ensures that the resulting predictor constructs regions that contain the truth with a user-specified level of probability. CANVI is agnostic to design decisions in formulating the candidate approximators and only requires access to samples from the forward model, permitting its use in likelihood-free settings. We prove lower bounds on the predictive efficiency of the regions produced by CANVI and explore how the quality of a posterior approximation relates to the predictive efficiency of prediction regions based on that approximation. Finally, we demonstrate the accurate calibration and high predictive efficiency of CANVI on a suite of simulation-based inference benchmark tasks and an important scientific task: analyzing galaxy emission spectra.

1. Introduction

In many scientific applications, such as in astrophysics, neuroscience, and particle physics (Papamakarios & Murray, 2016; Lueckmann et al., 2017; Greenberg et al., 2019; Deistler et al., 2022; Papamakarios et al., 2019; Boelts et al., 2022), posterior distributions $\mathcal{P}(\Theta \mid x)$ are sought over a large collection of x, typically on the order of 10,000 or more. In such scientific settings, (θ, x) pairs are assumed to come from nature, which has led to the growth of "simulation-based inference," in which a likelihood $\mathcal{P}(X \mid \theta)$ and prior $\mathcal{P}(\Theta)$ are posited and simulated to fit posteriors (Cranmer et al., 2020; Lueckmann et al., 2021). Even when the likelihood and prior are well specified, exact sampling from the posteriors is intractable, requiring running 10,000 separate MCMC chains.

In these settings, amortized variational inference is frequently employed. Variational inference (VI) has become a staple in Bayesian inference; however, it has been repeatedly noted that a major shortcoming of VI is its lack of any theoretical guarantees and tendency to produce biased posterior estimates (Blei et al., 2017; Murphy, 2022; Zhang et al., 2018; Yao et al., 2018). The most common metric for assessing the calibration of such inference algorithms is the "expected coverage." Having calibrated expected coverage is a necessary but not sufficient condition for conditional coverage, yet amortized variational approximations fail to even achieve this minimal requirement, despite significant work to remedy this shortcoming (Deistler et al., 2022; Delaunoy et al., 2022; Lemos et al., 2023; Delaunoy et al., 2023). A lack of such calibration limits the capacity to reach downstream scientific conclusions, highlighted in a recent meta-study of likelihood-free inference algorithms (Hermans et al., 2021b).

In applications where many posteriors need to be estimated, credible regions with *marginally* calibrated coverage can be sufficient for downstream scientific inquiries. For instance, in astrophysics, there is great interest in constraining the ΛCDM model, the current concordance model in cosmology (Gilman et al., 2021; Hezaveh et al., 2016; Vegetti et al., 2010; Vegetti & Koopmans, 2009; Hogg & Blandford, 1994). A recent work from this community, (Hermans et al., 2021a), leveraged Bayesian inference towards this end, obtaining approximate posteriors for the parameters of interest on each of 10,000 observations. Crucially, these posteriors were then used to produce credible intervals with *marginally* valid frequentist coverage, from which they made claims on the ΛCDM model.

[*]Senior author order was decided by a coin flip. [1]Department of Statistics, University of Michigan, Ann Arbor, USA. Correspondence to: Yash Patel <yppatel@umich.edu>.

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

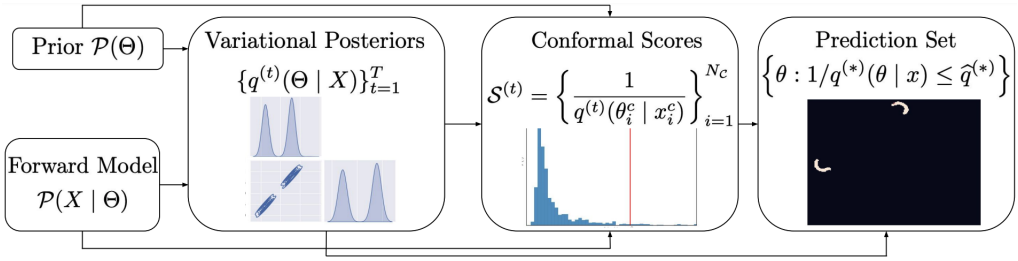


Figure 1. CANVI is a wrapper around variational inference requiring minimal implementation and computational overhead that produces prediction regions with guaranteed marginal calibration. Among a family of candidate amortized posterior approximators, CANVI can identify the approximator leading to the most efficient prediction regions. CANVI can be used in any setting where the forward model $\mathcal{P}(X \mid \Theta)$ can be sampled.

Our insight is that conformal prediction can be leveraged to provide variational approximators with marginal coverage guarantees and provide users new ways to measure the quality of variational approximators. In this manuscript, we present CANVI (Conformalized Amortized Neural Variational Inference), a novel, general framework for producing marginally calibrated, informative prediction regions from a collection of variational approximators. Such regions can be produced with minimal implementation and computational overhead, requiring only samples from the prior and $\mathcal{P}(X \mid \Theta)$, as shown in Figure 1. In Section 3.3, we provide theoretical analysis of the informativeness of the prediction regions produced by CANVI using a measure known as "predictive efficiency." High predictive efficiency is necessary to draw conclusions in downstream scientific inquiries and relates to asymptotic conditional coverage with the appropriate choice of score function. Finally, in Section 4, we show calibration and predictive efficiency across simulation-based inference benchmark tasks and an important scientific task: analyzing galaxy emission spectra.

2. Background

2.1. Variational Inference

Bayesian methods aim to sample the posterior distribution $\mathcal{P}(\Theta \mid X)$, typically using either MCMC or VI. VI has risen in popularity recently due to how well it lends itself to amortization. Given an observation X, variational inference transforms the problem of posterior inference into an optimization problem by seeking

$$\varphi^*(X) = \arg\min_{\varphi} D(q_\varphi(\Theta) || \mathcal{P}(\Theta \mid X)), \tag{1}$$

where D is a divergence and q_φ is a member of a variational family of distributions $\mathcal{Q}$ indexed by the free parameter φ. Normalizing flows have emerged as a particularly apt choice for $\mathcal{Q}$, as they are highly flexible and perform well empirically (Rezende & Mohamed, 2015; Agrawal et al., 2020). Amortized variational inference expands on this approach by training a neural network to approximate $\varphi^*(X)$. This leads to a variational posterior approximator $q(\Theta \mid X) = q_{\varphi^*(X)}(\Theta)$ that can be rapidly computed for any value X. The characteristics of φ^* depend in part on the variational objective, D. For instance, using a reverse-KL objective, i.e. $D_{KL}(q_\varphi(\Theta) || \mathcal{P}(\Theta \mid X))$, is known to produce mode-seeking posterior approximations, whereas using a forward-KL objective, i.e. $D_{KL}(\mathcal{P}(\Theta \mid X) || q_\varphi(\Theta))$, encourages mode-covering behavior (Murphy, 2023). Popular variational objectives include the Forward-Amortized Variational Inference (FAVI) objective (Ambrogioni et al., 2019; Bornschein & Bengio, 2014), the Evidence Lower Bound (ELBO), and the Importance Weighted ELBO (IWBO) (Burda et al., 2015).

2.2. Conformal Prediction

Given $\mathcal{D}_\mathcal{C} = \{(x_1, \theta_1), \ldots (x_N, \theta_N)\} \overset{\text{iid}}{\sim} \mathcal{P}(X, \Theta)$, conformal prediction (Angelopoulos & Bates, 2021; Shafer & Vovk, 2008) produces prediction regions with distribution-free guarantees. A prediction region is a mapping from observations of X to sets of possible values for Θ and is said to be marginally calibrated at the $1 - \alpha$ level if $\mathcal{P}(\Theta \notin \mathcal{C}(X)) \leq \alpha$.

Split conformal is one popular version of conformal prediction. In this approach, marginally calibrated regions $\mathcal{C}$ are designed using a "score function" $s(x, \theta)$. Intuitively, the score function should have the quality that $s(x, \theta)$ is smaller when it is more reasonable to guess that $\Theta = \theta$ given the observation $X = x$. For example, if one has access to a function $\hat{f}(x)$ which attempts to predict Θ from X, one might take $s(x, \theta) = \|\hat{f}(x) - \theta\|$. The score function is evaluated on each point of a subset of the dataset $\mathcal{D}_\mathcal{C}$, called the "calibration dataset," yielding $\mathcal{S} = \{s(x_i^c, \theta_i^c)\}_{i=1}^{N_c}$. Note that the calibration dataset cannot be used to pick the score function; if data is used to design the score function, it must independent of $\mathcal{D}_\mathcal{C}$. This is how "split conformal" gets its name: in typical cases, data are split into two parts, one used to design s and the other to perform calibration. We then de-

fine $\widehat{q}(\alpha)$ as the $\lceil (N_{\mathcal{C}}+1)(1-\alpha) \rceil / N_{\mathcal{C}}$ quantile of $\mathcal{S}$. For any future x, the set $\mathcal{C}(x) = \{\theta \mid s(x,\theta) \leq \widehat{q}(\alpha)\}$ satisfies $1-\alpha \leq \mathcal{P}(\Theta \in \mathcal{C}(X))$. This coverage guarantee arises from the exchangeability of the score of a future test point $s(x', \theta')$ with $\mathcal{S}$ and holds regardless $N_{\mathcal{C}}$: conformal guarantees are not asymptotic. If $s(X, \Theta)$ is jointly continuous, we additionally have that $\mathcal{P}(\Theta \in \mathcal{C}(X)) \leq 1-\alpha+1/(N_{\mathcal{C}}+1)$.

While the coverage guarantee holds for any score function, different choices may lead to more or less informative prediction regions (Shafer & Vovk, 2008). For example, the score $s(x,\theta) = 1$ leads to the uninformative prediction region of all possible values of Θ. Predictive efficiency is one way to quantify informativeness (Yang & Kuchibhotla, 2021; Sesia & Candès, 2020), defined as the expected inverse Lebesgue measure of $\mathcal{C}(X)$, $\mathbb{E}_X[\mathcal{L}(\mathcal{C}(X))^{-1}]$. Methods using conformal prediction often seek to identify prediction regions that are efficient and marginally calibrated.

2.3. Related Literature

Efforts to correct for miscalibration have been of great interest recently in light of its apparent omnipresence highlighted in (Hermans et al., 2021b). The notion of calibration studied therein, and the one we concentrate on here, centers on the expected coverage probability of the highest predictive density (HPD) of a posterior estimate $q(\Theta \mid X)$, defined for such a q, some fixed x, and pre-specified coverage level α to be the set $\mathrm{HPD}(q(\Theta \mid x), 1-\alpha)$ with smallest Lebesgue measure such that $\mathcal{P}_{\Theta \sim q(\Theta \mid x)}(\Theta \in \mathrm{HPD}(q(\Theta \mid x), 1-\alpha)) \geq 1-\alpha$. The expected coverage is then $\mathbb{E}_{X,\Theta \sim \mathcal{P}(X,\Theta)}[\mathbb{1}[\Theta \in \mathrm{HPD}(q(\Theta \mid X), 1-\alpha)]]$.

Such calibration has been studied across a number of posterior estimation techniques in the likelihood-free inference community, which generally fall into one of three categories. Neural posterior approximation (NPE) methods directly approximate the posterior density $\mathcal{P}(\Theta \mid X)$ (Papamakarios & Murray, 2016; Lueckmann et al., 2017; Greenberg et al., 2019; Deistler et al., 2022), neural likelihood estimation (NLE) methods estimate the assumed intractable likelihood $\mathcal{P}(X \mid \Theta)$ (Papamakarios et al., 2019; Boelts et al., 2022), and neural ratio estimation (NRE) methods estimate the $\mathcal{P}(X \mid \Theta)/\mathcal{P}(X)$ ratio (Hermans et al., 2020; Durkan et al., 2020b; Miller et al., 2022). Both NLE and NRE rely on MCMC for posterior sampling after estimation.

While these approaches differ in their estimation strategies, they all suffer miscalibration if naively employed. One suggestion advocated by (Hermans et al., 2021b) was ensembling, compatible with all the aforementioned posterior approximation strategies. Despite its improvement in empirical calibration, ensembling affords no guarantees on calibration and also dramatically increases the computational cost of training and inference alike. As an effort to address this lack of guarantees and computational cost, recent calibration efforts have focused on modifying the loss function to appropriately encourage conservatism in the learned posterior estimates, since the downstream scientific use cases can afford such conservatism, unlike underdispersed posteriors. In particular, (Delaunoy et al., 2022) took a first step in this direction by proposing a modification over the vanilla NRE formulation with a regularization term that results in more conservative posterior estimates. Such a method, however, has no guarantees and further relies on the selection of a tunable λ parameter, whose optimal selection is unknown without awareness of the true posterior. This approach was then extended to NPE and NLE in (Delaunoy et al., 2023) and (Falkiewicz et al., 2023), which both again suffer from the deficiencies of producing overly conservative posteriors and relying on the careful selection of λ.

3. Method

In response to the shortcomings highlighted in the previous section, we were interested in procuring a method that has (1) guarantees on calibration without tuning parameters, (2) minimal computational overhead, and (3) informative prediction regions. We demonstrate in Section 3.1 that leveraging conformal prediction on an amortized VI approximator $q(\Theta \mid X)$ immediately addresses points (1) and (2). We then study in the following sections how this naive application can be extended to the full CANVI algorithm by considering a collection of amortized VI approximators $\{q^{(1)}(\Theta \mid X), \ldots, q^{(T)}(\Theta \mid X)\}$ to address point (3). CANVI can be applied whenever $\mathcal{P}(X, \Theta)$ can be sampled. The coverage validity of CANVI is proven in Section 3.2, and analyses of its predictive efficiency in Section 3.5.

3.1. CANVI: Score Function

In the simplest case, CANVI takes as input a single amortized posterior approximator $q(\Theta \mid X)$. In traditional applications of split conformal, much concern is given to the loss of accuracy of the predictor q in having to reserve a subset of the training data for calibration. Here we have no such issues; we sample $\mathcal{D}_{\mathcal{C}} = \{(x_i^c, \theta_i^c)\}_{i=1}^{N_{\mathcal{C}}} \overset{\text{iid}}{\sim} \mathcal{P}(\Theta)\mathcal{P}(X \mid \Theta)$ from the joint distribution to produce a calibration dataset that can be arbitrarily large. Given $q(\Theta \mid X)$, we employ the following score, as used in (Angelopoulos & Bates, 2021):

$$s(x_i, \theta_i) = (q(\theta_i \mid x_i))^{-1}. \tag{2}$$

Denoting the $\lceil (N_{\mathcal{C}}+1)(1-\alpha) \rceil / N_{\mathcal{C}}$ quantile of the score distribution over $\mathcal{D}_C$ as $\widehat{q}_{\mathcal{C}}(\alpha)$, $\mathcal{C}(x) = \{\theta : 1/q(\theta \mid x) \leq \widehat{q}_{\mathcal{C}}(\alpha)\}$ is then marginally calibrated. It may be disjoint if the posterior is multimodal.

While other choices of score functions also result in regions $\mathcal{C}(x)$ that address both the lack of guarantees and computational cost of methods highlighted in Section 2.3, the particular choice of Equation (2) has the desirable

property that, if we recover the true posterior, that is $q(\Theta \mid X) = \mathcal{P}(\Theta \mid X)$, we recover the HPDs, namely $\mathcal{C}(x) = \text{HPD}(\mathcal{P}(\Theta \mid x), 1-\alpha)$, achieving conditional coverage. Note that the procedure described in this section and those that follow can be easily extended to the group conditional setting if such coverage is of interest, whose discussion we defer to Appendix A.

3.2. CANVI: Approximator Selection

To mitigate the risk of producing uninformative prediction regions, it is natural to explore multiple posterior approximations, $\{q^{(t)}(\Theta \mid X)\}_{t=1}^T$, since a poorly chosen approximator may lead to poor predictive efficiency. These posterior approximations could, for instance, differ in their choice of training objective, variational family, or hyperparameters. CANVI seeks to identify the variational approximator $q^{(t^*)}$ for which the efficiency is greatest, defined for a threshold τ to be

$$\ell(q, \tau) := \mathbb{E}_X \left[\mathcal{L}\left(\{\theta : 1/q(\theta \mid X) \leq \tau\} \right) \right], \tag{3}$$

We defer the discussion of estimating $\ell(q, \tau)$ to Section 3.4. Naively, one would expect by taking $(q^{(t^*)}, \widehat{q}_{\mathcal{C}}^{(t^*)}(\alpha))$ where $t^* := \arg\min_t \ell(q^{(t)}, \widehat{q}_{\mathcal{C}}^{(t)}(\alpha))$, we achieve maximal efficiency and retain coverage guarantees. However, defining $\mathcal{C}(x)$ with $\widehat{q}_{\mathcal{C}}^{(t^*)}(\alpha)$ fails to retain coverage guarantees, as the exchangeability of scores of future test points $s(x', \theta')$ with $\mathcal{S}$ is lost in conditioning on $\mathcal{D}_{\mathcal{C}}$ for selecting t^*.

We must, therefore, perform an *additional* recalibration step after selecting t^* to retain coverage guarantees. CANVI performs such recalibration using an additional dataset $\mathcal{D}_{\mathcal{R}}$ again constructed with i.i.d. draws from $\mathcal{P}(X, \Theta)$. We take $\mathcal{D}_{\mathcal{R}}$ to be the same size as $\mathcal{D}_{\mathcal{C}}$, i.e. $|\mathcal{D}_{\mathcal{R}}| = N_{\mathcal{C}}$. CANVI then computes the quantile $\widehat{q}_{\mathcal{R}}^{(*)}(\alpha) := \widehat{q}_{\mathcal{R}}^{(t^*)}(\alpha)$, which is used to define prediction regions $\mathcal{C}(x) := \{\theta : 1/q^{(t^*)}(\theta \mid x) \leq \widehat{q}_{\mathcal{R}}^{(*)}(\alpha)\}$. However, such regions require analysis to guarantee high efficiency, as we explore in Section 3.3.

The full CANVI framework is provided in Algorithm 1. The validity of the CANVI procedure follows directly from that of split conformal prediction, formally stated below and explicitly proven in Appendix B.

Lemma 3.1. *Let $\alpha \in (0, 1)$ and*

$$q^{(*)}(\Theta \mid X), \widehat{q}_{\mathcal{R}}^{(*)}(\alpha) = \text{CANVI}\left(\{q^{(t)}(\Theta \mid X)\}_{t=1}^T, \mathcal{P}(X, \Theta), 1-\alpha, N_{\mathcal{C}}, N_{\mathcal{T}} \right)$$

Let $(x', \theta') \sim \mathcal{P}(X, \Theta) \perp\!\!\!\perp \mathcal{D} \cup \mathcal{D}_{\mathcal{C}} \cup \mathcal{D}_{\mathcal{R}} \cup \mathcal{D}_{\mathcal{T}}$, with $\mathcal{D}$ being the data used to train $\{q^{(t)}(\Theta|X)\}_{t=1}^T$. Then $1-\alpha \leq \mathcal{P}(1/q^{(*)}(\theta'|x') \leq \widehat{q}_{\mathcal{R}}^{(*)}(\alpha))$.

Algorithm 1 CANVI: Note that VOLUMEEST is a volume estimator subroutine detailed in Section 3.4.

Procedure CPQuantile
Inputs: Posterior approximation $q(\Theta \mid X)$, Calibration set $\mathcal{D}_{\mathcal{C}}$, Desired coverage $1-\alpha$
$\mathcal{S} \leftarrow \{\frac{1}{q(\theta_i|x_i)}\}_{i=1}^{N_{\mathcal{C}}}$
Return $\frac{\lceil (N_{\mathcal{C}}+1)(1-\alpha) \rceil}{N_{\mathcal{C}}}$ quantile of $\mathcal{S}$

Procedure CANVI
Inputs: Posterior approximators $\{q^{(t)}(\Theta|X)\}_{t=1}^T$, Prior $\mathcal{P}(\Theta)$, Forward model $\mathcal{P}(X \mid \Theta)$, Desired coverage $1-\alpha$, Calibration size $N_{\mathcal{C}}$, Test size $N_{\mathcal{T}}$
$\mathcal{D}_{\mathcal{C}}, \mathcal{D}_{\mathcal{R}} \sim \mathcal{P}(X, \Theta), \mathcal{D}_{\mathcal{T}} \sim \mathcal{P}(X)$
for $t \in \{1, \ldots T\}$ **do**
  $\widehat{q}_{\mathcal{C}}^{(t)} \leftarrow$ CPQUANTILE$(q^{(t)}, \mathcal{D}_{\mathcal{C}}, 1-\alpha)$
  $\widehat{\ell}^{(t)} \leftarrow$ VOLUMEEST$(q^{(t)}, \mathcal{P}(\Theta), \widehat{q}_{\mathcal{C}}^{(t)}, \mathcal{D}_{\mathcal{T}})$
end for
$t^* \leftarrow \arg\min_t \widehat{\ell}^{(t)}$
$\widehat{q}_{\mathcal{R}}^{(*)}(\alpha) \leftarrow$ CPQUANTILE$(q^{(t^*)}, \mathcal{D}_{\mathcal{R}}, 1-\alpha)$
Return $q^{(*)}(\Theta \mid X), \widehat{q}_{\mathcal{R}}^{(*)}(\alpha)$

3.3. CANVI: Efficiency Analysis Assumptions

We now show that, with high probability, the pair CANVI produces $(q^{(*)}(\Theta|X), \widehat{q}_{\mathcal{R}}^{(*)}(\alpha))$ is the most efficient amongst the candidate posteriors considered. The concern is that the post-recalibrated quantile may result in significant degradation of the efficiency, i.e. $\ell(q^{(*)}, \widehat{q}_{\mathcal{R}}^{(*)}(\alpha)) \gg \ell(q^{(*)}, \widehat{q}_{\mathcal{C}}^{(*)}(\alpha))$. This tradeoff between coverage and efficiency was studied in (Yang & Kuchibhotla, 2021).

Recall $1-\alpha \leq \mathcal{P}(\Theta \in \mathcal{C}(X)) \leq 1-\alpha+1/(N_{\mathcal{C}}+1)$. Denote the CDF of the score function under the joint distribution $\mathcal{P}(X, \Theta)$ as $\mathcal{F}(s) := \mathcal{P}_{\Theta,X}(1/q(\Theta \mid X))$. The coverage guarantee, thus, implies $\widehat{q}(\alpha) \in [\mathcal{F}^{-1}(1-\alpha), \mathcal{F}^{-1}(1-\alpha+1/(N_{\mathcal{C}}+1))]$ for $\widehat{q}(\alpha)$ from *any* calibration set. In particular, $\widehat{q}_{\mathcal{C}}^{(*)}(\alpha)$ and $\widehat{q}_{\mathcal{R}}^{(*)}(\alpha)$ both lie in this range.

We bound the efficiency suboptimality by proceeding in two steps. We first demonstrate that the quantiles of $q^{(*)}(\Theta \mid X)$ under $\mathcal{D}_{\mathcal{C}}$ and $\mathcal{D}_{\mathcal{R}}$ are close by demonstrating the quantile range of $\mathcal{F}^{-1}$ is small. We then demonstrate the efficiency varies smoothly as a function of the quantile, allowing us to bound the resulting efficiency change. Formally, we state these assumptions respectively as follows, per (Yang & Kuchibhotla, 2021), which we then demonstrate follow from properties of the chosen variational families.

Assumption 3.2. For each t, the $\ell(q^{(t)}, \tau)$ is Lipschitz continuous in τ with constant L_W.

Assumption 3.3. For each t, $\exists\ r, \gamma \in (0, 1]$ such that $\mathcal{F}_t^{-1}(s)$ (inverse score CDF under $q^{(t)}$), is γ-Hölder continuous on $[1-\alpha, 1-\alpha+r]$ *with continuity constant* L_t.

3.3.1. Lipschitz Continuity of Efficiency

We now demonstrate Assumption 3.2 can be guaranteed with the appropriate selection of variational family by the end user. It suffices to demonstrate $\ell_x(q,\tau) := \mathcal{L}(\{\theta : 1/q(\theta \mid x) \leq \tau\})$ is L-Lipschitz continuous in τ for any $x \in \mathcal{X}$, proven in Appendix D.1.

For any fixed x, $\ell_x(q,\tau)$ is the intrinsic volume of the sublevel set of $1/q(\theta \mid x)$. We summarize and subsequently use relevant results from (Jubin, 2019) below. "Intrinsic volume" defines the notion of volume for a lower dimensional manifold embedded in a higher dimensional space. For a brief review of Riemannian manifolds, see Appendix C; a more complete presentation is available in (Lee & Lee, 2012). The $n-k$ degree intrinsic volume of a flat compact n-dimensional manifold N is

$$\mathcal{L}_{n-k}(N) = b_k \int_{\partial N} \operatorname{tr}\left(\bigwedge^{k-1} S\right) \operatorname{vol}_{\partial N}, \tag{4}$$

where $0 \leq k \leq n$, $b_k \in \mathbb{R}$, and S is the second fundamental form of ∂N in N. Lipschitz continuity of $\mathcal{L}_{n-k}(M_f^\tau)$ was established as follows. The level sets $M_f^\tau := f^{-1}((-\infty,\tau])$ are restricted to "regular values" of f, namely τ such that $f(x) = \tau \implies df(x) \neq 0$.

Theorem 3.4. *Let (M,g) be an n-dimensional Riemannian manifold, $f \in \mathcal{C}^3(M,\mathbb{R})$ bounded below, and τ a regular value of f equipped with the standard uniform $\mathcal{C}^3$ topology. Then, if $0 \leq k \leq n$, $\tau \to \mathcal{L}_{n-k}(M_f^\tau)$ is Lipschitz continuous (Jubin, 2019).*

The Lipschitz continuity of $\ell_x(q,\tau)$ then follows as a corollary for any variational families for which the density is sufficiently smooth, namely $q(\theta \mid x) \in \mathcal{C}^3(\mathbb{R}^n)$. The proof follows by taking $\ell_x(q,\tau) := \mathcal{L}_n((\mathbb{R}^n)_s^\tau)$ for $s(\theta) = 1/q(\theta|x)$, with $\tau = s(\theta)$ for $\{\theta : \nabla_\theta q(\theta|x) \neq 0\}$ being the set of regular values. This domain restriction is discussed more in Section 3.5. Notice $s(\theta) \in \mathcal{C}^3(\mathbb{R}^n)$ as both $f(x) := 1/x$ and $q(\theta \mid x)$ are $\in \mathcal{C}^3(\mathbb{R}^n)$ and $\mathcal{C}^3$ is closed under function composition. Formally,

Corollary 3.5. *Suppose for any $x \in \mathcal{X}$, $q(\theta \mid x) \in \mathcal{C}^3(\mathbb{R}^n)$ is bounded above and τ is a regular value of $q(\theta \mid x)$. Then, $\ell(q,\tau)$ is Lipschitz continuous in τ.*

Notably, this assumption on smoothness holds for most variational families used in practice, including highly expressive flow-based variational families (Köhler et al., 2021).

3.3.2. Continuity of Conformal Quantiles

We now discuss the validity of Assumption 3.3. Comparable assumptions are commonly used in the quantile estimation literature, as discussed in (Lei et al., 2018). The Hölder constant cannot be characterized in general, as it is intimately tied to specific details of the score distribution under $\mathcal{P}(X,\Theta)$. We, therefore, provide an explicit characterization for a particular family of distributions in Theorem 3.6 and defer extensions to a broader set of families to future work. Details for this proof are given in Appendix E.

Theorem 3.6. *Let Θ and X be zero-mean unit-variance Gaussian random variables with correlation ρ. Let $q^{(t)}(\theta|x) = \mathcal{N}(\theta; tx, 1-\rho^2)$. Let $\kappa := t^2 - 2t\rho + 1$ and $r > 0$. Then $F_t^{-1}(z)$, is 1-Hölder continuous on $[1-\alpha, 1-\alpha+r]$ with Hölder constant*

$$\frac{\kappa\Phi^{-1}(\frac{1-\alpha}{2})\sqrt{\exp\left(\frac{\kappa}{1-\rho^2}\Phi^{-1}(\frac{1-\alpha}{2})^2 - \frac{(1-\alpha)^2}{2}\right)}}{\sqrt{(1-\rho^2)/2}} \tag{5}$$

Notably, the Hölder constant is minimized in recovering the true posterior, as Equation (5) is minimized at $\varphi = \rho$.

3.4. CANVI: Volume Estimation

We now provide an estimation procedure for $\ell(q,\tau)$. Naively, we might expect taking the sample average of $\ell_{x_i}(q,\tau)$ over $\mathcal{D}_\mathcal{T} := \{x_i\}_{i=1}^{N_\mathcal{T}} \sim \mathcal{P}(X)$ would suffice. However, exact calculation of $\ell_{x_i}(q,\tau)$ requires a grid-discretization over $\operatorname{Supp}(\Theta|x_i)$, which is only feasible when the support has a known, small extent.

As a result, $\ell_x(q,\tau)$ is estimated using an importance-weighted Monte Carlo estimate over S samples from q. Such an estimator, however, suffers from high variance if q is underdispersed, as $\mathcal{C}(x)$ will cover regions of low variational density. To combat this issue, we use the well-known fact that $\mathcal{P}(\Theta \mid X)$ is narrower than $\mathcal{P}(\Theta)$ to construct a "mixed sampler," specifically with $z_j \sim \operatorname{Bern}(\lambda)$ and $\theta_j \sim q(\Theta|x_i)^{z_j}\mathcal{P}(\Theta)^{1-z_j}$, where the mixed density is now $\widetilde{q}(\theta_j) = \lambda q(\theta_j|x_i) + (1-\lambda)\mathcal{P}(\theta_j)$. The necessity of such mixing is dependent on the nature of the variational posterior with respect to the true posterior, which is unknown in practice. We, thus, average several estimates over $\{\lambda_k\} \in [0,1]$. Denoting the mixed density with λ_k as $\widetilde{q}_k$, for each x_i and λ_k, we make S draws $\{\theta_{jk}\}_{j=1}^S \sim \widetilde{q}_k(\Theta \mid x_i)$. We empirically demonstrate the necessity of such mixed sampling in Section 4.2.2. This procedure is summarized in Algorithm 2, with the final estimate $\widehat{\ell}(q,\tau)$ being

$$\frac{1}{K\mathcal{N}_\mathcal{T}} \sum_{i,k=1}^{\mathcal{N}_\mathcal{T},K} \frac{1}{S} \sum_{j=1}^{S} \frac{1}{\widetilde{q}_k(\theta_{jk}|x_i)} \mathbb{1}\left[\frac{1}{q(\theta_{jk}|x_i)} \leq \tau\right] \tag{6}$$

3.5. CANVI: Efficiency Proof

We now state the result of recovery of the optimal recalibrated approximator. To do so, we require the Monte Carlo estimate to be sufficiently well-behaved to recover the optimal pre-recalibration approximator.

Algorithm 2 VOLUMEEST

Procedure VolumeEst
Inputs: Posterior approximation $q(\Theta \mid X)$, Prior $\mathcal{P}(\Theta)$, CP quantile $\widehat{q}$, Test set $\mathcal{D}_{\mathcal{T}}$
for $i \in \{1, \ldots \mathcal{N}_{\mathcal{T}}\}, k \in \{1, \ldots K\}$ **do**
  for $j \in \{1, \ldots S\}$ **do**
    $z_j \sim \text{Bern}(k/K)$
    $\theta_j \sim q(\Theta \mid x_i)^{z_j} \mathcal{P}(\Theta)^{1-z_j}$
    $\widetilde{q}_j \leftarrow (k/K) q(\theta_j \mid x_i) + (1 - k/K)\mathcal{P}(\theta_j)$
  end for
  $V_{i,k} \leftarrow \frac{1}{S} \sum_{j=1}^{S} \frac{1}{\widetilde{q}_j} \mathbb{1}[1/q(\theta_j|x_i) \leq \widehat{q}]$
end for
Return $\frac{1}{K\mathcal{N}_{\mathcal{T}}} \sum_{i,k} V_{i,k}$

Assumption 3.7. If $t^* := \arg\min_{1 \leq t \leq T} \ell(q^{(t)}, \widehat{q}_{\mathcal{R}}^{(t)}(\alpha))$ and $\widehat{t}^* := \arg\min_{1 \leq t \leq T} \widehat{\ell}(q^{(t)}, \widehat{q}_{\mathcal{R}}^{(t)}(\alpha))$ for $\alpha \in (0, 1)$, then $\exists \Delta, \epsilon > 0$, such that with probability at least $1 - \epsilon$, $|\ell(q^{(\widehat{t}^*)}, \widehat{q}_{\mathcal{R}}^{(\widehat{t}^*)}(\alpha)) - \ell(q^{(t^*)}, \widehat{q}_{\mathcal{R}}^{(t^*)}(\alpha))| < \Delta$.

Important to note is that $\widehat{\ell}$ is only used to select t^*, after which we make claims on ℓ (i.e. *not* the estimate) for the recalibrated quantiles in Theorem 3.8. As with Assumption 3.3, this assumption is intimately tied to specific details of the score distribution under $\mathcal{P}(X, \Theta)$, making its restatement in more natural distributional properties of q impossible. We demonstrate its validity empirically across several posteriors in Section 4.2.

The proof of Theorem 3.8 now follows as an extension of Theorem 3 from (Yang & Kuchibhotla, 2021) and is explicitly provided in Appendix D.2. Notably, we use Corollary 3.5 to replace Assumption 3.2 with a more natural set of conditions for this context. This requires ensuring that, for any x, $\widehat{q}_{\mathcal{C}}^{(*)}(\alpha)$ and $\widehat{q}_{\mathcal{R}}^{(*)}(\alpha)$ are regular values of $s(\theta)$. We, thus, assume for any x and $\theta \neq 0$, $\mathcal{L}(\{\theta : \nabla_\theta q(\theta|x)\}) = 0$, which naturally holds for variational families used in practice, i.e. any non-piecewise constant density estimator.

Theorem 3.8. *Suppose for any $x \in \mathcal{X}$ and $t = 1, ..., T$, $q^{(t)}(\theta \mid x) \in \mathcal{C}^3(\mathbb{R}^n)$ is bounded above and for $\theta \neq 0$, $\mathcal{L}(\{\theta : \nabla_\theta q^{(t)}(\theta|x)\}) = 0$. Further assume $P(X, \Theta)$ is bounded above. Let $\alpha \in (0, 1)$ and*

$$q^{(*)}(\Theta \mid X), \widehat{q}_{\mathcal{R}}^{(*)}(\alpha) = \text{CANVI}\left(\{q^{(t)}(\Theta \mid X)\}_{t=1}^{T}, \mathcal{P}(X, \Theta), 1 - \alpha, N_{\mathcal{C}}, N_{\mathcal{T}}\right)$$

If, for $r \geq \max\{\sqrt{\log(4T/\delta)/2N_{\mathcal{C}}}, 2/N_{\mathcal{C}}\}$ and $\delta \in [0, 1]$, Assumption 3.3 holds and for $\Delta, \epsilon > 0$ Assumption 3.7 holds, then with probability at least $(1 - \epsilon)(1 - \delta)$,

$$\ell(q^{(*)}, \widehat{q}_{\mathcal{R}}^{(*)}(\alpha)) \leq \min_{1 \leq t \leq T} \ell(q^{(t)}, \widehat{q}_{\mathcal{R}}^{(t)}(\alpha)) + \Delta + 3L_W L_{[T]} \left[\left(\frac{\log(4T/\delta)}{N_{\mathcal{C}}} \right)^{\gamma/2} + \left(\frac{2}{N_{\mathcal{C}}} \right)^{\gamma} \right], \quad (7)$$

where γ, L_W, and $L_{[T]} = \max_{1 \leq t \leq T} L_t$ are constants defined in Assumptions 3.3 and 3.2.

Again, in any setting where it is possible to sample from $\mathcal{P}(X, \Theta)$, $N_{\mathcal{C}}$ can be made arbitrarily large, tightening the bound in Equation 7. Practitioners can, thus, focus on obtaining efficient predictors knowing that CANVI will make the optimal selection with high probability.

4. Experiments

As discussed, the main advantages of CANVI over alternative strategies are its guaranteed calibration without the need for tuning parameters, minimal computational overhead, and informative prediction regions. For this reason, we present three experiments herein. The first (Section 4.1) seeks to validate the first two claims by comparing BNRE, BNRE C, NPE, NRE, NRE C, and Ratio BNPE to the vanilla version of CANVI i.e. when it is applied to a single $q(\Theta \mid X)$, where no recalibration is necessary.

Notably, the informativeness of prediction regions, the focus of the third claim, is only of interest once a predictor is calibrated, meaning the comparison with alternatives is only meaningful when they are nearly calibrated. As demonstrated in Section 4.1, however, the alternatives fail to consistently demonstrate calibration, rendering such a comparison moot. For this reason, the following experiment (Section 4.2) focuses on demonstrating that applying the full CANVI procedure to a collection $\{q^{(t)}(\Theta \mid X)\}_{t=1}^{T}$ both retains coverage under recalibration and ultimately recovers the most efficient predictor. The latter hinges upon the validity of Assumption 3.7 for the Monte Carlo efficiency estimator in Equation (6). We, therefore, demonstrate in the subsequent experiments that this estimator exhibits the desired estimation consistency when applied in settings where posterior estimators are trained to different epochs (Section 4.2.1) or with different training objectives (Section 4.2.2). Finally, we demonstrate CANVI can computationally scale up to scientific problems of interest in Section 4.3.

In all experiments, coverage for variational posteriors is assessed using Monte Carlo estimation, namely by constructing the highest density credible region per x_i. That is, for a given x_i, the ζ such that $\{\theta_j \mid q(\theta_j \mid x_i) \geq \zeta\}$ captures $1 - \alpha$ of the probability mass is estimated by drawing $\{\theta_j\}_{j=1}^{N} \sim q(\Theta|x_i)$ and finding the $1 - \alpha$ quantile of $\{q(\theta_j|x_i)\}_{j=1}^{N}$. Coverage of the true parameter θ can be assessed by checking if $q(\theta \mid x_i) \geq \zeta$. Details are provided in Appendix G, and code is available at `https://github.com/yashpatel5400/canvi.git`.

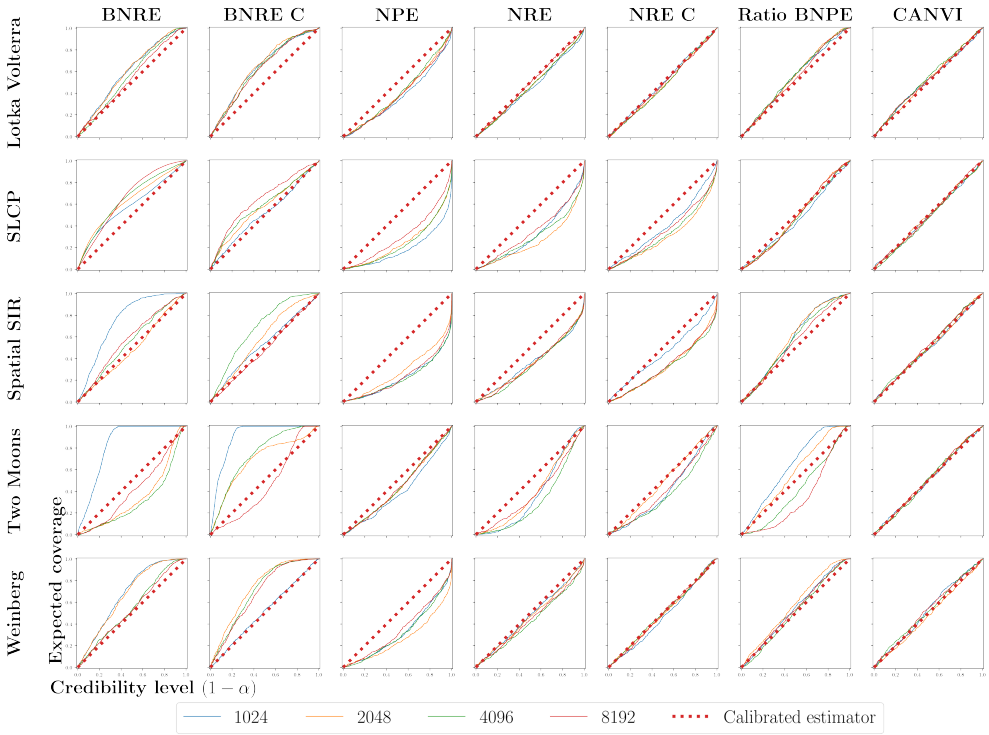


Figure 2. Calibration on the SBI benchmarks across different calibration strategies. Perfect calibration corresponds to the highlighted $y = x$ curve. Conservative prediction regions lie above this calibrated line and overconfident ones below. Conformalized lines (CANVI) are difficult to distinguish, as they all lie along the desired $y = x$ curve.

4.1. Coverage Calibration

We evaluate on the standard SBI benchmark tasks, highlighted in (Delaunoy et al., 2023). For full descriptions of the tasks, refer to Appendix F. Again following the precedent from previous SBI works, we present the calibration of the models trained over several simulation budgets ($|\mathcal{D}|$). CANVI was applied to an NPE, in which $\mathcal{D}_C$ was taken to be 10% of the simulation budgets and the remainder used for training. Calibration was completed in **under one second** for each task. Coverage was assessed 1,000 i.i.d. samples. Figure 2 demonstrates the miscalibration of the alternative approaches and the correction afforded with CANVI.

4.2. Predictive Efficiency

4.2.1. Training Epochs

We now study the application of CANVI to a collection of posteriors. In this experiment, $q^{(t)}$ is taken to be the t-th iterate of training a Neural Spline Flow family against the $\mathcal{L}_{\text{FAVI}}$ objective (Durkan et al., 2019). Generally, we expect efficiency to improve with training iterates, as $q(\Theta \mid x)$ better approximates $\mathcal{P}(\Theta \mid x)$; however, the most efficient iterate may not occur at $t = T$. Selecting an intermediate t is comparable to the practice of retaining the training iterate with the best validation performance in prediction tasks.

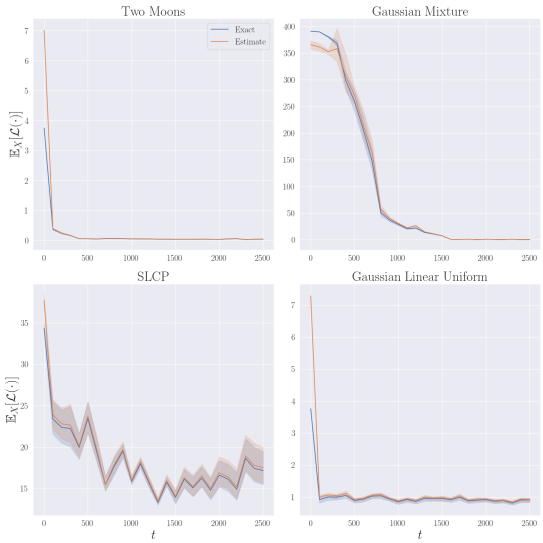


Figure 3. $\ell(q, \tau)$ and $\widehat{\ell}(q, \tau)$ for $\tau = \widehat{q}_{\mathcal{C}}^{(t)}(\alpha)$. Error bars across test batches are plotted.

We first study the validity of Assumption 3.7 by comparing $\widehat{\ell}(q, \tau)$ to $\ell(q, \tau)$. To compare to $\ell(q, \tau)$, we restrict experiments to projected posteriors of $\tilde{\Theta} = (\theta_1, \theta_2)$ for the Two Moons, Gaussian Mixture, SLCP, and Gaussian Linear

Table 1. Coverage rates and standard errors for θ before (rows 1-4) and after conformalization (rows 5-8) by CANVI, for ARCH (left table) and SED (right table), assessed by checking for inclusion of θ in the $1-\alpha$ highest density region. Non-conformalized regions were estimated empirically from batches of 1000 i.i.d. samples per point. $\ell(q, \widehat{q}_{\mathcal{C}}(0.05))$ were computed with explicit gridding, discretizing each dimension into 200 bins; such estimation was intractable for $\Theta \in \mathbb{R}^{11}$ for SEDs. $\widehat{\ell}$ was estimated using Algorithm 2, with $K = 1$ and $K = 10$ mixing discretizations.

$1-\alpha$	ELBO	IWBO	FAVI
0.50	0.0007 (0.0008)	0.1031 (0.0110)	0.5514 (0.0127)
0.75	0.0044 (0.0018)	0.1994 (0.0115)	0.7534 (0.0127)
0.90	0.0195 (0.0044)	0.3263 (0.0122)	0.8797 (0.0065)
0.95	0.0396 (0.0061)	0.4074 (0.0186)	0.9260 (0.0083)
0.50	0.4970 (0.0124)	0.5019 (0.0167)	0.5086 (0.0144)
0.75	0.7488 (0.0139)	0.7559 (0.0126)	0.7565 (0.0092)
0.90	0.8978 (0.0080)	0.9036 (0.0111)	0.9005 (0.0110)
0.95	0.9496 (0.0071)	0.9548 (0.0052)	0.9487 (0.0081)
$\ell(q, \widehat{q}_{\mathcal{C}}(0.05))$	1.3184 (0.2178)	1.4211 (0.1128)	0.7451 (0.0641)
$\widehat{\ell}_{K=1}(q, \widehat{q}_{\mathcal{C}}(0.05))$	50.6595 (8.1313)	69.6484 (2.2495)	19.3635 (0.8868)
$\widehat{\ell}_{K=10}(q, \widehat{q}_{\mathcal{C}}(0.05))$	1.4151 (0.2181)	1.5206 (0.1114)	0.8402 (0.0656)

$1-\alpha$	ELBO	IWBO	FAVI
0.50	0.1683 (0.0088)	0.6185 (0.0111)	0.4986 (0.0180)
0.75	0.4556 (0.0146)	0.7978 (0.0079)	0.7550 (0.0105)
0.90	0.5803 (0.0106)	0.8881 (0.0064)	0.9028 (0.0092)
0.95	0.6824 (0.0144)	0.9248 (0.0083)	0.9532 (0.0085)
0.50	0.4901 (0.0156)	0.4937 (0.0101)	0.5024 (0.0203)
0.75	0.7485 (0.0163)	0.7513 (0.0139)	0.7542 (0.0163)
0.90	0.8985 (0.0098)	0.9013 (0.0096)	0.9022 (0.0067)
0.95	0.9499 (0.0084)	0.9510 (0.0061)	0.9461 (0.0059)
$\widehat{\ell}(q, \widehat{q}_{\mathcal{C}}(0.05))$	∞	1.3849 (1.2357) $\times 10^9$	5.3732 (3.3302) $\times 10^6$

Uniform tasks, for which explicit gridding was tractable. $\widehat{\ell}(q^{(t)}, \widehat{q}_{\mathcal{C}}^{(t)}(\alpha))$ was estimated for a fixed $\alpha = 0.05$ and t taken every 100 training steps with 5 batches of 100 test points ($|\mathcal{D}_{\mathcal{T}}| = 100$) using $S = 10,000$ importance-weighted i.i.d. samples for each of $K = 10$ mixed samplers. From Figure 3, we see that $\widehat{\ell}(q, \tau)$ tracks closely to $\ell(q, \tau)$ across all tasks, giving credence to Assumption 3.7. We visualize the credible regions in Appendix H.

4.2.2. Training Objectives

We now similarly study $q^{(t)}$ across training objectives, taking one iterate of q trained against each of $\mathcal{L}_{\text{FAVI}}$, $\mathcal{L}_{\text{ELBO}}$, and $\mathcal{L}_{\text{IWBO}}$ ($K = 10$), giving us three amortized posteriors as input for CANVI, for a lag-one ARCH model:

$$y^{(m)} = \theta_1 y^{(m-1)} + e^{(m)},\ e^{(m)} = \xi^{(m)} \sqrt{0.2 + \theta_2 \left(e^{(m-1)}\right)^2},$$

where $y^{(0)} = 0$, $e^{(0)} = 0$, $M = 100$, and the $\xi^{(m)}$ are independent standard normal random variables (Thomas et al., 2022), detailed in Appendix F.9.

Table 1 shows that prior to conformalization, the variational posteriors are generally miscalibrated: training by the ELBO or IWBO results in significant under-coverage, as targeting either is known to find solutions that are mode-seeking. While the variational posterior obtained by FAVI is better calibrated, it still needs correction. As multiple posterior approximators were considered, CANVI had to be applied with recalibration. Table 1 shows that the recalibrated $1-\alpha$ prediction regions are nearly perfectly calibrated. Importantly, correction by CANVI can result in either larger or smaller $1-\alpha$ regions, depending on the direction of miscalibration. In settings where the variational posterior is overdispersed, applying CANVI results in smaller $1-\alpha$ density regions, explicitly shown in Appendix I.7. Table 1 also demonstrates using the better calibrated $\mathcal{L}_{\text{FAVI}}$-trained approximation results in higher efficiency compared to the $\mathcal{L}_{\text{ELBO}}$ and $\mathcal{L}_{\text{IWBO}}$ counterparts.

We also demonstrate the necessity of using mixed sampling for $\widehat{\ell}$. In particular, we compare the estimates when using only the variational posterior as a sampler ($\widehat{\ell}_{K=1}$) and when averaging 10 mixed samplers ($\widehat{\ell}_{K=10}$), from which we observe the estimator accuracy greatly improves with mixing, especially in the underdispersed IWBO and ELBO cases.

4.3. Galaxy Spectral Energy Distributions

We now present the application of CANVI to an important scientific problem. The spectrum of an astronomical object is measured via a spectrograph, which records the flux across a large grid of wavelength values (York et al., 2000; Abareshi et al., 2022), which we simulate with the Probabilistic Value-Added Bright Galaxy Survey simulator (PROVABGS). PROVABGS maps $\theta \in \mathbb{R}^{11}$ to galaxy spectra, detailed further and visualized in Appendix J.

A mixture of 20 Gaussian distributions was used as the variational posterior and trained against the $\mathcal{L}_{\text{FAVI}}, \mathcal{L}_{\text{ELBO}}$, and $\mathcal{L}_{\text{IWBO}}$ objectives, as in Section 4.2.2. Table 1 shows that the ELBO and IWBO tend to be overly concentrated, failing to contain the entire parameter vector θ in the $1-\alpha$ highest-density region often. FAVI, on the other hand, is reasonably well-calibrated. After applying CANVI, all three methods achieve nearly perfect calibration across a range of desired confidence levels. Of course, the utility of these corrected regions depends on the level of information contained in the original model. For the ELBO or IWBO cases, the corrected regions achieve statistical validity, but are too large to be informative. Notably, the underdispersion of the ELBO approximator led to $\widehat{q}_{\mathcal{C}}(0.05) = 0$ (up to machine precision), resulting in a volume estimate of ∞. For FAVI, on the other hand, application of CANVI results in statistical guarantees with minimal alterations to the high-density regions.

5. Discussion

We have presented CANVI, a framework for producing marginally calibrated, efficient prediction regions from a collection of variational approximators with minimal overhead in simulation-based inference settings. We view such guarantees on marginal coverage as an important first step toward increasing the utility of such inference algorithms for downstream applications, suggesting many directions for extension. Of immediate interest would be extending CANVI to cases of misspecified forward models $\mathcal{P}(X \mid \Theta)$ by leveraging work on conformal prediction under distribution shift (Tibshirani et al., 2019; Barber et al., 2022). Further, leveraging CANVI over functional spaces may enable guarantees over the full posterior distributions.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Abareshi, B., Aguilar, J., Ahlen, S., Alam, S., Alexander, D. M., Alfarsy, R., Allen, L., Prieto, C. A., Alves, O., Ameel, J., et al. Overview of the instrumentation for the Dark Energy Spectroscopic Instrument. *The Astronomical Journal*, 164(5):207, 2022.

Agrawal, A., Sheldon, D. R., and Domke, J. Advances in black-box VI: Normalizing flows, importance weighting, and optimization. *Advances in Neural Information Processing Systems*, 33:17358–17369, 2020.

Ambrogioni, L., Güçlü, U., Berezutskaya, J., Borne, E., Güçlütürk, Y., Hinne, M., Maris, E., and Gerven, M. Forward amortized inference for likelihood-free variational marginalization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 777–786. PMLR, 2019.

Angelopoulos, A. N. and Bates, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. Conformal prediction beyond exchangeability. *arXiv preprint arXiv:2202.13415*, 2022.

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

Boelts, J., Lueckmann, J.-M., Gao, R., and Macke, J. H. Flexible and efficient simulation-based inference for models of decision-making. *Elife*, 11:e77220, 2022.

Bornschein, J. and Bengio, Y. Reweighted wake-sleep. *arXiv preprint arXiv:1406.2751*, 2014.

Burda, Y., Grosse, R., and Salakhutdinov, R. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

Cranmer, K., Brehmer, J., and Louppe, G. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

Deistler, M., Goncalves, P. J., and Macke, J. H. Truncated proposals for scalable and hassle-free simulation-based inference. *arXiv preprint arXiv:2210.04815*, 2022.

Delaunoy, A., Hermans, J., Rozet, F., Wehenkel, A., and Louppe, G. Towards reliable simulation-based inference with balanced neural ratio estimation. *arXiv preprint arXiv:2208.13624*, 2022.

Delaunoy, A., Miller, B. K., Forré, P., Weniger, C., and Louppe, G. Balancing simulation-based inference for conservative posteriors. *arXiv preprint arXiv:2304.10978*, 2023.

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

Durkan, C., Bekasov, A., Murray, I., and Papamakarios, G. nflows: normalizing flows in PyTorch, November 2020a. URL https://doi.org/10.5281/zenodo.4296287.

Durkan, C., Murray, I., and Papamakarios, G. On contrastive learning for likelihood-free inference. In *International conference on machine learning*, pp. 2771–2781. PMLR, 2020b.

Falkiewicz, M., Takeishi, N., Shekhzadeh, I., Wehenkel, A., Delaunoy, A., Louppe, G., and Kalousis, A. Calibrating neural simulation-based inference with differentiable coverage probability. *arXiv preprint arXiv:2310.13402*, 2023.

Gilman, D., Bovy, J., Treu, T., Nierenberg, A., Birrer, S., Benson, A., and Sameie, O. Strong lensing signatures of self-interacting dark matter in low-mass haloes. *Monthly Notices of the Royal Astronomical Society*, 507(2):2432–2447, 2021.

Greenberg, D., Nonnenmacher, M., and Macke, J. Automatic posterior transformation for likelihood-free inference. In *International Conference on Machine Learning*, pp. 2404–2414. PMLR, 2019.

Gupta, C., Kuchibhotla, A. K., and Ramdas, A. Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognition*, 127:108496, 2022.

Hahn, C., Kwon, K., Tojeiro, R., Siudek, M., Canning, R. E., Mezcua, M., Tinker, J. L., Brooks, D., Doel, P., Fanning, K., et al. The DESI PRObabilistic Value-added Bright Galaxy Survey (PROVABGS) mock challenge. *The Astrophysical Journal*, 945(1):16, 2023.

Hermans, J., Begy, V., and Louppe, G. Likelihood-free mcmc with amortized approximate ratio estimators. In *International conference on machine learning*, pp. 4239–4248. PMLR, 2020.

Hermans, J., Banik, N., Weniger, C., Bertone, G., and Louppe, G. Towards constraining warm dark matter with stellar streams through neural simulation-based inference. *Monthly Notices of the Royal Astronomical Society*, 507 (2):1999–2011, 2021a.

Hermans, J., Delaunoy, A., Rozet, F., Wehenkel, A., and Louppe, G. Averting a crisis in simulation-based inference. *arXiv preprint arXiv:2110.06581*, 2021b.

Hezaveh, Y. D., Dalal, N., Marrone, D. P., Mao, Y.-Y., Morningstar, W., Wen, D., Blandford, R. D., Carlstrom, J. E., Fassnacht, C. D., Holder, G. P., et al. Detection of lensing substructure using ALMA observations of the dusty galaxy SDP.81. *The Astrophysical Journal*, 823(1): 37, 2016.

Hogg, D. W. and Blandford, R. The gravitational lens system B1422+231: dark matter, superluminal expansion and the Hubble constant. *Monthly Notices of the Royal Astronomical Society*, 268(4):889–893, 1994.

Jubin, B. Intrinsic volumes of sublevel sets. *arXiv preprint arXiv:1903.01592*, 2019.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Köhler, J., Krämer, A., and Noé, F. Smooth normalizing flows. *Advances in Neural Information Processing Systems*, 34:2796–2809, 2021.

Lee, J. M. and Lee, J. M. *Smooth manifolds*. Springer, 2012.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Lemos, P., Coogan, A., Hezaveh, Y., and Perreault-Levasseur, L. Sampling-based accuracy testing of posterior estimators for general inference. *arXiv preprint arXiv:2302.03026*, 2023.

Lueckmann, J.-M., Goncalves, P. J., Bassetto, G., Öcal, K., Nonnenmacher, M., and Macke, J. H. Flexible statistical inference for mechanistic models of neural dynamics. *Advances in neural information processing systems*, 30, 2017.

Lueckmann, J.-M., Boelts, J., Greenberg, D., Goncalves, P., and Macke, J. Benchmarking simulation-based inference. In *International Conference on Artificial Intelligence and Statistics*, pp. 343–351. PMLR, 2021.

Miller, B. K., Weniger, C., and Forré, P. Contrastive neural ratio estimation. *arXiv preprint arXiv:2210.06170*, 2022.

Murphy, K. P. *Probabilistic machine learning: an introduction*. MIT press, 2022.

Murphy, K. P. *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023. URL probml.ai.

Papamakarios, G. and Murray, I. Fast ε-free inference of simulation models with bayesian conditional density estimation. *Advances in neural information processing systems*, 29, 2016.

Papamakarios, G., Sterratt, D., and Murray, I. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 837–848. PMLR, 2019.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

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

Sesia, M. and Candès, E. J. A comparison of some conformal quantile regression methods. *Stat*, 9(1):e261, 2020.

Shafer, G. and Vovk, V. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

Thomas, O., Dutta, R., Corander, J., Kaski, S., and Gutmann, M. U. Likelihood-Free Inference by Ratio Estimation. *Bayesian Analysis*, 17(1):1 – 31, 2022. doi: 10.1214/20-BA1238. URL https://doi.org/10.1214/20-BA1238.

Tibshirani, R. J., Foygel Barber, R., Candes, E., and Ramdas, A. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

Vegetti, S. and Koopmans, L. Statistics of mass substructure from strong gravitational lensing: quantifying the mass fraction and mass function. *Monthly Notices of the Royal Astronomical Society*, 400(3):1583–1592, 2009.

Vegetti, S., Koopmans, L., Bolton, A., Treu, T., and Gavazzi, R. Detection of a dark substructure through gravitational imaging. *Monthly Notices of the Royal Astronomical Society*, 408(4):1969–1981, 2010.

Yang, Y. and Kuchibhotla, A. K. Finite-sample efficient conformal prediction. *arXiv preprint arXiv:2104.13871*, 2021.

Yao, Y., Vehtari, A., Simpson, D., and Gelman, A. Yes, but did it work?: Evaluating variational inference. In *International Conference on Machine Learning*, pp. 5581–5590. PMLR, 2018.

York, D. G., Adelman, J., Anderson Jr, J. E., Anderson, S. F., Annis, J., Bahcall, N. A., Bakken, J., Barkhouser,

R., Bastian, S., Berman, E., et al. The Sloan Digital Sky Survey: Technical summary. *The Astronomical Journal*, 120(3):1579, 2000.

Zhang, C., Bütepage, J., Kjellström, H., and Mandt, S. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):2008–2026, 2018.

A. Group Conditional CANVI

We now discuss how CANVI can be easily extended to provide a stronger notion of group conditional coverage over the purely marginal coverage over $\mathcal{X}$ that was provided in the paper. Here, instead of defining a global $\widehat{q}$, a collection $\{\widehat{q}_i\}_{i=1}^K$ is obtained by defining centroids over the $\mathcal{X}$ space $\{c_i\}_{i=1}^K$, constructing balls of radius ϵ around them $\mathcal{B}_\epsilon(c_i) \subset \mathcal{X}$, and performing calibration using $\mathcal{D}_{\mathcal{C}}^i := \{(x_j, \theta_j)\}_{j=1}^{\mathcal{N}_C}$ with $x_j \in \mathcal{B}_\epsilon(c_i)$.

To compare, we define a *global* quantile $\widehat{q}$ using KN samples. Specifically, we take $N = 100,000$ and $K = 5$, meaning calibration was performed using 500,000 i.i.d. samples for the *overall* calibration set. Coverage was assessed over $\alpha \in [0, 1]$ discretized at steps of .05 over $K = 5$ regions. Coverage was assessed over 10 batches of 100,000 i.i.d. test samples.

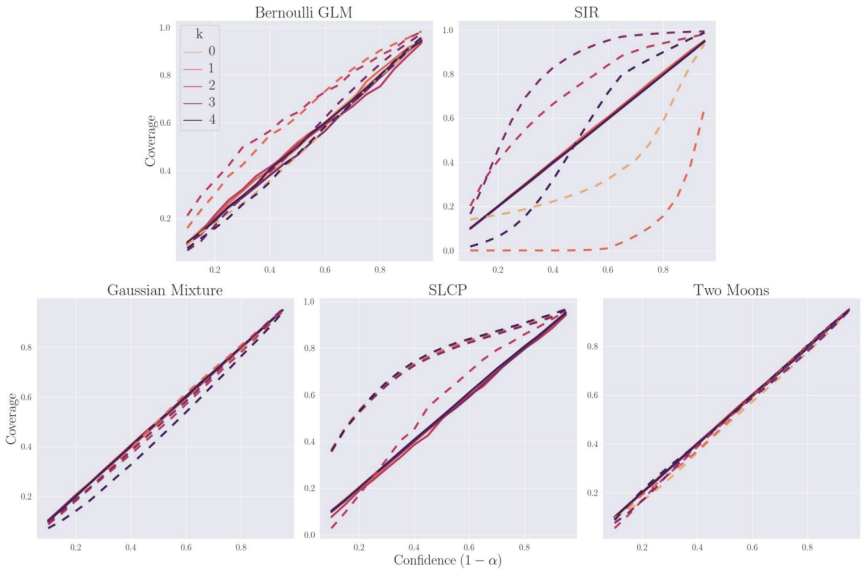


Figure 4. Calibration on the SBI benchmarks with overall quantile and region-specific quantiles, respectively the dashed and solid lines. Region-specific conformalized lines are slightly difficult to distinguish, as they all lie along the desired $y = x$ curve. Error bars from coverage assessments across test batches are plotted, although they are difficult to see due to the low variance between estimates across batches.

Figure 4 demonstrates the miscalibration of the regions produced using the overall quantile and subsequent correction with region-specific calibration. We additionally plot the calibration across the X space of the Two Moons task to demonstrate the non-uniformity of coverage across regions and subsequent corrections in Figure 5.

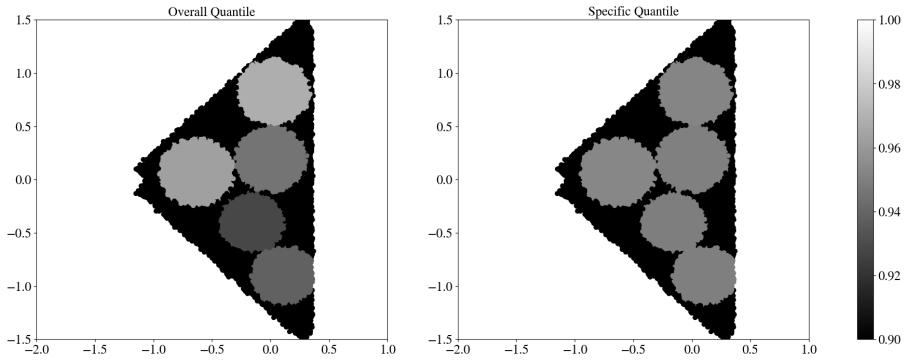


Figure 5. Coverage for Two Moons with the overall and region-specific quantiles with $\alpha = 0.05$.

B. CANVI Validity

Lemma B.1. *Let $\alpha \in (0,1)$ and*

$$q^{(*)}(\Theta \mid X), \widehat{q}^{(*)}_{\mathcal{R}}(\alpha) = \\ \text{CANVI}\left(\{q^{(t)}(\Theta \mid X)\}_{t=1}^{T}, \mathcal{P}(X,\Theta), 1-\alpha, N_{\mathcal{C}}, N_{\mathcal{T}}\right)$$

Let $(x', \theta') \sim \mathcal{P}(X,\Theta) \perp\!\!\!\perp \mathcal{D} \cup \mathcal{D}_{\mathcal{C}} \cup \mathcal{D}_{\mathcal{R}} \cup \mathcal{D}_{\mathcal{T}}$, with $\mathcal{D}$ being the data used to train $\{q^{(t)}(\Theta|X)\}_{t=1}^{T}$. Then $1-\alpha \leq \mathcal{P}(1/q^{(*)}(\theta'|x') \leq \widehat{q}^{(*)}_{\mathcal{R}}(\alpha))$.

Proof. Consider the score function $s^{(*)}(x,\theta) := 1/q^{(*)}(\theta \mid x)$. Observe that $(x',\theta') \cup \mathcal{D}_{\mathcal{R}}$ are jointly sampled i.i.d. from $\mathcal{P}(X,\Theta)$, independent of the datasets used to design $s^{(*)}(x,\theta)$, namely $\mathcal{D} \cup \mathcal{D}_{\mathcal{C}} \cup \mathcal{D}_{\mathcal{T}}$. Denoting $\mathcal{S}_{\mathcal{R}} := \{s^{(*)}(x_i,\theta_i)\}_{(x_i,\theta_i)\in\mathcal{D}_{\mathcal{R}}}$, scores $s^{(*)}(x',\theta') \cup \mathcal{S}_{\mathcal{R}}$, thus, too are i.i.d and, hence, exchangeable. The coverage guarantee then follows from the general theory of conformal prediction, presented in (Angelopoulos & Bates, 2021). □

C. Riemannian Manifolds

Let $(\mathcal{M}, g)$ denote a Riemannian manifold, with g denoting the metric tensor associated with this space. By definition, a manifold $\mathcal{M}$ is locally isomorphic to Euclidean space, from which we can define the notion of a tangent space $\mathcal{T}_z\mathcal{M}$ at each point $z \in \mathcal{M}$. The metric tensor g then defines a *local* notion of distance, namely $g : \mathcal{T}_z\mathcal{M} \times \mathcal{T}_z\mathcal{M} \to \mathbb{R}$. This, therefore, induces a local notion of length, namely for $x \in \mathcal{T}_z\mathcal{M}$, $||x||_z = \sqrt{g(x,x)}$.

Global distances over the manifold, therefore, can then be denoted by integrating such a local notion across a path γ. Concretely, a path is defined between $a, b \in \mathcal{M}$ by $\gamma : [0,1] \to \mathcal{M}$ such that $\gamma(0) = a, \gamma(1) = b$. The length of such a path is then $\ell(\gamma) := \int_0^1 ||\gamma'(t)||_{\gamma(t)} dt$. A natural choice of distance, therefore, is the minimum length of a constant velocity path, formally

$$d(a,b) = \inf_{\gamma : ||\gamma'(t)||=1, \gamma(0)=a, \gamma(1)=b} \ell(\gamma) \tag{8}$$

D. CANVI: Efficiency Analysis

D.1. Lipschitz Continuity of Efficiency

Lemma D.1. *Let $\ell(q,\tau)$ be as defined in Equation 3. If $\ell_x(q,\tau)$ is L-Lipschitz continuous in τ for any $x \in \mathcal{X}$, then $\ell(q,\tau)$ is L-Lipschitz continuous.*

Proof.

$$|\ell(q,\tau_1) - \ell(q,\tau_2)| = |\mathbb{E}_X\left[\mathcal{L}\left(\{\theta : 1/q(\Theta \mid X) \leq \tau_1\}\right) - \mathcal{L}\left(\{\theta : 1/q(\Theta \mid X) \leq \tau_2\}\right)\right]| \\ = \left|\int \mathcal{P}(x)\left[\ell_x(q,\tau_1) - \ell_x(q,\tau_2)\right]\right| \leq \int \mathcal{P}(x)\left|\ell_x(q,\tau_1) - \ell_x(q,\tau_2)\right| dx \leq L\left|\tau_1 - \tau_2\right| \int \mathcal{P}(x)dx = L\left|\tau_1 - \tau_2\right|,$$

completing the proof as desired. □

D.2. Predictive Efficiency of CANVI

The proof emerges through a reduction of CANVI to a special case of the "Validity First Conformal Prediction for the Smallest Prediction Set" (VFCP) algorithm presented in the (Yang & Kuchibhotla, 2021). The VFCP algorithm is provided for convenience in Algorithm 3. Note that we made the appropriate replacements in notations to match those presented in the main body of our paper for clarity.

To further underscore the parallel, we present an immaterially modified version of VFCP, where it is assumed the data are split *prior* to the algorithm execution into disjoint sets $\mathcal{D}, \mathcal{D}_{\mathcal{C}}$, and $\mathcal{D}_{\mathcal{R}}$. We assume input prediction methods $\mathcal{A}^{(1)}, ..., \mathcal{A}^{(T)}$ have been trained on $\mathcal{D}$. Finally, the presentation in (Yang & Kuchibhotla, 2021) allows for generic definitions of the measure function of $\mathcal{C}(x)$, referred to as "Width" therein. We present it with the particular Lebesgue measure function as presented in the main body to avoid confusion. These three modifications in the algorithm presentation have no manifest

effects in the proof of (Yang & Kuchibhotla, 2021), meaning all results as presented there apply with the appropriate choices of inputs in Algorithm 3.

We similarly consider an immaterially modified form of CANVI, that is, one in which $\mathcal{D}_{\mathcal{C}}$ and $\mathcal{D}_{\mathcal{R}}$ are pre-generated in precisely the fashion outlined Algorithm 1 and provided as input. Thus, inputs to CANVI parallel those of VFCP, namely $\{q^{(t)}(\Theta \mid X)\}_{t=1}^{T}, 1-\alpha, D_{\mathcal{C}}$, and $D_{\mathcal{R}}$.

Intuitively, the proof follows from the fact that CANVI is a special case of Algorithm 3, where the particular structural assumptions we make imply the corresponding assumptions as stated in the theorem from (Yang & Kuchibhotla, 2021). Algorithm 3 is presented using the so-called "nested set formulation" of conformal prediction. Briefly, the nested sets formulation starts by requiring the user to specify a *set-valued* function F_r instead of a score function. Sets must be nested over increasing r. Despite being seemingly more expressive, the equivalence of the nested set and split score conformal methods was demonstrated in (Gupta et al., 2022). In particular, for a chosen score function $s(x, \theta)$, the corresponding nested set function would be $F_{\tau}^{(t)} = \{(x, \theta) \mid s^{(t)}(x, \theta) \leq \tau\}$, where $s^{(t)}$ denotes the score function as defined with $\mathcal{A}^{(t)}$. The presentation of Algorithm 3 using the nested set formulation was, thus, a means to allow for the employment of relevant proof strategies from classical learning theory. We now proceed through the formal equivalence of CANVI and a special case of VFCP.

Algorithm 3 Validity First Conformal Prediction (VFCP) (Yang & Kuchibhotla, 2021)

1: **Procedure:** VFCP
2: **Inputs:** Predictors $\{\mathcal{A}^{(t)}\}_{t=1}^{T}$, Target coverage $1-\alpha$, Calibration set $D_{\mathcal{C}}$, Recalibration set $D_{\mathcal{R}}$, Score $s(x, \theta)$
3: Using $\mathcal{A}^{(t)}$, construct an increasing (nested) sequence of sets $\{F_{\tau}^{(t)}\}_{\tau \in \mathcal{T}}$, where $\mathcal{T} \subset \mathbb{R}$ and

$$F_{\tau}^{(t)} = \{(x, \theta) \mid s^{(t)}(x, \theta) \leq \tau\},$$

where $s^{(t)}$ denotes the score function as defined with $\mathcal{A}^{(t)}$
4: Compute conformal prediction set $\mathcal{C}^{(t)}$ based on $\{F_{\tau}^{(t)}\}_{\tau \in \mathcal{T}}$. Specifically, for each $(x_i, \theta_i) \in \mathcal{D}_{\mathcal{C}}$ and $t \in \{1, 2, ..., T\}$, denote its corresponding score as

$$s^{(t)}(x_i, \theta_i) := \inf_{\tau \in \mathcal{T}} \{(x_i, \theta_i) \in F_{\tau}^{(t)}\}$$

5: Compute the corresponding conformal prediction set as

$$\mathcal{C}_{\mathcal{C}}^{(t)} := \{(x, \theta) : s^{(t)}(x, \theta) \leq \widehat{q}_{\mathcal{C}}^{(t)}(\alpha)\},$$

where $\widehat{q}_{\mathcal{C}}^{(t)}(\alpha)$ is the $\lceil (|\mathcal{D}_{\mathcal{C}}| + 1)(1-\alpha) \rceil$-th largest element of $\{s^{(t)}(x_i, \theta_i)\}_{i \in \mathcal{D}_{\mathcal{C}}}$
6: Let $\mathcal{C}_{\mathcal{C}}^{(t)}(x) := \{\theta : (x, \theta) \in \mathcal{C}_{\mathcal{C}}^{(t)}\}$. Set

$$t^* := \arg\min_{1 \leq t \leq T} \mathbb{E}_X[\mathcal{L}(\mathcal{C}_{\mathcal{C}}^{(t)}(X))]$$

7: For each $(x_i, \theta_i) \in \mathcal{D}_{\mathcal{R}}$, define the conformal score

$$s^{(*)}(x_i, \theta_i) := \inf_{\tau \in \mathcal{T}} \{(x_i, \theta_i) \in F_{\tau}^{(t^*)}\}$$

8: Compute the corresponding conformal prediction set as

$$\mathcal{C}_{\mathcal{R}}^{(*)} := \{(x, \theta) : s^{(*)}(x, \theta) \leq \widehat{q}_{\mathcal{R}}^{(*)}(\alpha)\},$$

where $\widehat{q}_{\mathcal{R}}^{(*)}(\alpha) := \lceil (|\mathcal{D}_{\mathcal{R}}| + 1)(1-\alpha) \rceil$-th largest element of $\{s^{(*)}(x_i, \theta_i)\}_{i \in \mathcal{D}_{\mathcal{R}}}$
9: Return the prediction set $\mathcal{C}_{\mathcal{R}}^{(*)}$

We first state for reference the original theorem for VFCP, which we leverage in the proof of theorem.

Theorem D.2. *Suppose Assumption 3.2 holds. Let $\alpha \in (0, 1)$, $\mathcal{D}_{\mathcal{C}}$ and $\mathcal{D}_{\mathcal{R}}$ be drawn i.i.d. from $\mathcal{P}(X, \Theta)$, and*

$$\mathcal{C}_{\mathcal{R}}^{(*)} = \text{VFCP}\left(\mathcal{A}^{(t)}, 1-\alpha, D_{\mathcal{C}}, D_{\mathcal{T}}, s(x, \theta)\right)$$

If, for $r \geq \max\{\sqrt{\log(4T/\delta)/2N_{\mathcal{C}}}, 2/N_{\mathcal{C}}\}$ *and* $\delta \in [0,1]$*, Assumption 3.3 holds, then with probability at least* $(1-\delta)$*,*

$$\mathbb{E}_X[\mathcal{L}(\mathcal{C}_{\mathcal{R}}^{(*)}(X))] \leq \min_{1\leq t\leq T} \mathbb{E}_X[\mathcal{L}(\mathcal{C}_{\mathcal{R}}^{(t)}(X))] + 3L_W L_{[T]} \left[\left(\frac{\log(4T/\delta)}{N_{\mathcal{C}}}\right)^{\gamma/2} + \left(\frac{2}{N_{\mathcal{C}}}\right)^{\gamma}\right], \tag{9}$$

where γ*,* L_W*, and* $L_{[T]} = \max_{1\leq t\leq T} L_t$ *are constants defined in Assumptions 3.3 and 3.2 (Yang & Kuchibhotla, 2021).*

We now present the proof of our theorem. The proof proceeds in two steps. We first demonstrate the CANVI and VFCP algorithms are equivalent if we assume access to the exact $\ell(q,\tau)$, which, coupled with the assumptions imposed on q, allows us to directly leverage the results of Theorem D.2. We then demonstrate, under Assumption 3.7, we recover the desired bound even if t^* is chosen with $\widehat{\ell}(q,\tau)$.

Theorem D.3. *Suppose for any* $x \in \mathcal{X}$ *and* $t = 1,...,T$*,* $q^{(t)}(\theta \mid x) \in \mathcal{C}^3(\mathbb{R}^n)$ *is bounded above and for* $\theta \neq 0$*,* $\mathcal{L}(\{\theta : \nabla_\theta q^{(t)}(\theta|x)\}) = 0$*. Further assume* $P(X,\Theta)$ *is bounded above. Let* $\alpha \in (0,1)$ *and*

$$q^{(*)}(\Theta \mid X), \widehat{q}_{\mathcal{R}}^{(*)}(\alpha) = \\ \text{CANVI}\left(\{q^{(t)}(\Theta \mid X)\}_{t=1}^T, \mathcal{P}(X,\Theta), 1-\alpha, N_{\mathcal{C}}, N_{\mathcal{T}}\right)$$

If, for $r \geq \max\{\sqrt{\log(4T/\delta)/2N_{\mathcal{C}}}, 2/N_{\mathcal{C}}\}$ *and* $\delta \in [0,1]$*, Assumption 3.3 holds and for* $\Delta, \epsilon > 0$ *Assumption 3.7 holds, then with probability at least* $(1-\epsilon)(1-\delta)$*,*

$$\ell(q^{(*)}, \widehat{q}_{\mathcal{R}}^{(*)}(\alpha)) \leq \min_{1\leq t\leq T} \ell(q^{(t)}, \widehat{q}_{\mathcal{R}}^{(t)}(\alpha)) + \Delta + 3L_W L_{[T]} \left[\left(\frac{\log(4T/\delta)}{N_{\mathcal{C}}}\right)^{\gamma/2} + \left(\frac{2}{N_{\mathcal{C}}}\right)^{\gamma}\right], \tag{10}$$

where γ*,* L_W*, and* $L_{[T]} = \max_{1\leq t\leq T} L_t$ *are constants defined in Assumptions 3.3 and 3.2.*

Proof. Let $\mathcal{C}_{\mathcal{R}}^{(*)} = \text{VFCP}\left(\{q^{(t)}(\Theta \mid X)\}_{t=1}^T, 1-\alpha, D_{\mathcal{C}}, D_{\mathcal{T}}, 1/q(\theta \mid x)\right)$. We first wish to demonstrate $\mathcal{C}_{\mathcal{C},\mathcal{R}}^{(t),\text{VFCP}}(x) = \mathcal{C}_{\mathcal{C},\mathcal{R}}^{(t),\text{CANVI}}(x)$ for any fixed x, where we use the $\mathcal{C},\mathcal{R}$ condensed notation to mean this equality holds both under $\mathcal{D}_{\mathcal{C}}$ and $\mathcal{D}_{\mathcal{R}}$. This equivalence can be shown in demonstrating the equivalence in corresponding scores, as the resulting empirical score distributions and hence quantiles over $\mathcal{D}_{\mathcal{C}}$ and $\mathcal{D}_{\mathcal{R}}$ and finally future prediction regions follow to then be equivalent by the algorithm structure.

This equivalence follows as a straightforward instance of the equivalence of the set-valued and standard conformal prediction frameworks. In particular, we recover the original score formulation, as:

$$s^{(t)}(x_i,\theta_i) := \inf_{\tau\in\mathcal{T}}\{(x_i,\theta_i) \in F_\tau^{(t)}\} = \inf_{\tau\in\mathcal{T}}\{(x_i,\theta_i) \in \{(x,\theta) \mid 1/q^{(t)}(\theta \mid x) \leq \tau\}\} = 1/q^{(t)}(\theta_i \mid x_i).$$

The bound under access to the exact $\ell(q,\tau)$ then follows from Theorem D.2 under demonstration of the appropriate assumptions. Assumption 3.3 holds by assumption. Assumption 3.2 holds for any $\vartheta^{(t)} := \{\theta : \nabla_\theta q^{(t)}(\theta|x)\}$ by Corollary 3.5. In the application of Assumption 3.2 for the proof of Theorem D.2, it suffices for $\mathcal{F}_t^{-1}(1-\alpha) \in \vartheta^{(t)}$ and $\mathcal{F}_t^{-1}(1-\alpha+1/(N_{\mathcal{C}}+1)) \in \vartheta^{(t)}$. Since $\mathcal{L}(\{\theta : \nabla_\theta q^{(t)}(\theta|x)\}) = 0$ and $\mathcal{P}(X,\Theta)$ is bounded above, this means $\mathcal{P}_{X,\Theta}(\mathcal{F}_t^{-1}(1-\alpha) \in \vartheta^{(t)}) = 1$ and $\mathcal{P}_{X,\Theta}(\mathcal{F}_t^{-1}(1-\alpha+1/(N_{\mathcal{C}}+1)) \in \vartheta^{(t)}) = 1$. Thus, if t^* is chosen in CANVI using $\ell(q,\tau)$, by Theorem D.2, with probability $1(1-\delta) = 1-\delta$,

$$\ell(q^{(*)}, \widehat{q}_{\mathcal{R}}^{(*)}(\alpha)) \leq \min_{1\leq t\leq T} \ell(q^{(t)}, \widehat{q}_{\mathcal{R}}^{(t)}(\alpha)) + 3L_W L_{[T]} \left[\left(\frac{\log(4T/\delta)}{N_{\mathcal{C}}}\right)^{\gamma/2} + \left(\frac{2}{N_{\mathcal{C}}}\right)^{\gamma}\right].$$

The extension of this to the case of interest, where t^* is chosen using $\widehat{\ell}(q,\tau)$, is now a straightforward application of Assumption 3.7, from which we have that $\exists\, \Delta, \epsilon > 0$, such that with probability at least $1-\epsilon$

$$\left|\ell(q^{(\widehat{t}^*)}, \widehat{q}_{\mathcal{R}}^{(\widehat{t}^*)}(\alpha)) - \ell(q^{(t^*)}, \widehat{q}_{\mathcal{R}}^{(t^*)}(\alpha))\right| < \Delta \implies \ell(q^{(\widehat{t}^*)}, \widehat{q}_{\mathcal{R}}^{(\widehat{t}^*)}(\alpha)) < \ell(q^{(t^*)}, \widehat{q}_{\mathcal{R}}^{(t^*)}(\alpha)) + \Delta. \tag{11}$$

Switching back to denoting $\ell(q^{(*)}, \widehat{q}_{\mathcal{R}}^{(*)}(\alpha)) := \ell(q^{(\widehat{t}^*)}, \widehat{q}_{\mathcal{R}}^{(\widehat{t}^*)}(\alpha))$, this implies that, with probability $(1-\epsilon)(1-\delta)$,

$$\ell(q^{(*)}, \widehat{q}_{\mathcal{R}}^{(*)}(\alpha)) \leq \ell(q^{(t^*)}, \widehat{q}_{\mathcal{R}}^{(t^*)}(\alpha)) + \Delta \leq \min_{1\leq t\leq T} \ell(q^{(t)}, \widehat{q}_{\mathcal{R}}^{(t)}(\alpha)) + \Delta + 3L_W L_{[T]} \left[\left(\frac{\log(4T/\delta)}{N_{\mathcal{C}}}\right)^{\gamma/2} + \left(\frac{2}{N_{\mathcal{C}}}\right)^{\gamma}\right],$$

completing the proof as desired.

□

E. Gaussian Hölder Continuity

Theorem E.1. *Let Θ and X be zero-mean unit-variance Gaussian random variables with correlation ρ. Let $q^{(t)}(\theta|x) = \mathcal{N}(\theta; tx, 1-\rho^2)$. Let $\kappa := t^2 - 2t\rho + 1$ and $r > 0$. Then $F_t^{-1}(z)$, is 1-Hölder continuous on $[1-\alpha, 1-\alpha+r]$ with Hölder constant*

$$\frac{\kappa\Phi^{-1}(\frac{1-\alpha}{2})\sqrt{\exp\left(\frac{\kappa}{1-\rho^2}\Phi^{-1}(\frac{1-\alpha}{2})^2 - \frac{(1-\alpha)^2}{2}\right)}}{\sqrt{(1-\rho^2)/2}} \tag{12}$$

Proof. Notice that in the bivariate Gaussian case, we have closed forms for the following:

$$\Theta \mid x \sim \mathcal{N}(\rho x, 1-\rho^2) \qquad X \mid \theta \sim \mathcal{N}(\rho\theta, 1-\rho^2)$$
$$\Theta \sim \mathcal{N}(0,1) \qquad X \sim \mathcal{N}(0,1).$$

We wish to find the distribution of $s(X,\Theta) = 1/q(\Theta \mid X)$ jointly over X, Θ to find $F_t^{-1}(z)$ explicitly. The CDF of this score can be computed as follows:

$$\mathcal{P}\left(1/q_t(\Theta \mid X) \le q\right) = \mathcal{P}\left(\sqrt{2\pi(1-\rho^2)}e^{\frac{(tX-\Theta)^2}{2(1-\rho^2)}} \le q\right)$$
$$= \mathcal{P}\left(R^2 \le 2(1-\rho^2)\log\left(\sqrt{\frac{q^2}{2\pi(1-\rho^2)}}\right)\right) = \mathcal{P}\left(R \le \sqrt{(1-\rho^2)\log\left(\frac{q^2}{2\pi(1-\rho^2)}\right)}\right),$$

where $R := |tX - \Theta|$. From the above calculation, $F_t^{-1}(z)$ must satisfy:

$$\mathcal{P}\left(R \le \sqrt{(1-\rho^2)\log\left(\frac{F_t^{-1}(z)^2}{2\pi(1-\rho^2)}\right)}\right) = z.$$

Notice now that, since (X,Θ) are bivariate Gaussian:

$$tX - \Theta \sim \mathcal{N}(0, t^2+1-2t\rho) \implies R \sim \text{HalfNormal}(t^2+1-2t\rho).$$

Therefore, the z quantile of R is $\sqrt{t^2+1-2t\rho}\Phi^{-1}\left(1-\frac{\alpha}{2}\right)$. Solving for $F_t^{-1}(z)$ in this quantile produces the final threshold:

$$\sqrt{(1-\rho^2)\log\left(\frac{F_t^{-1}(z)^2}{2\pi(1-\rho^2)}\right)} = \sqrt{t^2+1-2t\rho}\Phi^{-1}\left(1-\frac{\alpha}{2}\right)$$
$$\implies \log\left(\frac{F_t^{-1}(z)^2}{2\pi(1-\rho^2)}\right) = \frac{t^2+1-2t\rho}{1-\rho^2}\left(\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right)^2$$
$$\implies F_t^{-1}(z) = \sqrt{2\pi(1-\rho^2)\exp\left(\frac{t^2+1-2t\rho}{1-\rho^2}\left(\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right)^2\right)}.$$

The Hölder constant follows from bounding the derivative of $F_t^{-1}(z)$, namely

$$\frac{\kappa\Phi^{-1}(\frac{z}{2})\sqrt{\exp\left(\frac{\kappa}{1-\rho^2}\Phi^{-1}(\frac{z}{2})^2 - \frac{(z)^2}{2}\right)}}{\sqrt{(1-\rho^2)/2}}$$

This expression is monotonically decreasing in z and hence maximized for $z = 1-\alpha$ in the interval, giving the desired expression.

□

F. Simulation-Based Inference Benchmarks

The benchmark tasks are a subset of those provided by (Lueckmann et al., 2021). For convenience, we provide brief descriptions of the tasks curated by this library; however, a more comprehensive description of these tasks can be found in their manuscript.

F.1. Gaussian Linear

10-dimensional Gaussian model with a Gaussian prior:

$$\textbf{Prior}: \mathcal{N}(0, 0.1 \odot I)$$
$$\textbf{Simulator}: x \mid w \sim \mathcal{N}(x \mid w, 0.1 \odot I)$$

F.2. Gaussian Linear Uniform

10-dimensional Gaussian model with a uniform prior:

$$\textbf{Prior}: \mathcal{U}(-1, 1)$$
$$\textbf{Simulator}: x \mid w \sim \mathcal{N}(x \mid w, 0.1 \odot I)$$

F.3. SLCP with Distractors

Simple Likelihood Complex Posterior (SLCP) with Distractors has uninformative dimensions in the observation over the standard SLCP task:

$$\textbf{Prior}: \mathcal{U}(-3, 3)$$
$$\textbf{Simulator}: x \mid w = p(y) \text{ where } p \text{ reorders}$$
$$y \text{ with a fixed random order}$$
$$y_{[1:8]} \sim \mathcal{N}\left(\begin{bmatrix} w_1 \\ w_2 \end{bmatrix}, \begin{bmatrix} w_3^4 & w_3^2 w_4^2 \tanh(w_5) \\ w_3^2 w_4^2 \tanh(w_5) & w_4^4 \end{bmatrix}\right),$$
$$y_{9:100} \sim \frac{1}{20} \sum_{i=1}^{20} t_2(\mu^i, \Sigma^i), \mu^i \sim \mathcal{N}(0, 15^2 I),$$
$$\Sigma^i_{j,k} \sim \mathcal{N}(0, 9), \Sigma^i_{j,j} = 3e^a, a \sim \mathcal{N}(0, 1),$$

F.4. Bernoulli GLM Raw

10-parameter GLM with Bernoulli observations and Gaussian prior. Observations are not sufficient statistics, unlike the standard "Bernoulli GLM" task:

$$\textbf{Prior}: \beta \sim \mathcal{N}(0, 2), f \sim \mathcal{N}(0, (F^T F)^{-1})$$
$$F_{i,i-2} = 1, F_{i,i-1} = -2$$
$$F_{i,i} = 1 + \sqrt{\frac{i-1}{9}}, F_{i,j} = 0; i \leq j$$
$$\textbf{Simulator}: x^{(i)} \mid w \sim \text{Bern}(\eta(v_T^{(i)} f + \beta)),$$
$$\eta(\odot) = \exp(\odot)/(1 + \exp(\odot))$$

F.5. Gaussian Mixture

A mixture of two Gaussians, with one having a much broader covariance structure:

$$\textbf{Prior}: \beta \sim \mathcal{U}(-10, 10)$$
$$\textbf{Simulator}: x \mid w \sim 0.5\mathcal{N}(x \mid w, I) + 0.5\mathcal{N}(x \mid w, .01I)$$

F.6. Two Moons

Task with a posterior that has both global (bimodal) and local (crescent-shaped) structure:

$$\textbf{Prior}: \beta \sim \mathcal{U}(-1, 1)$$
$$\textbf{Simulator}: x \mid w =$$
$$\begin{bmatrix} r\cos(\alpha) + 0.25 \\ r\sin(\alpha) \end{bmatrix} + \begin{bmatrix} -|w_1 + w_2|/\sqrt{2} \\ (-w_1 + w_2)/\sqrt{2} \end{bmatrix}$$
$$\alpha \sim \mathcal{U}(-\pi/2, \pi/2), r \sim \mathcal{N}(0.1, 0.01^2)$$

F.7. SIR

Epidemiology model with S (susceptible), I (infected), and R (recovered). A contact rate β and mean recovery rate of γ are used as follows:

$$\textbf{Prior}: \beta \sim \text{LogNormal}(\log(0.4), 0.5),$$
$$\gamma \sim \text{LogNormal}(\log(1/8), 0.2)$$
$$\textbf{Simulator}: x = (x^{(i)})_{i=1}^{10}; x^{(i)} \mid w \sim \text{Bin}(1000, \frac{I}{N}),$$
$$\text{where } I \text{ is simulated from:}$$
$$\frac{dS}{dt} = -\beta\frac{SI}{N}, \qquad \frac{dI}{dt} = \beta\frac{SI}{N} - \gamma I, \qquad \frac{dR}{dt} = \gamma I$$

F.8. Lotka-Volterra

An ecological model commonly used in describing dynamics of competing species. w parameterizes this interaction as $w = (\alpha, \beta, \gamma, \delta)$:

$$\textbf{Prior}: \alpha \sim \text{LogNormal}(-.125, 0.5)$$
$$\beta \sim \text{LogNormal}(-3, 0.5), \gamma \sim \text{LogNormal}(-.125, 0.5)$$
$$\delta \sim \text{LogNormal}(-3, 0.5)$$
$$\textbf{Simulator}: x = (x^{(i)})_{i=1}^{10},$$
$$x_{1,i} \mid w \sim \text{LogNormal}(\log(X), 0.1),$$
$$x_{2,i} \mid w \sim \text{LogNormal}(\log(Y), 0.1)$$
$$\text{where } X, Y \text{ is simulated from:}$$
$$\frac{dX}{dt} = \alpha X - \beta XY, \qquad \frac{dY}{dt} = -\gamma Y + \delta XY$$

F.9. ARCH

The two-dimensional parameter $\theta = (\theta_1, \theta_2)$ includes both an autoregressive component (θ_1) and a component controlling the level of conditional noise (θ_2). Given a full realization of the time series $y_{1:T}$, we aim to amortize inference over θ.

Priors are taken to be $\theta_1 \sim \text{Unif}(-1, 1)$ and $\theta_2 \sim \text{Unif}(0, 1)$. One important change from the model of (Thomas et al., 2022) is that we fix $e^{(0)} = 0$ rather than drawing this quantity from a standard Gaussian.

The Adam optimizer was used with a learning rate of 0.0001 for 25,000 training steps for each of the three methods: IWBO ($K = 10$), ELBO, FAVI. Due to constraints arising from the uniform priors on the parameters, the IWBO and ELBO implementations rely on logit transformations of the latent random variables θ to avoid zero-density regions that result in undefined gradients. The encoder network is trained to learn distributions on the unconstrained space of the transformed random variable θ', and visualizations are produced by performing the inverse transformation. While FAVI avoids these issues because it is likelihood-free, for an apples-to-apples comparison we also implement FAVI on the unconstrained latent space as well.

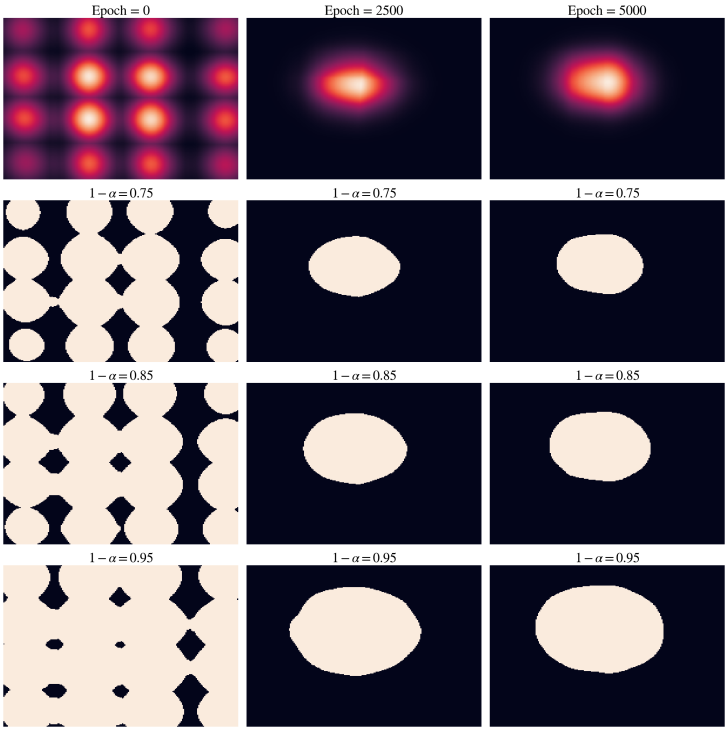
G. Training Details

All encoders were implemented in PyTorch (Paszke et al., 2019) with a Neural Spline Flow architecture. The NSF was built using code from (Durkan et al., 2020a). Specific architecture hyperparameter choices were taken to be the defaults from (Durkan et al., 2020a) and are available in the code. Optimization was done using Adam (Kingma & Ba, 2014) with a learning rate of 10^{-3} over 5,000 training steps. Minibatches were drawn from the corresponding prior $\mathcal{P}(\Theta)$ and simulator $\mathcal{P}(X \mid \Theta)$ as specified per task in the preceding section. Training these models required between 10 minutes and two hours using an Nvidia RTX 2080 Ti GPUs for each of the SBI tasks.

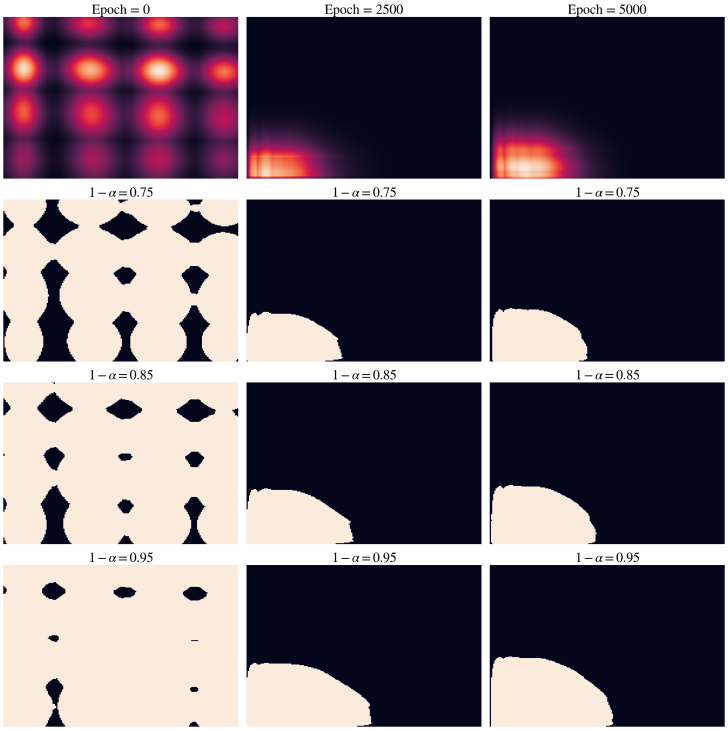
H. Prediction Regions

Credible regions for $q_\varphi(\Theta \mid x)$ (for a single $x \sim \mathcal{P}(X)$) on a subset of the SBI benchmark tasks are plotted below over varying degrees of training, as indicated in the figures. As expected, the efficiency of the prediction regions improves over training across all tasks, resulting in *smaller* regions for each target coverage $1 - \alpha$.

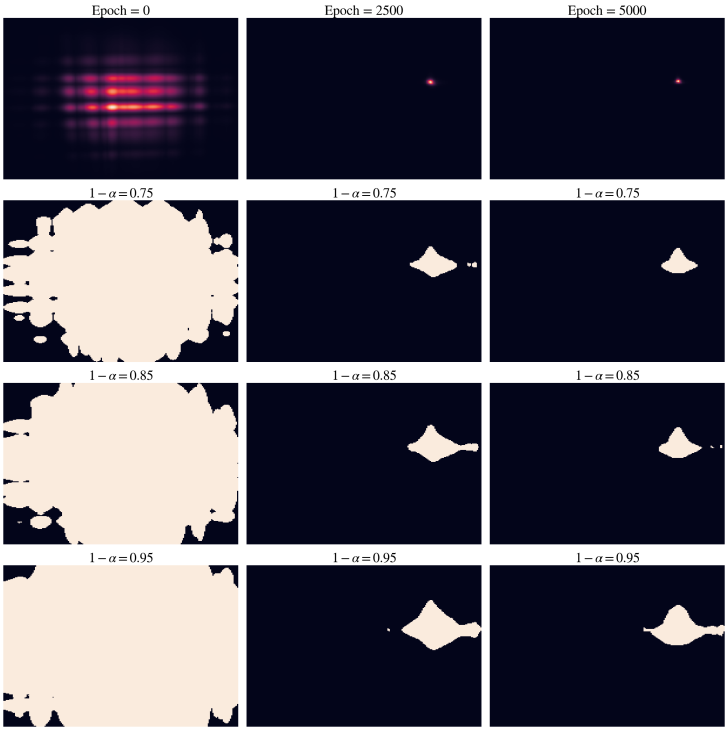
H.1. Gaussian Linear



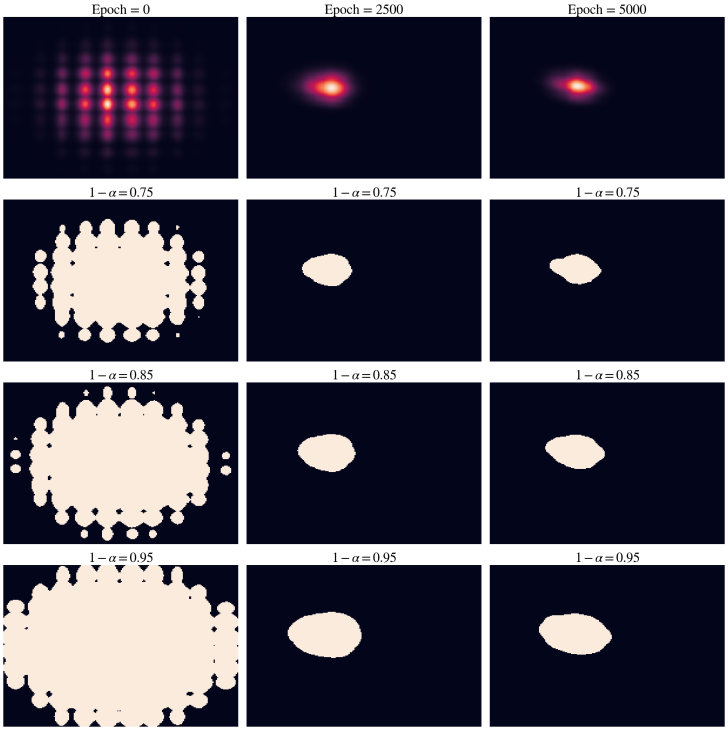
H.2. Gaussian Linear Uniform



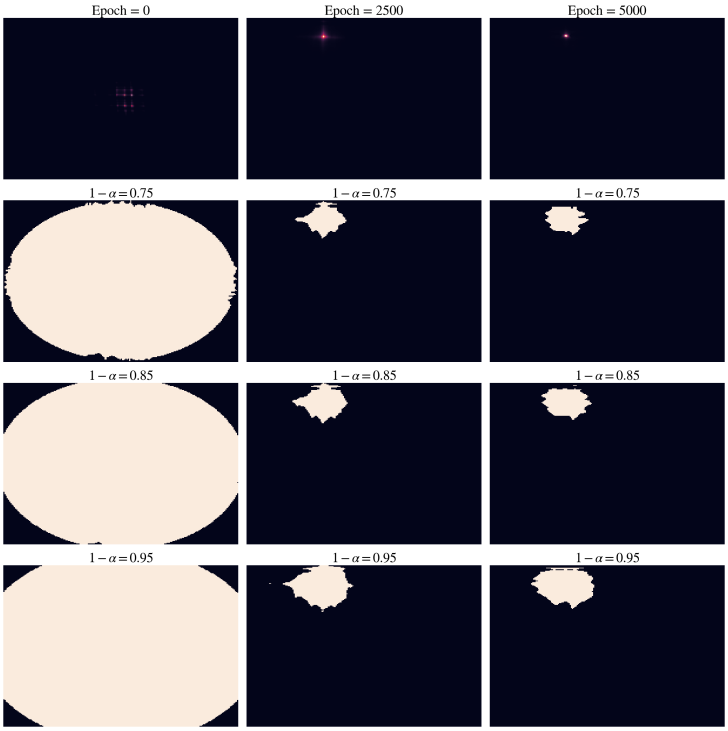
H.3. SLCP



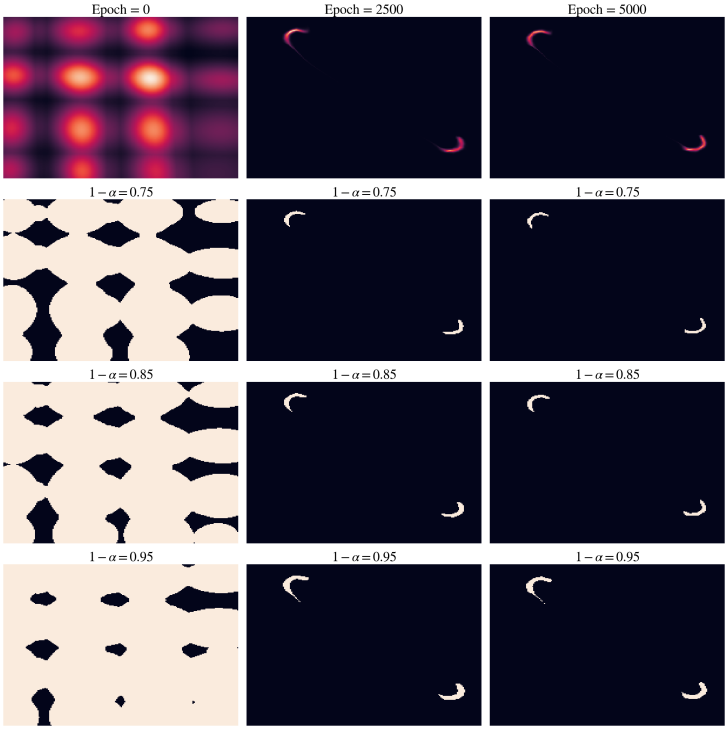
H.4. Bernoulli GLM Raw



H.5. Gaussian Mixture



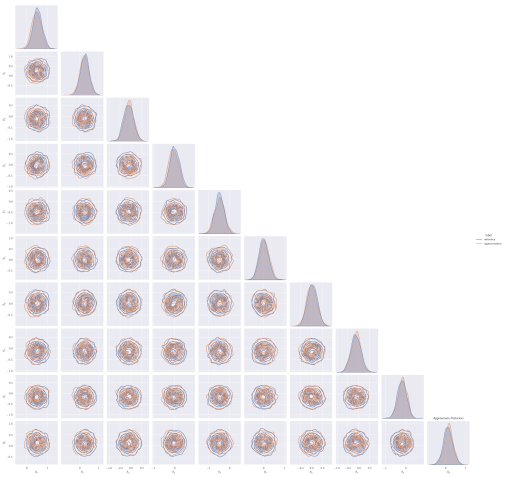
H.6. Two Moons



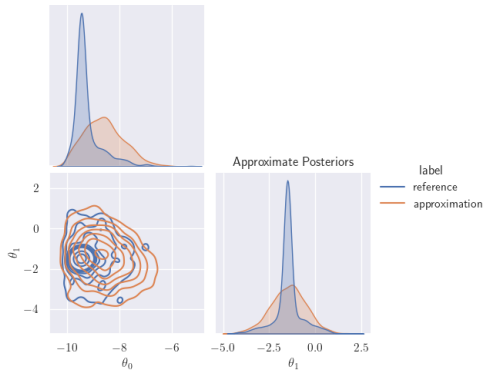
I. Posteriors

We provide visualizations of approximate and reference posteriors (produced with MCMC from (Lueckmann et al., 2021)) to justify the overdispersion claims made on the variational approximation procedure.

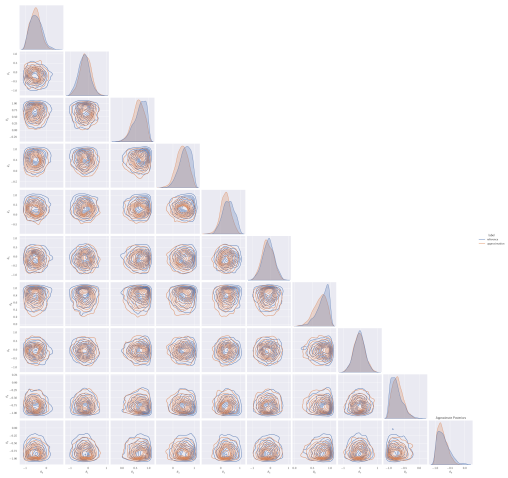
I.1. Gaussian Linear



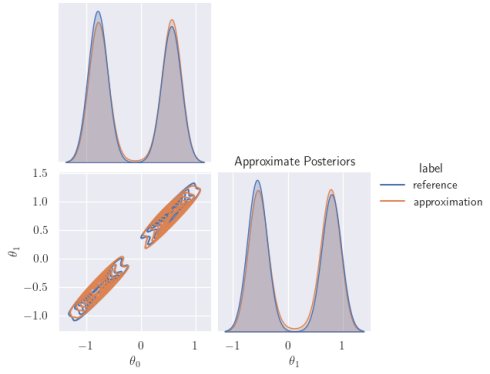
I.2. Gaussian Mixture



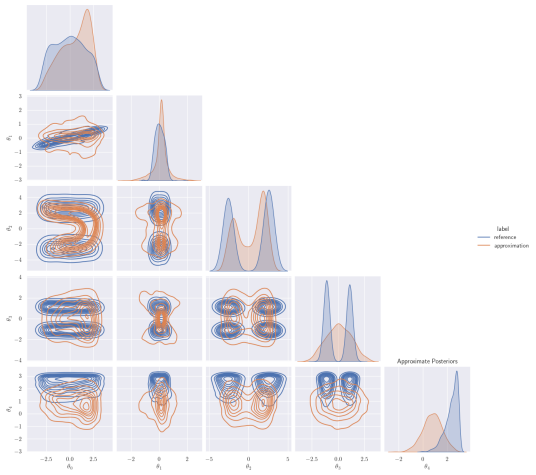
I.3. Gaussian Linear Uniform



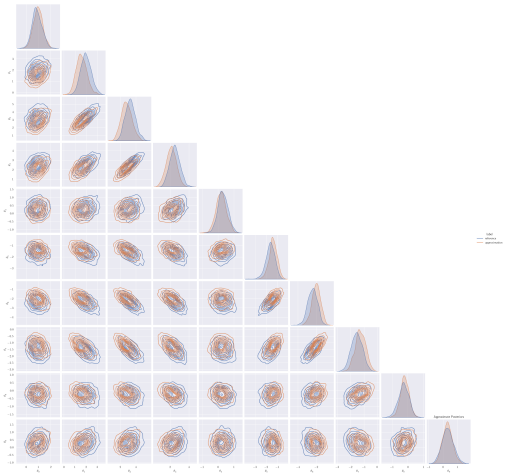
I.4. Two Moons



I.5. SLCP

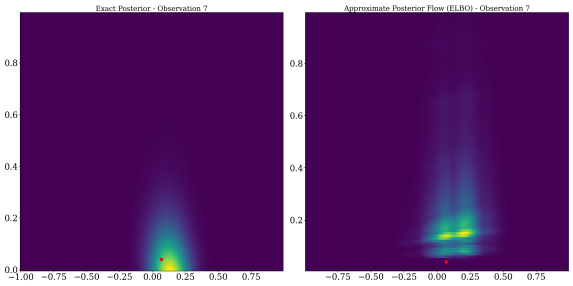


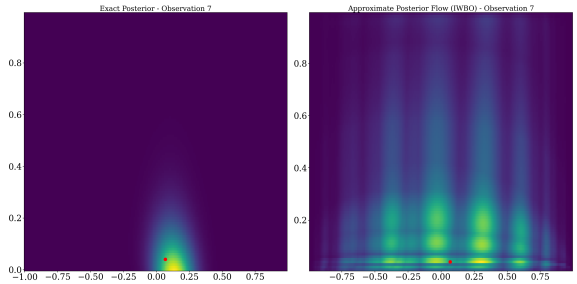
I.6. Bernoulli GLM

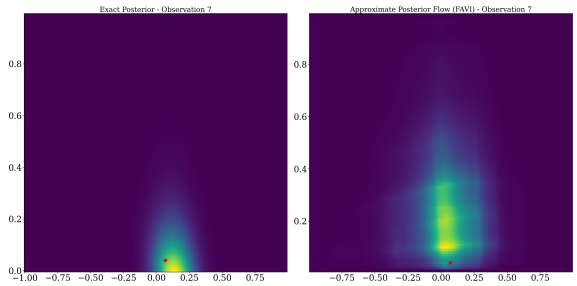


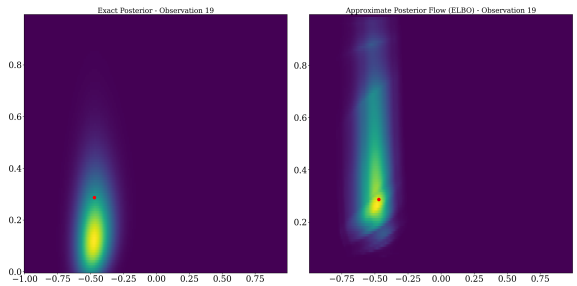
I.7. ARCH

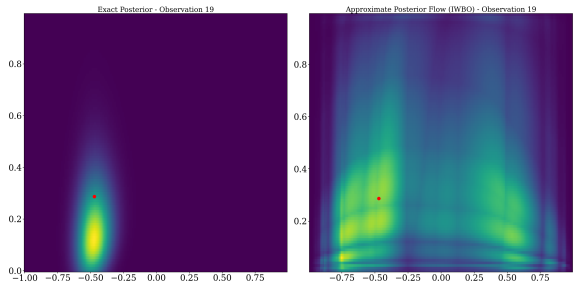
For selected points from the training sets, we show the exact and approximate posteriors obtained by training according to the ELBO, IWBO, and FAVI objectives. The ground-truth parameter value is noted in red.

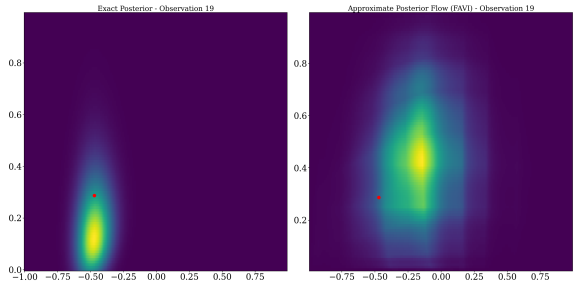



Exact Posterior - Observation 7
Approximate Posterior Flow (IWBO) - Observation 7


Exact Posterior - Observation 7
Approximate Posterior Flow (FAVI) - Observation 7


Exact Posterior - Observation 19
Approximate Posterior Flow (ELBO) - Observation 19


Exact Posterior - Observation 19
Approximate Posterior Flow (IWBO) - Observation 19


Exact Posterior - Observation 19
Approximate Posterior Flow (FAVI) - Observation 19

Depending on the miscalibration of the variational posterior on either per-objective or per-point basis, the conformalized $1-\alpha$ high-density region (HDR) either shrinks or expands, and can be visualized in two dimensions. Figure 6, Figure 7 show examples where the 50% prediction regions obtained from the FAVI-learned variational posterior shrinks and grows after applying CANVI.

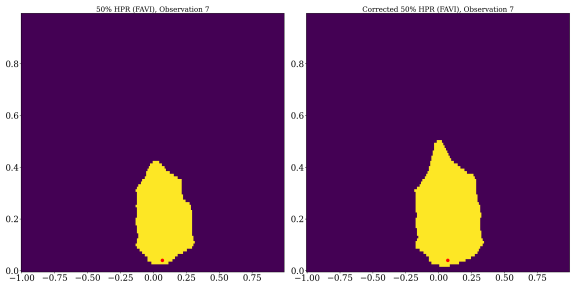


Figure 6. 50% prediction region, observation 7, before and after applying CANVI.

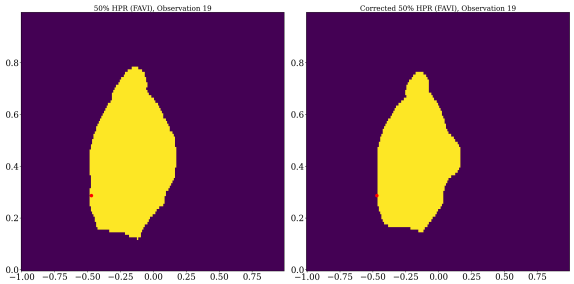


Figure 7. 50% prediction region, observation 8, before and after applying CANVI.

We visualize the efficiency of the iterates q_φ across training iterations in Figure 8. Recalling the form of Equation 6, it is unsurprising that FAVI tends to be more efficient, as it trains using simulated data (over which Equation 6 is computed). To estimate Equation 6, at every 500 training steps, we simulate 20 θ, x pairs from the forward model. A larger number of Monte Carlo samples can be used but at an increased cost. As the resulting estimates are noisy, we smooth the resulting series with a Savitzky-Golay filter with a window length of 10 and third-degree polynomial order for better visualization.

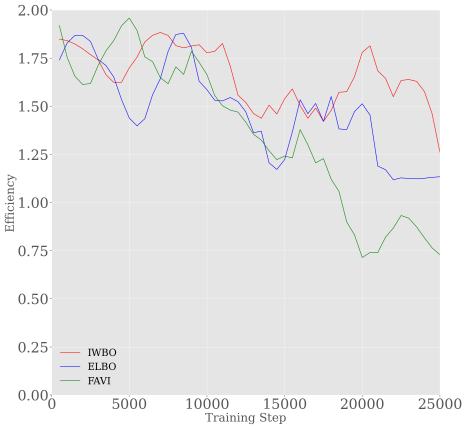


Figure 8. Efficiency estimates for variational posteriors trained by ELBO, IWBO, and FAVI across training iterations.

J. SED Experimental Details

The PROVABGS emulator (Section 4.3) was trained to minimize the MSE using normalized simulated PROVABGS outputs with fixed log stellar mass parameter (Hahn et al., 2023). Training data were generated from PROVABGS with a fixed magnitude parameter ($\theta_0 = 10.5$), resampled onto a 5 Angstrom grid, and normalized to integrate to one. After training, forward passes through the emulator are significantly faster than the base simulator. We provide two simulated draws from our emulator in Figure 9. We use the recommended priors from (Hahn et al., 2023) on the remaining eleven parameters. As these are highly constrained (uniform priors, vastly different scales, and a 4-dimensional vector on the simplex), we similarly operate on an unconstrained, 10-dimensional space by invertible transformations. All three methods were trained for 10,000 steps using the Adam optimizer with learning rate 0.0001. $K = 1000$ was used for the IWBO.

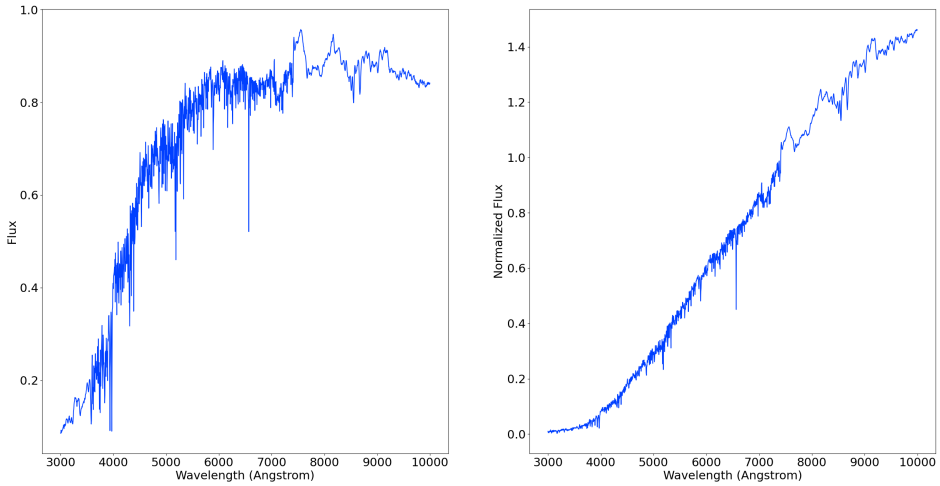


Figure 9. Two example draws from our neural network emulator of PROVABGS.

We let the output of our simulator be a mean parameter $\mu \in \mathbb{R}^{1400}$, and generate observed data independently binwise via $x_i \sim \mathcal{N}(\mu_i, |\mu_i|^2\sigma^2)$ for a fixed hyperparameter $\sigma = .1$ and all $i = 1, \ldots, 1400$. We adopt this noise model for simplicity, as it imposes a fixed signal-to-noise ratio (SNR) of $1/\sigma = 10$ across all spectra and wavelength bins.