Variational Inference with Continuously-Indexed Normalizing Flows

Anthony Caterini[1] **Rob Cornish**[1] **Dino Sejdinovic**[1] **Arnaud Doucet**[1]

[1]Department of Statistics, University of Oxford

Abstract

Continuously-indexed flows (CIFs) have recently achieved improvements over baseline normalizing flows on a variety of density estimation tasks. CIFs do not possess a closed-form marginal density, and so, unlike standard flows, cannot be plugged in directly to a variational inference (VI) scheme in order to produce a more expressive family of approximate posteriors. However, we show here how CIFs can be used as part of an *auxiliary* VI scheme to formulate and train expressive posterior approximations in a natural way. We exploit the conditional independence structure of multi-layer CIFs to build the required auxiliary inference models, which we show empirically yield low-variance estimators of the model evidence. We then demonstrate the advantages of CIFs over baseline flows in VI problems when the posterior distribution of interest possesses a complicated topology, obtaining improved results in both the Bayesian inference and surrogate maximum likelihood settings.

1 INTRODUCTION

Variational inference (VI) has emerged as a fast, albeit biased, alternative to Markov chain Monte Carlo for Bayesian inference. VI methods attempt to minimize the KL divergence from a parametrized family of distributions to a true posterior over latent variables. The expressiveness of this family is essential for good performance, with underexpressive models leading to both increased bias and underestimation of posterior variance (Yin and Zhou, 2018).

If the density of the approximate posterior is available in closed-form, then the variational family is said to be *explicit*. Explicit models allow for straightforward estimation of the VI objective, but can often lead to reduced expressiveness, which limits their performance overall. Mean-field VI (Blei et al., 2017), for example, imposes restrictive independence assumptions between the latent variables of interest.

Normalizing flows (Tabak et al., 2010; Rezende and Mohamed, 2015) provide an alternative family of explicit density models that yield improved expressiveness compared with mean field alternatives. These methods push samples from a simple base distribution (typically Gaussian) through parametrized bijections to produce complex, yet still exact, density models. Normalizing flows have performed well in tasks requiring explicit density models (e.g. (Louizos and Welling, 2017; Papamakarios et al., 2017; Ho et al., 2019)), including VI, where flows have demonstrated the ability to improve the quality of approximate posteriors (Rezende and Mohamed, 2015; Durkan et al., 2019).

Although normalizing flows can directly improve the expressiveness of mean-field VI schemes, their inherent bijectivity remains quite restrictive. We can overcome this limitation by instead using continuously-indexed flows (CIFs) (Cornish et al., 2020). CIFs relax the bijectivity constraint of standard normalizing flows by augmenting them with continuous index variables, thus parametrizing an *implicit* density model defined as the marginalization over these additional indexing variables. Beyond being well-grounded theoretically, CIFs also have empirically demonstrated the ability to outperform relevant normalizing flow baselines in the context of density estimation, and thus it is sensible to investigate the performance of CIFs in VI.

A difficulty in applying CIFs to VI – and implicit models more generally – is that their marginal distribution is intractable, precluding evaluation of the standard VI objective. However, conveniently, CIFs still admit a tractable *joint* distribution over the variables of interest (latent variables in VI) and the auxiliary indexing variables. We can therefore appeal to the framework of *auxiliary variational inference* (AVI) (Agakov and Barber, 2004), which facilitates the training of implicit models with tractable joint densities as variational inference models. CIFs also *already* prescribe a model for inferring auxiliary variables – typically required

Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2021), PMLR 161:44–53.

in AVI schemes – suggesting that CIFs are a natural fit here. AVI methods more generally have shown improved expressiveness over the explicit counterparts in several settings (Burda et al., 2016; Yin and Zhou, 2018), and are becoming more popular with the rise of implicit models overall (Tran et al., 2017; Lawson et al., 2019; Kleinegesse et al., 2020), suggesting that this framework is able to overcome any supposed drawbacks associated with not having access to explicit densities.

In this work, we show that these benefits are also realized when CIFs are applied within the AVI framework. We first describe how CIFs can be used as the variational family in AVI, naturally incorporating the components of CIF models designed for density estimation, and we explain how we can also *amortize* these inference models. We then empirically demonstrate the advantages of using CIFs over standard normalizing flows for modelling posteriors with complicated topologies, and additionally how CIFs can facilitate maximum likelihood estimation of the parameters of complex latent-variable generative models.

2 CONTINUOUSLY-INDEXED FLOWS FOR VARIATIONAL INFERENCE

In this section we first review necessary background on variational inference (VI) – including auxiliary variational inference (AVI) – and continuously-indexed flows (CIFs). We then describe how CIFs naturally fit in as a class of auxiliary variational posteriors, and extend to include amortization. We summarize the results of this section in Algorithm 1.

2.1 VARIATIONAL INFERENCE

Given a joint probability density $p_{X,Z}$, with observed data $X \in \mathcal{X}$ and latent variable $Z \in \mathcal{Z}$, variational inference (VI) provides us with a means to approximate the intractable posterior $p_{Z|X}(\cdot \mid x)$. This is accomplished by introducing a parametrized approximate posterior[1] q_Z, and maximizing the evidence lower bound (ELBO)

$$\mathcal{L}_1(x) \coloneqq \mathbb{E}_{z\sim q_Z}[\log p_{X,Z}(x,z) - \log q_Z(z)] \tag{1}$$

with respect to the parameters of q_Z. This is equivalent to minimizing the KL divergence between q_Z and the true posterior $p_{Z|X}(\cdot \mid x)$.

Explicit VI methods, such as mean-field approaches or normalizing flow models, define q_Z in such a way that it can be evaluated pointwise. Although this approach is computationally convenient, the expressiveness of the resulting methods can often be limited. To improve on this, implicit methods define q_Z typically through some type of sampling process with intractable marginal distribution, such as the pushforward of a simple distribution through an unrestricted deep neural network. These methods can be quite powerful but also challenging to optimize, especially in the context of VI (Tran et al., 2017), as we lose the tractability of (1).

Auxiliary Variational Inference In contexts where q_Z is obtained as $q_Z(z) \coloneqq \int q_{Z,U}(z,u)\,\mathrm{d}u$ for some joint density $q_{Z,U}$ that can be sampled from and evaluated pointwise, its parameters can be learned via *auxiliary variational inference* (AVI) (Agakov and Barber, 2004). We refer to U here as an *auxiliary* variable. These approaches introduce an auxiliary inference distribution $r_{U|Z}$ and optimize

$$\mathcal{L}_2(x) \coloneqq \mathbb{E}_{(z,u)\sim q_{Z,U}}\left[\log \frac{p_{X,Z}(x,z)\cdot r_{U|Z}(u \mid z)}{q_{Z,U}(z,u)}\right]. \tag{2}$$

Key to this approach is the fact that $\mathcal{L}_1(x) \geq \mathcal{L}_2(x)$, and that this bound is tight when $r_{U|Z} = q_{U|Z}$, which holds because

$$\mathcal{L}_1(x) = \mathcal{L}_2(x) + \mathbb{E}_{z\sim q_Z}\left[D_{\mathrm{KL}}(q_{U|Z}(\cdot|z)||r_{U|Z}(\cdot|z))\right]. \tag{3}$$

As such, optimizing the parameters of $r_{U|Z}$ jointly with those of $q_{Z,U}$ will encourage learning better approximations to the true posterior $p_{Z|X}$. Note that, although we now are optimizing a lower bound on $\mathcal{L}_1$, we are also optimizing over a larger family of approximate posteriors which may end up yielding a better optimum (cf. Proposition 3.1 below).

2.2 CONTINUOUSLY-INDEXED FLOWS

We now describe in detail the *continuously-indexed flow* (CIF) model (Cornish et al., 2020), which we intend to incorporate into an AVI scheme. CIFs define a density q_Z over $\mathcal{Z}$ as the Z-marginal of

$$W \sim q_W, \quad U \sim q_{U|W}(\cdot \mid W), \quad Z = G(W;U), \tag{4}$$

where q_W is a noise distribution over $\mathcal{Z}$, $q_{U|W}$ is a conditional distribution over $\mathcal{U}$ describing an auxiliary indexing variable, and $G : \mathcal{Z}\times\mathcal{U} \to \mathcal{Z}$ is a function such that $G(\cdot;u)$ is a bijection for each $u \in \mathcal{U}$. For all $z \in \mathcal{Z}$, the density model q_Z is then given by the intractable integral $q_Z(z) \coloneqq \int q_{Z,U}(z,u)\,\mathrm{d}u$ over the tractable joint density $q_{Z,U}$ given by

$$\begin{aligned} q_{Z,U}(z,u) &= q_W\big(G^{-1}(z;u)\big) \\ &\quad\times q_{U|W}\big(u \mid G^{-1}(z;u)\big)\left|\det \mathrm{D}_z G^{-1}(z;u)\right| \end{aligned} \tag{5}$$

for all $z \in \mathcal{Z}$ and $u \in \mathcal{U}$, where G^{-1} denotes the inverse of G (and $\mathrm{D}_z G^{-1}$ the Jacobian of G^{-1}) with respect to its first argument z (see Appendix A for a derivation). Typically, $q_{U|W}$ is chosen to be conditionally Gaussian with mean and covariance as the outputs of neural networks taking the conditioning variables W and Z as input, and

$$G(w;u) \coloneqq e^{s(u)} \odot (g(w) + t(u)), \tag{6}$$

[1] We may also *amortize* q_Z and replace it with the conditional $q_{Z|X}$, especially when using VI to facilitate generative modelling. Further discussion on amortization is deferred to Subsection 2.4.

where $g : \mathcal{Z} \to \mathcal{Z}$ is some base bijection, $s, t : \mathcal{U} \to Z$ are arbitrary neural networks, and $\odot$ denotes elementwise multiplication. Cornish et al. (2020) used the model (4) in the context of density estimation to model the generative process of a set of i.i.d. data.

Multi-layer CIFs Cornish et al. (2020) also propose to improve the expressiveness of (4) by taking the noise distribution q_W to be a CIF model itself. Applying this recursively L times, we can take q_Z to be the W_L-marginal in the following model:

$$W_0 \sim q_{W_0}, \quad U_\ell \sim q_{U_\ell|W_{\ell-1}}(\cdot \mid W_{\ell-1}) \\ W_\ell = G_\ell(W_{\ell-1}; U_\ell), \tag{7}$$

where $\ell \in \{1, \dots, L\}$. Here q_{W_0} is typically a mean-field Gaussian and each $G_\ell : \mathcal{Z} \times \mathcal{U} \to \mathcal{Z}$ is bijective in its first argument. Practically, multi-layer CIF models have demonstrated far more representational power than single-layer versions, although we note that we can still view this multi-layer model as an instance of (4) for certain choices of $q_{U|W}$ and G (as in Appendix B).

Auxiliary Inference Distribution The intractability of q_Z arising from both (4) and (7) precludes direct maximum likelihood estimation. Cornish et al. (2020) therefore introduce an auxiliary *backward* distribution, either $r_{U|Z}$ or $r_{U_{1:L}|Z}$ respectively, to enable training of CIFs through an amortized ELBO. Particularly noteworthy is the structure of this distribution in the multi-layer case. The optimal choice for $r_{U_{1:L}|Z}$ would be $q_{U_{1:L}|Z}$, which can be shown to factorize as $q_{U_{1:L}|Z}(u_{1:L} \mid z) = \prod_{\ell=1}^{L} q_{U_\ell|W_\ell}(u_\ell \mid w_\ell)$, where $w_L \coloneqq z$ and $w_\ell \coloneqq G_{\ell+1}^{-1}(w_{\ell+1}; u_{\ell+1})$ recursively for $\ell \in \{1, \dots, L-1\}$. Although this gives us the form of $q_{U_{1:L}|Z}$, the backward distributions $q_{U_\ell|W_\ell}$ are not generally available in closed form. However this does at least motivate defining $r_{U_{1:L}|Z}$ to have the same form, which can be done by introducing (reparametrizable) densities $r_{U_\ell|W_\ell}$ and setting

$$r_{U_{1:L}|Z}(u_{1:L} \mid z) \coloneqq \prod_{\ell=1}^{L} r_{U_\ell|W_\ell}(u_\ell \mid w_\ell) \tag{8}$$

with w_ℓ defined as above. The densities for $r_{U_\ell|W_\ell}$ are also taken to be parametrized conditional Gaussians. This structured inference procedure induces a natural weight-sharing scheme between the forward and backward directions of the model, as both are defined using G_ℓ.

2.3 CIF MODELS IN AVI

We can use CIFs as the family of approximate posteriors q_Z in VI by appealing to the framework of AVI. Starting with the single-layer version, we see from (5) that CIFs admit a tractable joint distribution $q_{Z,U}$ over latent and auxiliary variables. We can then plug this distribution into (2), noting also that CIFs already prescribe a form for $r_{U|Z}$ and thus are a natural fit within an AVI scheme. However, we must take one additional step to formulate an objective amenable to optimization, as naïvely substituting $q_{Z,U}$ into (2) produces an expectation over a distribution containing the parameters of G itself. To address this, we show in Appendix A how to rewrite this as an expectation over $q_{W,U}$ rather than $q_{Z,U}$, obtaining the objective

$$\mathbb{E}_{(w,u)\sim q_{W,U}}\left[\log \frac{p_{X,Z}(x,z)\cdot r_{U|Z}(u \mid z)}{q_{W,U}(w,u)\cdot|\det \mathrm{D}_w G(w;u)|^{-1}}\right], \tag{9}$$

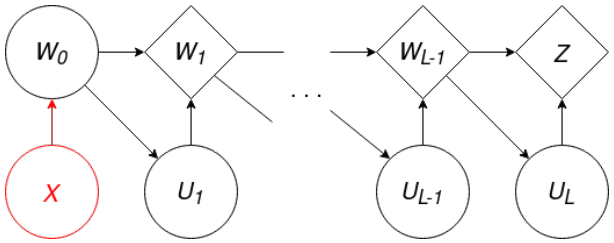
where we write $z \coloneqq G(w;u)$ for readability. We always select $q_{W,U}$ to be reparametrizable (Kingma and Welling, 2014), which makes the objective straightforward to optimize via stochastic gradient descent with respect to the parameters of q, r, and G. Note however that $r_{U|Z}$ need not necessarily be reparametrizable. This "direction" of reparametrization contrasts with CIF models for density estimation which require $r_{U|Z}$ – not $q_{U|W}$ – to be reparametrizable. Further discussion demonstrating that CIF models in density estimation and VI can be viewed as "opposites" of each other is provided in Subsection 3.3 and Appendix D.

Multi-layer CIFs in AVI We can also use multi-layer CIFs as past of an AVI scheme. Now, each of the $q_{U_\ell|W_{\ell-1}}$ distributions in (7) is chosen to be reparametrizable, again contrasting with (Cornish et al., 2020) which does not require reparametrizable distributions here. Figure 1a graphically displays the joint model $q_{Z,U_{1:L}}$. We also adopt the form of $r_{U_{1:L}|Z}$ from (8) and demonstrate this auxiliary inference procedure in Figure 1b, although we do not require the individual $r_{U_\ell|W_\ell}$ distributions to be reparametrizable. Being able to have $r_{U_{1:L}|Z}$ match the structure of the true auxiliary posterior $q_{U_{1:L}|Z}$ is likely useful in lowering the variance of estimators of the ELBO and gradients thereof.

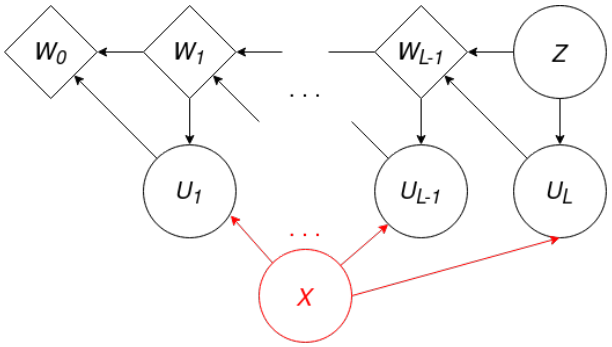
We can now substitute our definitions for $q_{Z,U_{1:L}}$ (implied by (7)) and $r_{U_{1:L}|Z}$ into (2) to derive an optimization objective for training multi-layer CIFs as the approximate posterior in VI. We again must be careful about reparametrization, as we need to write the objective as an expectation over $q_{W_0,U_{1:L}}$ instead of $q_{Z,U_{1:L}}$ to be able optimize all parameters of q, analogously to (9). Further details on how to do this are provided in Appendix B, with the full objective given in (1) there. Algorithm 1 describes how to compute an unbiased estimator of this objective, from which we can then obtain unbiased gradients via automatic differentiation.

2.4 AMORTIZATION

VI methods can also be used to provide a surrogate objective for maximum likelihood estimation of the parameters of latent-variable models, particularly for deep generative models such as the variational auto-encoder



(a) Sampling $Z \sim q_{Z|X}$ as defined in (4)



(b) Sampling $U_{1:L} \sim r_{U_{1:L}|Z,X}$ as defined in (8)

Figure 1: Diagrams demonstrating how to sample from the CIF approximate posterior (left) and the auxiliary inference model (right). The red highlighting corresponds to amortization – these can be ignored for models not requiring amortization.

(VAE) (Kingma and Welling, 2014; Rezende et al., 2014). In these settings, the goal is to maximize the marginal log-likelihood $\sum_i \log p_X(x_i)$ over the observed data $\{x_i\}_i$, with respect to the parameters of p, where $\log p_X(x) := \log \int p_{X,Z}(x,z)\,\mathrm{d}z$ is a density model containing parametrized $p_{X,Z}(x,z)$. This integral is often intractable, and thus we resort to maximizing the ELBO (1) (or (2)) with respect to both the parameters of p and q as it bounds the marginal log-likelihood from below. In this case, we would like to *amortize* the cost of variational inference across an entire dataset, rather than compute a brand new approximate posterior for each datapoint, and so we parametrize our variational distribution as an explicit function of the data.

We can readily incorporate amortization into the single-layer CIF by replacing q_W with $q_{W|X}$ in (4), and $r_{U|Z}$ with $r_{U|Z,X}$ in (9) since the true auxiliary posterior $q_{U|Z,X}$ will now carry an explicit dependence on the data X. For multi-layer CIFs, it is again straightforward to incorporate amortization into the model for q by replacing q_{W_0} with $q_{W_0|X}$ in (7). Additional care must be taken when constructing the *auxiliary* inference model r, however, as the explicit dependence on data will appear in each term of the factorization of the true auxiliary posterior $q_{U_{1:L}|Z,X}$. We thus structure $r_{U_{1:L}|Z,X}$ similarly:

$$r_{U_{1:L}|Z,X}(u_{1:L} \mid z, x) := \prod_{\ell=1}^{L} r_{U_\ell|W_\ell,X}(u_\ell \mid w_\ell, x),$$

where w_ℓ is as defined in (8). The full amortized objective is given in (2) in the Appendix. Figure 1 graphically demonstrates how to incorporate amortization into both q and r, while Algorithm 1 includes a provision for this case as well.

Algorithm 1 Unbiased L-layer CIF ELBO estimator

function ELBO(x, amortized)
  if amortized **then**
    $q_0 \leftarrow q_{W_0|X}(\cdot \mid x)$
  else
    $q_0 \leftarrow q_{W_0}$
  $w_0 \sim q_0$
  $\Delta \leftarrow -\log q_0(w)$
  for $\ell = 1, \ldots, L$ **do**
    $u \sim q_{U_\ell|W_{\ell-1}}(\cdot \mid w_{\ell-1})$
    $w_\ell \leftarrow G_\ell(u; w_{\ell-1})$
    if amortized **then**
      $r_\ell \leftarrow r_{U_\ell|W_\ell,X}(\cdot \mid w_\ell, x)$
    else
      $r_\ell \leftarrow r_{U_\ell|W_\ell}(\cdot \mid w_\ell)$
    $\Delta \leftarrow \Delta + \log r_\ell(u) - \log q_{U_\ell|W_{\ell-1}}(u \mid w_{\ell-1})$
      $+ \log |\det \mathrm{D}G_\ell(w_{\ell-1}; u)|$
  return $\Delta + \log p_{X,Z}(x, w_L)$

3 COMPARISON TO RELATED WORK

In this section we first compare against methods using explicit normalizing flow models for variational inference, then move on to a discussion of implicit VI methods, and lastly compare the structure of CIFs in basic density estimation to CIFs in VI.

3.1 NORMALIZING FLOWS FOR VI

Normalizing flows (NFs) originally became popular as a method for increasing the expressiveness of explicit variational inference models (Rezende and Mohamed, 2015). NF methods define q_Z as the Z-marginal of

$$W \sim q_W, \qquad Z = g(W), \tag{10}$$

where $g : \mathcal{Z} \to \mathcal{Z}$ is a bijection. We can equivalently write q_Z as $q_Z := g_\# q_W$, where $g_\# q_W$ denotes the *pushforward* of the distribution q_W under the map g. Using the change of variable formula, we can rewrite (1) here as

$$\mathcal{L}_1(x) = \mathbb{E}_{w \sim q_W}\left[\log \frac{p_{X,Z}(x, g(w))}{q_W(w) \cdot |\det \mathrm{D}g(w)|^{-1}}\right]. \tag{11}$$

This objective is a simplified version of the CIF VI objective (9). The following proposition, which we adapt here to the VI setting from Cornish et al. (2020, Proposition 4.1), shows that generalizing from (11) to (9) is beneficial, as a CIF model trained by this auxiliary bound will perform at least as well in inference as its corresponding baseline flow trained via maximization of (11).

Proposition 3.1. *Assume a CIF inference model with components $q^{\phi}_{U|W}$, $r^{\phi}_{U|Z}$, and G_ϕ is parametrized by $\phi \in \Phi$, with associated objective* (9) *denoted as $\mathcal{L}^{\phi}_2$. Suppose there exists $\psi \in \Phi$ such that for some bijection g, $G_\psi(\cdot; u) = g(\cdot)$ for all $u \in \mathcal{U}$. Similarly, suppose $q^{\psi}_{U|W}$ and $r^{\psi}_{U|Z}$ are such that, for some density ρ on $\mathcal{U}$, $q^{\psi}_{U|W}(\cdot \mid w) = r^{\psi}_{U|Z}(\cdot \mid z) = \rho(\cdot)$ for all $w, z \in \mathcal{Z}$. For a given $x \in \mathcal{X}$, if $\mathcal{L}^{\phi}_2(x) \geq \mathcal{L}^{\psi}_2(x)$,*

$$D_{KL}\Big(q^{\phi}_Z \mid\mid p_{Z|X}(\cdot \mid x)\Big) \leq D_{KL}\big(g_{\#}q_W \mid\mid p_{Z|X}(\cdot \mid x)\big)\,.$$

The proof of this result, from which we also see that $\mathcal{L}^{\psi}_2(x) = \mathcal{L}_1(x)$ (where $\mathcal{L}_1(x)$ is as written in (11)) for all $x \in \mathcal{X}$, is provided in Appendix C. This shows that optimizing a CIF using the auxiliary ELBO $\mathcal{L}_2$ will produce at least as good of an inference model (as measured by the KL divergence) as a baseline normalizing flow optimized using the marginal ELBO (11), in the limit of infinite samples from the inference model. Note that our choices of G from (6) and $q_{U|W}$ and $r_{U|Z}$ as conditionally Gaussian will usually entail the conditions of Proposition 3.1, since for example we have $G(w; u) = g(w)$ in (6) if the final layer weights in the s and t networks are zero. We also empirically confirm that Proposition 3.1 holds in the experiments.

Beyond the discussion above, we also note that the bijectivity constraint of baseline normalizing flows can lead to problems when modelling a density that is concentrated on a region with complicated topological structure (Cornish et al., 2020, Corollary 2.2), and may cause flows to become numerically non-invertible in this case (Behrmann et al., 2020). Many models such as neural spline flows (NSFs) (Durkan et al., 2019) and *universal* flows (Huang et al., 2018; Jaini et al., 2019) have been proposed to improve expressiveness within the standard framework based on a single bijection. CIFs, on the other hand, use auxiliary variables to provide a mechanism for circumventing the limitations of using a single bijection, but lose analytical tractability as a result.

3.2 IMPLICIT VI METHODS

Several other AVI methods exist that, like our approach, also require the specification of parametrized auxiliary inference distribution $r_{U|Z}$. Hierarchical variational models (HVMs) (Ranganath et al., 2016) are one such example, which take $q_{Z,U}(z, u) \coloneqq q_{Z|U}(z \mid u) \cdot q_U(u)$ for parametrized distributions $q_{Z|U}$ and q_U both analytically tractable. Although both CIFs and HVMs specify tractable $q_{Z,U}$, the CIF joint distribution (5) does not admit such a simple factorization, which may therefore increase expressiveness. Furthermore, unlike CIFs, HVMs do not admit a natural mechanism for matching the auxiliary inference model $r_{U|Z}$ to the structure of the true auxiliary posterior $q_{U|Z}$ when considering multiple levels of hierarchy.

Related to these are approaches are Hamiltonian-based VI methods (Salimans et al., 2015; Caterini et al., 2018), which build $q_{Z,U}$ by numerically integrating Hamiltonian dynamics, inducing a flow that is bijective now on the extended space $\mathcal{Z} \times \mathcal{U}$ instead of just $\mathcal{Z}$. In contrast, CIFs can be used to augment any type of normalizing flow (not just Hamiltonian dynamics), and are not restricted to a specific family of bijections G. Hamiltonian methods also suffer from greatly increasing computational requirements as the number of parameters in $p_{X,Z}$ grows, since they require $\mathrm{D}_z \log p_{X,Z}(x, z)$ at every flow step.

There also exist methods which that do not parametrize $r_{U|Z}$, but instead build an auxiliary inference distribution in VI by drawing extra samples from the approximate posterior q_Z and re-weighting (as noted in Lawson et al. (2019)). These methods, including the importance-weighted autoencoder (IWAE) (Burda et al., 2016) and semi-implicit variational inference (Yin and Zhou, 2018), effectively perform inference over an extended space consisting of K copies of the original latent space (Domke and Sheldon, 2018). These approaches may thus require far more memory to train than parametrized AVI methods, and often require care to ensure the variance of estimators of the objective (and gradients thereof) is controlled (Rainforth et al., 2018b; Tucker et al., 2019). That being said, it may be possible to combine multi-sample bounds with CIF models using a framework such as the one in Sobolev and Vetrov (2019), which demonstrates how to use IWAE-like approaches within HVMs.

A separate class of implicit VI models proposes expressive but intractable joint densities requiring density ratio estimation to train (Huszár, 2017; Tran et al., 2017). CIFs, along with other AVI methods, avoid density ratio estimation by instead constructing a tractable joint density $q_{Z,U}$.

3.3 CIFS FOR DENSITY ESTIMATION

As mentioned earlier, CIFs were originally proposed as a model for density estimation (DE), a setting in which we have access to a set of observed data $\{x_i\}_i$ over which we would like to build a density model p_X maximizing the marginal likelihood. This constitutes the key distinction between this work and Cornish et al. (2020): here, we only use CIFs for parametrizing an *inference* model q_Z, assuming we *already* have access to a forward density model $p_{X,Z}$.

However, the inference procedure required to *train* CIFs for DE is actually very closely related to the model (4). In particular, if we relabel the forward CIF model for DE as r

(instead of p used by Cornish et al. (2020)), the single-layer CIF density estimation objective is equivalent to

$$\mathbb{E}_{(x,u)\sim q_{X,U}}\left[\log \frac{r_Z(G(x;u)) \cdot r_{U|Z}(u \mid G(x;u))}{q_X^*(x) \cdot q_{U|X}(u \mid x) \cdot |\det \mathrm{D}_x G(x;u)|^{-1}}\right], \tag{12}$$

where q_X^* is the unknown data-generating distribution from which we have i.i.d. samples, and $q_{X,U}(x,u) \coloneqq q_X^*(x) \cdot q_{U|X}(u \mid x)$. See Appendix D for a derivation. Comparing this with (9), we see that CIFs for density estimation may be interpreted as performing AVI targeting r_Z with an amortized inference model defined as the Z-marginal of

$$X \sim q_X^*, \quad U \sim q_{U|X}(\cdot \mid X), \quad Z = G(X;U). \tag{13}$$

Furthermore, despite the aesthetic similarities between (7) of Cornish et al. (2020) defining p for DE, and (4) here defining q for VI, it is actually the q models that share a natural correspondence with each other. In both cases, q refers to an inference model that must be reparametrized, whereas neither p in DE nor r here require this. We might even consider using a CIF as the inference distribution for a CIF density model, which may yield additional benefits from added compositionality, although we leave these considerations as future work.

4 EXPERIMENTS

In this section, we investigate using CIFs to build more expressive variational models in posterior sampling and maximum likelihood estimation of generative models. We compare inference models based on the Masked Autoregressive Flow (MAF) (Papamakarios et al., 2017) and the autoregressive variant of the Neural Spline Flow (NSF) (Durkan et al., 2019) to CIF-based extensions. Both of these baseline models empirically provide good performance in general-purpose density estimation. We use the ADAM optimizer (Kingma and Ba, 2015) throughout. Hyperparameters for all experiments are available in Appendix E. Code will be made available at `https://github.com/anthonycaterini/cif-vi`.

4.1 TOY MIXTURE OF GAUSSIANS

Our first example looks at using VI to sample from a toy mixture of Gaussians. Given component means $\{\mu_k\}_k$ and covariances $\{\Sigma_k\}_k$, we directly define the "posterior"[2] $p_{Z|X}(z \mid x) \coloneqq \sum_{k=1}^K \mathcal{N}(z; \mu_k, \Sigma_k)/K$, where K is the total number of components, so that the joint target is $p_{X,Z}(x,z) \propto p_{Z|X}(z \mid x)$. We work in two dimensions with component means adequately spaced out in a square lattice. Although the support of $p_{Z|X}$ is all of $\mathbb{R}^2$, it is concentrated on a subset of K disconnected components, which is not homeomorphic to $\mathbb{R}^2$, and thus we anticipate difficulties in using just a normalizing flow as the approximate posterior. We compare baseline NSF models to CIF-based extensions.

The initial distribution for both the NSF and CIF models is given by $q_W \coloneqq \mathcal{N}(0, \sigma_0^2 \mathbf{I})$, with σ_0 taken as either a fixed hyperparameter or a trainable variational parameter. The CIF extension includes an auxiliary variable $u \in \mathbb{R}$ at each layer, conditional Gaussian distributions for $q_{U_\ell|W\ell-1}$ and $r_{U_\ell|W_\ell}$ parametrized by small neural networks, and a single small two-headed neural network to output s and t in (6) at each layer, adding only 8.5% more parameters on top of the baseline NSF model.

Marginal ELBO Estimator For all experiments in this section, we will measure the trained models on estimates of the marginal ELBO (1). When using an explicit variational method, such as an NSF, this is readily estimated by basic Monte Carlo (MC) with N i.i.d. samples $z^{(i)} \sim q_Z$ for $i \in \{1, \dots, N\}$:

$$\widehat{\mathcal{L}}(x) \coloneqq \frac{1}{N} \sum_{i=1}^N \log \frac{p_{X,Z}(x, z^{(i)})}{q_Z(z^{(i)})}. \tag{14}$$

However, recall that in implicit methods q_Z is not available in closed form, which precludes direct evaluation of (14). Thus, we first must build an estimator of $q_Z(z)$ for all $z \in \mathcal{Z}$ to use within (14). We can do this via importance sampling, taking M i.i.d. samples $u^{(j)} \sim r_{U|Z}(\cdot \mid z)$ for $j \in \{1, \dots, M\}$ from our trained auxiliary inference model:

$$q_Z(z) \approx \frac{1}{M} \sum_{j=1}^M \frac{q_{Z,U}(z, u^{(j)})}{r_{U|Z}(u^{(j)} \mid z)} =: \widehat{q}_Z(z). \tag{15}$$

The full estimator of the marginal ELBO for auxiliary models is then obtained by substituting (15) into (14); this is written out in full in Appendix F. Although this estimator is positively biased (because it includes the negative logarithm of an unbiased MC estimator), it is still *consistent*, and its bias is naturally controlled by the training procedure which encourages $r_{U|Z}$ to match the intractable $q_{U|Z}$. We can mitigate any further bias by increasing M (Rainforth et al., 2018a). A table displaying estimates of the marginal ELBO on a single trained model for various choices of N and M is also available in Appendix F; we choose $N = 10{,}000$ and $M = 100$ based on these results.

Results For our first experiment, we select $K = 9$ and fix σ_0 throughout training to either $0.1, 1$, or 10. We train both NSF baselines and CIF-NSF extensions with three different random seeds for each setting of σ_0. We show a kernel density estimation of the approximate posterior of the average case model on each configuration in Figure 2 and report the average of the marginal ELBO estimates across all three runs in the titles of the plots. We can clearly see

[2]Note that there is no data x in this example – we define the "posterior" directly. Details are in Appendix E.

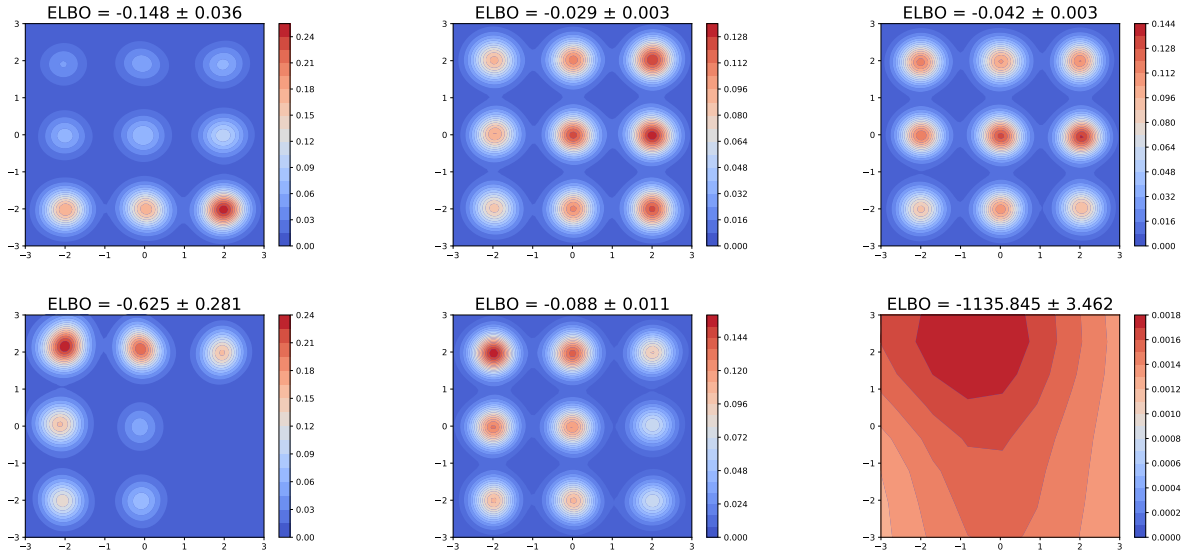


Figure 2: Samples from the trained inference models visualized using a KDE plot for a range of σ_0 values. We ran each configuration 3 times, displaying the average case of the three runs in the image, with the average plus/minus standard error of the marginal ELBO across the three runs shown in the title of the plot (higher is better). Models in the top row are CIF-NSFs, and those in the bottom row are baseline NSFs. We can see that when $\sigma_0 = 0.1$, the NSF does not have enough initial noise to consistently cover the target, and when $\sigma_0 = 10$, the NSF has too much noise and cannot locate the target. The CIF-NSF at least locates each mode in all cases and provides higher-quality approximations across the board.

from both the ELBO values and the plots themselves that the CIF extensions are more consistently producing higher-quality variational approximations across the range of σ_0, as form of (6) allows the model to directly control the noise of the outputted samples. The NSF baselines only produce reliable models for $\sigma_0 = 1$.

In this example it is quite clear how the parametrization of the CIF model "cleans up" a major deficiency of the baseline method by rescaling the initial noise. However, we might also allow σ_0 to be learned as part of the overall variational inference procedure to further probe the effectiveness of CIFs, and we experiment with this on a more challenging problem ($K = 16$). We find that the trained CIF models again outperform the baseline NSFs (estimated marginal ELBO over 3 runs of $\mathbf{-0.116 \pm 0.021}$ for CIFs vs. -0.562 ± 0.008 for basleine NSFs), thus demonstrating the increased expressiveness of CIFs beyond just rescaling.

4.2 GENERATIVE MODELLING OF IMAGES

For our second example, we use amortized variational inference to facilitate the training of a generative model of image data in the style of the variational auto-encoder (VAE) method (Kingma and Welling, 2014). We attempt to build models of the MNIST (LeCun et al., 1998) and Fashion-MNIST (Xiao et al., 2017) datasets, which both contain 256-bit greyscale images of size 28×28. We employ dynamic binarization of these greyscale images at each training step. The likelihood function to describe an image relies on a neural network "decoder" $\pi : \mathcal{Z} \to [0,1]^d$, such that

$$Z \sim \mathcal{N}(0, \mathbf{I}), \quad X \sim \bigotimes_{j=1}^{d} \text{Ber}(\cdot \mid \pi_j(Z))$$

is the generative process for an image X. In our experiments, we consider two different types of decoders: a small convolutional network with only one hidden layer, and a larger convolutional network with several residual blocks as in e.g. Durkan et al. (2019). For the experiments with the smaller decoder, we use a 20-dimensional latent space $\mathcal{Z}$, and for the larger decoder, we increase to 32 dimensions.

Inference Methods We consider several models of inference to aid in surrogate maximum likelihood estimation of the parameters of π. First we consider a VAE inference model, where

$$q_{Z|X}(\cdot \mid x) \coloneqq \mathcal{N}\left(\mu_Z(x), \text{diag}\, \sigma_Z^2(x)\right)$$

with an "encoder" neural network taking in image data x and outputting both μ_Z and $\log \sigma_Z$. The encoder that we use in all experiments is a single-hidden-layer convolutional network which "matches" the structure of the small decoder; we keep the encoder small since the VAE here is just a base upon which we build more complicated inference models. We also consider an importance-weighted version of

Table 1: Test-set average marginal log-likelihood (plus/minus one standard error) over three runs. Runs that are within one standard error of the best-performing model are shown in bold.

Model	Small Target		Large Target	
	MNIST	Fashion-MNIST	MNIST	Fashion-MNIST
VAE	-94.83 ± 0.05	-238.54 ± 0.11	-86.27 ± 0.04	-229.72 ± 0.03
IWAE ($K = 5$)	-93.14 ± 0.10	-237.03 ± 0.05	-84.23 ± 0.09	-227.80 ± 0.02
Small MAF	-91.98 ± 0.19	-237.09 ± 0.15	-83.41 ± 0.09	-228.74 ± 0.24
Large MAF	-92.68 ± 0.26	-237.57 ± 0.03	-83.38 ± 0.12	-228.72 ± 0.27
CIF-MAF	-90.87 ± 0.05	$\mathbf{-236.31 \pm 0.14}$	$\mathbf{-82.70 \pm 0.12}$	$\mathbf{-227.64 \pm 0.05}$
Small NSF	-91.12 ± 0.15	-236.65 ± 0.17	-83.06 ± 0.05	-228.58 ± 0.18
Large NSF	$\mathbf{-90.79 \pm 0.02}$	$\mathbf{-236.48 \pm 0.13}$	-83.12 ± 0.10	-228.46 ± 0.07
CIF-NSF	$\mathbf{-90.82 \pm 0.09}$	$\mathbf{-236.48 \pm 0.20}$	-83.31 ± 0.17	-228.54 ± 0.12

this VAE model (IWAE) with $K = 5$ importance samples (Burda et al., 2016), which we find roughly matches the computation time per epoch of the flow-based inference methods below.

The first flow-based model that we consider is a 5-layer masked autoregressive flow (MAF) (Papamakarios et al., 2017), which is equivalent to an inverse autoregressive flow (IAF) (Kingma et al., 2016) when removing the hypernetworks producing the flow parameters. We also run experiments with a 10-layer neural spline flow (NSF) (Durkan et al., 2019), for which we clip the norm of the gradients to a maximum of 5 – as suggested for tabular density estimation – for increased stability of training. Additional hyperparameter settings for each flow are available in Appendix E. As alluded to previously, for each of the flow-based methods we will use the small VAE encoder as a base distribution $q_{W_0|X}$ to project the image data into the dimension of the latent space; we do this rather than using a large VAE encoder as the base distribution in the large target experiments (as is typically done) to force the flow models to handle more of the inference. We also consider two baseline variants for each model, a larger and smaller version, which we control by changing the number of hidden channels in the autoregressive maps.

Finally, we consider amortized CIF-based extensions of the *smaller* variants of the flow models mentioned above, so that in the end our CIF models have approximately the same total number of parameters as the larger baseline flows. We use a 2-dimensional u at each flow step. We include parametrized conditional Gaussian distributions for $q_{U_\ell|W_{\ell-1}}$ and $r_{U_\ell|W_\ell,X}$ at each layer $\ell \in \{1, \ldots, L\}$, with additional care taken in the structure of the r network to combine vector inputs W_ℓ with image inputs X – details are provided in Appendix E.3.3. We use a single neural network at each layer to parametrize s_ℓ and t_ℓ appearing in G_ℓ.

Results The results of the experiment are available in Table 1. We use the standard importance-sampling based estimator of the marginal likelihood from Rezende et al. (2014, Appendix E) with 1,000 samples, which we find empirically produces low-variance estimates for the small target model[3] as noted in Appendix E.5. We see that, in each experiment, CIF models are either producing the best average performance as measured by test-set estimated average marginal likelihood, or are within error bars of the best. Importantly, we note that CIFs are outperforming the baseline models which they are built directly on top of across the board: CIF-MAF and CIF-NSF significantly improve upon Small MAF and Small NSF, respectively. This justifies the claims of Proposition 3.1, demonstrating that we are not penalized for using the auxiliary objective instead of the standard ELBO.

We also can see that the CIF models produce better results than the IWAE models, which can themselves be seen as a method for auxiliary VI as previously mentioned. Despite IWAE methods being more parameter-efficient, we found that increasing K for IWAE significantly increased training time per epoch over the CIF models.

5 CONCLUSION AND DISCUSSION

In this work, we have presented continuously-indexed flows (CIFs) as a novel parametrization of an approximate posterior for use within variational inference (VI). We did this by naturally incorporating the CIF model into the framework of AVI. We have shown that the theoretical and empirical benefits of CIFs over baseline flow models extend to the VI setting, as CIFs outperform baseline flows in both sampling from complicated target distributions and facilitating maximum likelihood estimation of parametrized latent-variable models. We now add a brief further discussion on CIFs in

[3] We expect the same low-variance behaviour to translate to the larger target model, but did not run this for computational reasons.

VI and consider some directions for future work.

Modelling Discrete Distributions One issue with CIFs for VI (indeed, CIFs more generally) is that they are currently only designed to model continuous distributions, unlike e.g. HVMs. It may be possible however to alleviate this constraint by using discrete flows (Hoogeboom et al., 2019) as a component of the overall CIF model, although it remains to be seen if the theoretical and empirical benefits of CIFs over baseline flows would extend to this case.

CIFs in Other Applications This work can serve as a template for applying CIFs more generally in applications where NFs have proven effective, such as compression (Ho et al., 2019) and approximate Bayesian computation (Papamakarios et al., 2019). These approaches may require the formulation of appropriate, application-specific surrogate objectives, but the expressiveness gains could overcome the additional costs (as in VI and density estimation) and could therefore be investigated.

Acknowledgements

Anthony Caterini is a Commonwealth Scholar supported by the U.K. Government. Rob Cornish is supported by the Engineering and Physical Sciences Research Council (EPSRC) through the Bayes4Health programme Grant EP/R018561/1. Arnaud Doucet is supported by the EPSRC CoSInES (COmputational Statistical INference for Engineering and Security) grant EP/R034710/1

References

Felix V Agakov and David Barber. An auxiliary variational method. In *International Conference on Neural Information Processing*, pages 561–566. Springer, 2004.

Jens Behrmann, Paul Vicol, Kuan-Chieh Wang, Roger B. Grosse, and Jörn-Henrik Jacobsen. On the invertibility of invertible neural networks, 2020. URL `https://openreview.net/forum?id=BJlVeyHFwH`.

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

Yuri Burda, Roger B. Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. In *4th International Conference on Learning Representations*, 2016.

Anthony L Caterini, Arnaud Doucet, and Dino Sejdinovic. Hamiltonian variational auto-encoder. In *Advances in Neural Information Processing Systems*, pages 8167–8177, 2018.

Rob Cornish, Anthony L Caterini, George Deligiannidis, and Arnaud Doucet. Relaxing bijectivity constraints with continuously-indexed normalising flows. In *International Conference on Machine Learning*, 2020.

Justin Domke and Daniel R Sheldon. Importance weighting and variational inference. In *Advances in Neural Information Processing Systems*, pages 4470–4479, 2018.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. In *Advances in Neural Information Processing Systems*, pages 7509–7520, 2019.

Jonathan Ho, Evan Lohn, and Pieter Abbeel. Compression with flows via local bits-back coding. In *Advances in Neural Information Processing Systems*, pages 3874–3883, 2019.

Emiel Hoogeboom, Jorn Peters, Rianne van den Berg, and Max Welling. Integer discrete flows and lossless compression. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Chin-Wei Huang, David Krueger, Alexandre Lacoste, and Aaron Courville. Neural autoregressive flows. In *International Conference on Machine Learning*, pages 2078–2087, 2018.

Ferenc Huszár. Variational inference using implicit distributions. *arXiv preprint arXiv:1702.08235*, 2017.

Priyank Jaini, Kira A Selby, and Yaoliang Yu. Sum-of-squares polynomial flow. In *International Conference on Machine Learning*, pages 3009–3018, 2019.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR*, 2015.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *2nd International Conference on Learning Representations*, 2014.

Diederik P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems*, pages 4743–4751, 2016.

Steven Kleinegesse, Christopher Drovandi, and Michael U Gutmann. Sequential Bayesian experimental design for implicit models via mutual information. *arXiv preprint arXiv:2003.09379*, 2020.

John Lawson, George Tucker, Bo Dai, and Rajesh Ranganath. Energy-inspired models: Learning with sampler-induced distributions. In *Advances in Neural Information Processing Systems*, pages 8499–8511, 2019.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Christos Louizos and Max Welling. Multiplicative normalizing flows for variational bayesian neural networks. In *International Conference on Machine Learning*, pages 2218–2227. PMLR, 2017.

George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems*, pages 2338–2347, 2017.

George Papamakarios, David Sterratt, and Iain Murray. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 837–848, 2019.

Tom Rainforth, Rob Cornish, Hongseok Yang, Andrew Warrington, and Frank Wood. On nesting Monte Carlo estimators. In *International Conference on Machine Learning*, pages 4267–4276. PMLR, 2018a.

Tom Rainforth, Adam Kosiorek, Tuan Anh Le, Chris Maddison, Maximilian Igl, Frank Wood, and Yee Whye Teh. Tighter variational bounds are not necessarily better. In *International Conference on Machine Learning*, pages 4277–4285, 2018b.

Rajesh Ranganath, Dustin Tran, and David Blei. Hierarchical variational models. In *International Conference on Machine Learning*, pages 324–333, 2016.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*, pages 1530–1538, 2015.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pages 1278–1286, 2014.

Tim Salimans, Diederik P. Kingma, and Max Welling. Markov chain Monte Carlo and variational inference: Bridging the gap. In *International Conference on Machine Learning*, pages 1218–1226, 2015.

Artem Sobolev and Dmitry P Vetrov. Importance weighted hierarchical variational inference. In *Advances in Neural Information Processing Systems*, pages 603–615, 2019.

Esteban G Tabak, Eric Vanden-Eijnden, et al. Density estimation by dual ascent of the log-likelihood. *Communications in Mathematical Sciences*, 8(1):217–233, 2010.

Dustin Tran, Rajesh Ranganath, and David Blei. Hierarchical implicit models and likelihood-free variational inference. In *Advances in Neural Information Processing Systems*, pages 5523–5533, 2017.

George Tucker, Dieterich Lawson, Shixiang Gu, and Chris J. Maddison. Doubly reparameterized gradient estimators for Monte Carlo objectives. In *7th International Conference on Learning Representations, ICLR*, 2019.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Mingzhang Yin and Mingyuan Zhou. Semi-implicit variational inference. In *International Conference on Machine Learning*, pages 5660–5669, 2018.