Review and Critical Analysis

Variational Inference with Normalizing Flows

Rezende, D. J., & Mohamed, S. (ICML 2015)

https://arxiv.org/abs/1505.05770

1 Summary

This paper addresses a fundamental bottleneck in Variational Inference (VI): the expressiveness of the approximate posterior distribution. Standard VI often relies on the mean-field assumption (fully factorized Gaussians) for computational tractability, which severely limits the model's ability to capture complex, multi-modal, or correlated posteriors. Rezende and Mohamed introduce Normalizing Flows (NF) into the VI framework, a method to construct arbitrarily complex posterior distributions by transforming a simple base density through a sequence of invertible mappings. The authors derive specific parameterized transformations—Planar and Radial flows—that allow for the linear-time computation of Jacobian determinants, enabling scalable, amortized inference in Deep Latent Gaussian Models (DLGMs).

2 Theoretical Framework

2.1 The Normalizing Flow Mechanism

The core methodology relies on the change of variables theorem. If we transform a random variable $\mathbf{z}_0$ with distribution $q_0(\mathbf{z}_0)$ through a chain of K invertible mappings f_k, the log-density of the final variable $\mathbf{z}_K$ is given by:

$$\ln q_K(\mathbf{z}_K) = \ln q_0(\mathbf{z}_0) - \sum_{k=1}^{K} \ln \left| \det \frac{\partial f_k}{\partial \mathbf{z}_{k-1}} \right| \tag{1}$$

In general, computing the determinant of the Jacobian $\mathbf{J} = \partial f / \partial \mathbf{z}$ has a computational cost of $\mathcal{O}(D^3)$, which is prohibitive for high-dimensional latent spaces.

2.2 Efficient Transformations: Planar and Radial Flows

To overcome the cubic complexity of the determinant, the authors introduce specific families of transformations where the determinant can be computed in $\mathcal{O}(D)$ time.

- **Planar Flows:** Defined as $f(\mathbf{z}) = \mathbf{z} + \mathbf{u}h(\mathbf{w}^T\mathbf{z} + b)$. By applying the Matrix Determinant Lemma, the determinant of the Jacobian simplifies to $|1 + \mathbf{u}^T\psi(\mathbf{z})\mathbf{w}|$, which requires only simple dot products. This effectively acts as a contraction or expansion of the density along the direction perpendicular to the hyperplane defined by $\mathbf{w}$.
- **Radial Flows:** Deform the space around a specific reference point, allowing for local radial expansions and contractions (useful for multimodal distributions).

2.3 Flow-Based Free Energy Bound

The authors modify the Evidence Lower Bound (ELBO) to account for the flow. The new objective function, which they term the flow-based free energy, explicitly includes the Jacobian terms:

$$\mathcal{F}(\mathbf{x}) = \mathbb{E}_{q_0(\mathbf{z}_0)}\left[\ln p(\mathbf{x}, \mathbf{z}_K) - \ln q_0(\mathbf{z}_0) + \sum_{k=1}^{K} \ln \left|\det \frac{\partial f_k}{\partial \mathbf{z}_{k-1}}\right|\right] \tag{2}$$

This formulation allows the variational parameters (the flow weights) to be optimized jointly with the model parameters via stochastic gradient descent (SGD), fitting seamlessly into the VAE training pipeline.

3 Empirical Evaluation

The paper provides a comprehensive evaluation of the representational capacity of flows:

1. **2D Toy Potentials:** Visualizations demonstrate that starting from a standard Gaussian, normalizing flows can morph the density into complex, non-Gaussian shapes (e.g., spirals, bimodal distributions) that mean-field approximations fundamentally cannot capture.
2. **Deep Latent Gaussian Models (DLGMs):** Applied to MNIST and CIFAR-10, the authors show that replacing a standard diagonal Gaussian posterior with a Planar Flow posterior significantly improves the variational lower bound and test-set log-likelihood.
3. **Comparison to NICE:** The method is compared to volume-preserving flows (NICE). The results indicate that Planar flows, which are non-volume preserving (allowing compression/expansion of density), achieve better performance with fewer parameters.

4 Critical Analysis

4.1 Strengths

- **Breaking the Mean-Field Assumption:** This is the definitive strength of the paper. It provides a practical, scalable way to move beyond the "Gaussian assumption" in VAEs without resorting to expensive MCMC methods.
- **Unified Perspective:** The authors skillfully connect their discrete flows to continuous-time dynamics, framing Langevin dynamics and Hamiltonian Monte Carlo (HMC) as specific instances of infinitesimal flows. This theoretical bridging enriches the understanding of both sampling and variational methods.
- **Computational Efficiency:** The reliance on the Matrix Determinant Lemma ensures that the added complexity is linear in dimensionality, making the method viable for the high-dimensional latent spaces common in deep learning.

4.2 Limitations and Weaknesses

- **Topological Constraints:** While Planar and Radial flows are efficient, they are topologically limited. A Planar flow applies a global slicing operation. Transforming a single mode into clearly separated, disjoint modes (e.g., the "checkerboard" pattern) requires a very long composition of flows ($K \gg 10$), which can lead to vanishing gradients during training.

- **High-Dimensional Scaling:** The paper primarily evaluates on relatively small images (MNIST/CIFAR). In very high dimensions, the "contraction" behavior of Planar flows might be insufficient to model complex dependencies compared to later architectures like Autoregressive Flows (IAF/MAF) or Coupling Layers (RealNVP).
- **Parameter Efficiency:** While $O(D)$ is efficient per layer, constructing a highly flexible posterior might require a deep flow (K large), significantly increasing the number of parameters to optimize compared to a simple Gaussian encoder.

5 Conclusion

Rezende and Mohamed (2015) presents a foundational contribution to probabilistic deep learning. By making the approximate posterior flexible and learnable via the normalizing flow framework, they resolved a major criticism of Variational Inference—the bias introduced by simple posterior approximations. This work laid the groundwork for a rich sub-field of research into invertible neural networks and more advanced flow architectures.