Iterative Inference Models

Joseph Marino[1], Yisong Yue[1], Stephan Mandt[2]
[1]California Institute of Technology, [2]Disney Research

abstract

We introduce ***iterative inference models*** for deep latent variable models, which learn to *infer the approximate posterior of a generative model by iteratively encoding approximate posterior gradients*. Our contributions are:

- Generalization of inference models to iterative inference.
- Demonstration that neural networks can approximate inference updates.
- Theoretical justification for "top-down" inference in latent variable models.

background

Set-up: *Variational Inference in Deep Latent Variable Models*

Observations $\mathbf{x}$, Latent Variables $\mathbf{z}$
Latent Variable Model $p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z}) p_\theta(\mathbf{z})$
Model Parameters θ
Approximate Posterior $q(\mathbf{z}|\mathbf{x})$
Approx. Posterior Distribution Parameters (e.g. mean) λ
Evidence Lower Bound (ELBO) $\mathcal{L} \leq \log p_\theta(\mathbf{x})$

Variational EM [1] alternates between directly maximizing $\mathcal{L}$ w.r.t. the approx. posterior parameters λ and model parameters θ. λ is updated as:

$$\lambda_t = \lambda_{t-1} + \alpha \nabla_\lambda \mathcal{L}$$

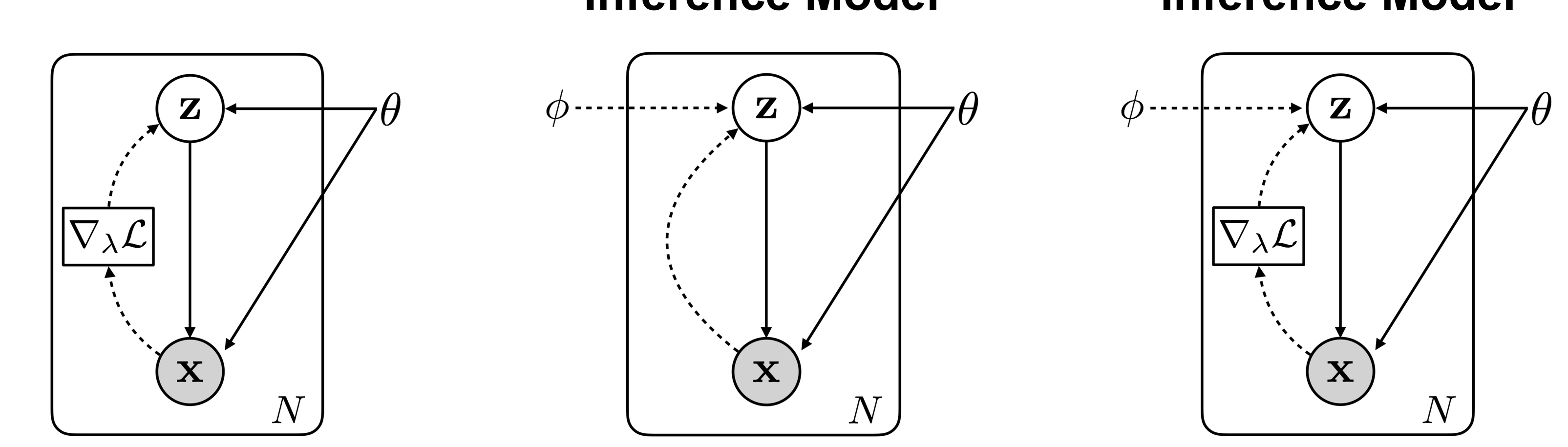
Standard Inference Models, as used in VAEs [2, 3], are separate models, with amortized parameters ϕ, that output estimates of λ. The parameters ϕ and θ are updated jointly. λ is given as:

$$\lambda = f_\phi(\mathbf{x})$$

Iterative Inference Models are inference models that update the approx. posterior parameters using stochastic estimates of the approx. posterior gradient. Again, ϕ and θ are updated jointly. λ is updated as:

$$\lambda_t = f_\phi(\lambda_{t-1}, \nabla_\lambda \mathcal{L})$$

Variational EM | **Standard Inference Model** | **Iterative Inference Model**



For latent Gaussian models, where $\lambda \equiv \{\mu_q, \sigma_q^2\}$, the approximate posterior gradients take the following form:

$$\nabla_{\mu_q} \mathcal{L} = \mathbb{E}_{\mathcal{N}(\epsilon;\mathbf{0},\mathbf{I})} \left[\frac{\partial \mu_\mathbf{x}}{\partial \mu_q}^\intercal \frac{\mathbf{x} - \mu_\mathbf{x}}{\sigma_\mathbf{x}^2} - \frac{\mu_q + \sigma_q \odot \epsilon - \mu_p}{\sigma_p^2} \right]$$

$$\nabla_{\sigma_q^2} \mathcal{L} = \mathbb{E}_{\mathcal{N}(\epsilon;\mathbf{0},\mathbf{I})} \left[\frac{\partial \mu_\mathbf{x}}{\partial \sigma_q^2}^\intercal \frac{\mathbf{x} - \mu_\mathbf{x}}{\sigma_\mathbf{x}^2} - \left(\operatorname{diag} \frac{\epsilon}{2\sigma_q} \right)^\intercal \frac{\mu_q + \sigma_q \odot \epsilon - \mu_p}{\sigma_p^2} \right] - \frac{\mathbf{1}}{2\sigma_q^2}$$

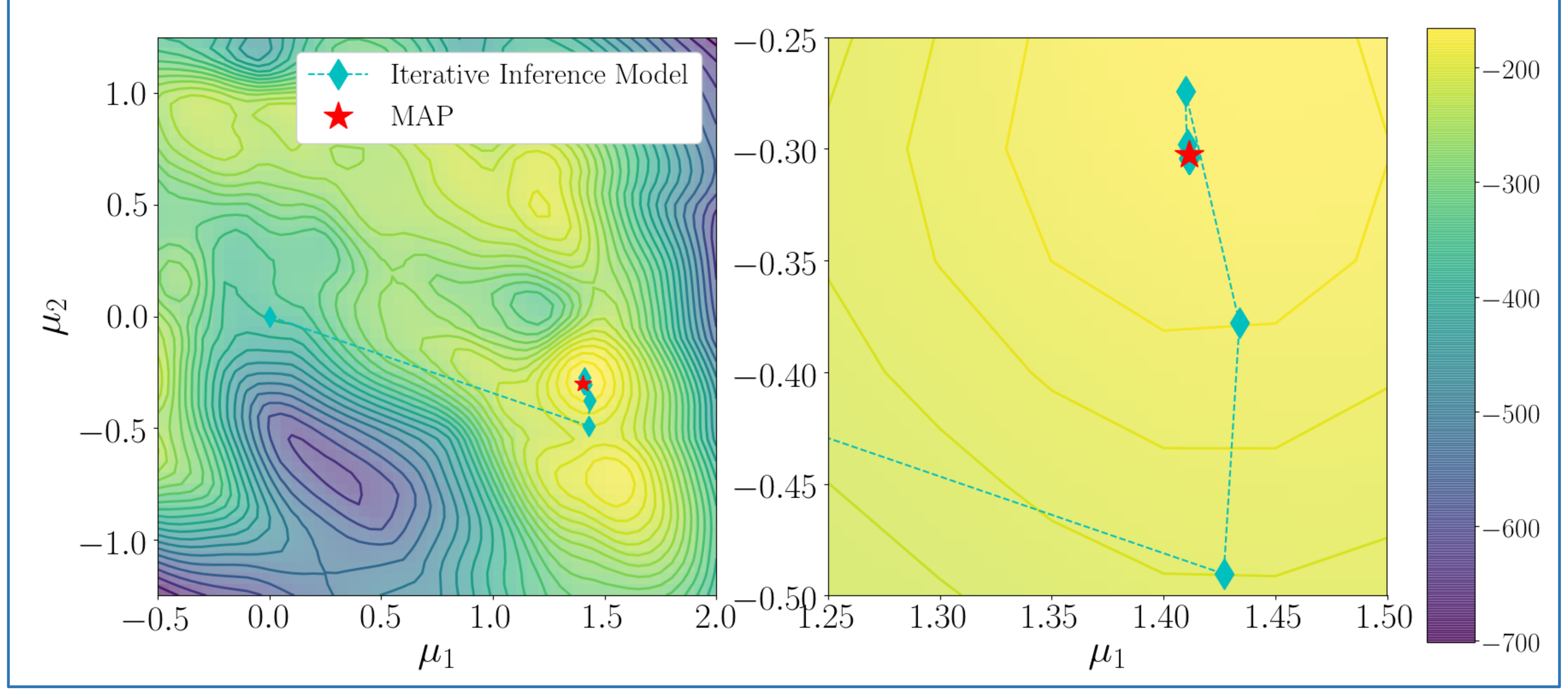
These consist of **bottom-up errors** $\varepsilon_\mathbf{x} \equiv (\mathbf{x} - \mu_\mathbf{x}) / \sigma_\mathbf{x}^2$ and **top-down errors** $\varepsilon_\mathbf{z} \equiv (\mathbf{z} - \mu_p) / \sigma_p^2$. We can encode these terms instead of gradients, approximating the other terms in the gradients with the model:

$$\lambda_t = f_\phi(\lambda_{t-1}, \varepsilon_{\mathbf{x},t-1}, \varepsilon_{\mathbf{z},t-1})$$

Under certain assumptions, **models of this form generalize standard inference models.** Additionally, the form of these gradients **provides theoretical justification for "top-down" inference** procedures in hierarchical latent variable models [4]. See our paper for further details.

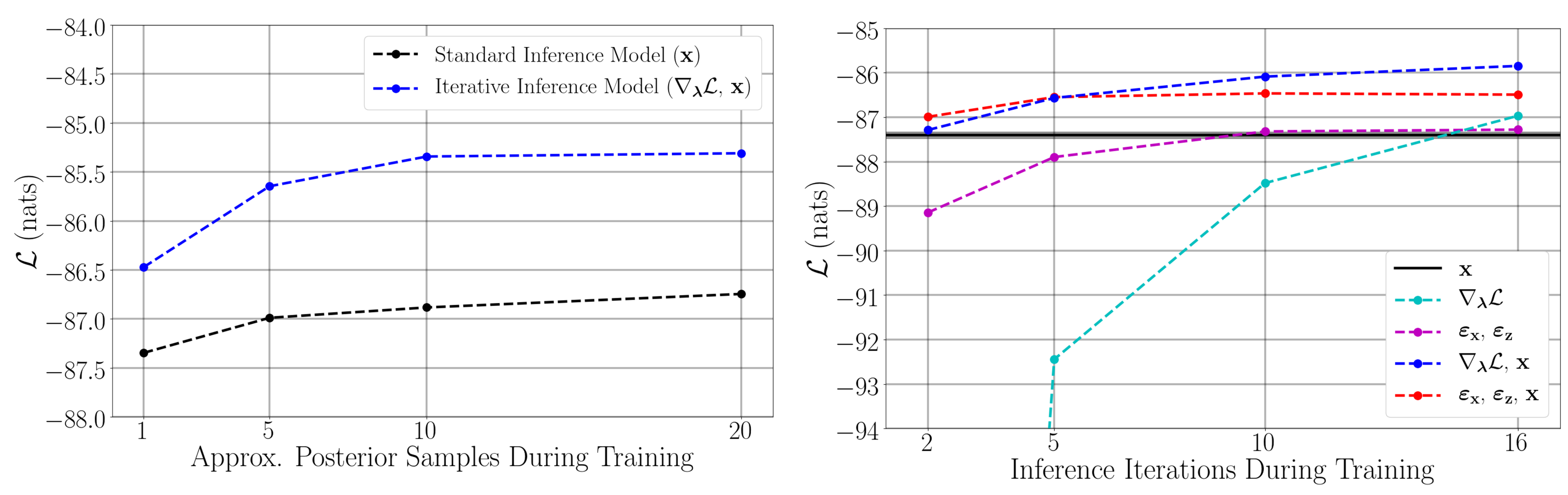
results

We trained 2D latent variable models on MNIST to visualize inference optimization directly. **Iterative inference models adaptively adjust update step sizes to the approximate posterior parameters**. As shown below, the model arrives at a near-optimal estimate in a small number of steps.



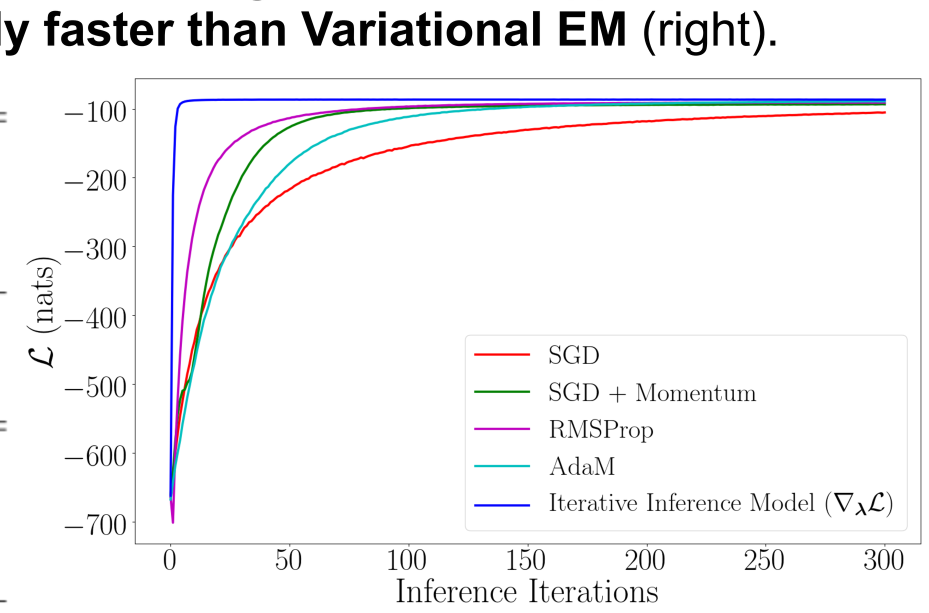
Inference optimization can be visualized through reconstructions. In the following figure, we visualize **reconstructions improving over inference iterations** on MNIST and CIFAR-10. See our paper for additional examples.

Iterative inference models improve significantly with **additional approximate posterior samples and inference iterations**. These provide more precise updates and additional update steps.



Iterative inference models quantitatively outperform standard inference models with *identical architectures* on MNIST *(nats)* and CIFAR-10 *(bits/dim)* (left). Iterative inference models **converge to similar approximate posterior estimates significantly faster than Variational EM** (right).

	$-\log p(\mathbf{x}) \approx$
MNIST	
One-Level Model	
Standard (VAE)	84.14 ± 0.02
Iterative	$\mathbf{83.84 \pm 0.05}$
Hierarchical Model	
Standard (VAE)	82.63 ± 0.01
Iterative	$\mathbf{82.457 \pm 0.001}$
CIFAR-10	
One-Level Model	
Standard (VAE)	5.823 ± 0.001
Iterative	$\mathbf{5.71 \pm 0.02}$



discussion

We have shown that iterative inference models **1)** generalize and outperform standard inference models, **2)** outperform variational EM inference optimization, and **3)** qualitatively learn to perform optimization.

Work remains to be done in extending these models to larger architectures and data sets. Iterative inference models also naturally extend to dynamic latent variable models, where approximate posterior estimates are updated relative to previous estimates.

contact

Joseph Marino
California Institute of Technology
Email: jmarino@caltech.edu
Website: http://joelouismarino.github.io/

references

1. Radford M Neal and Geoffrey E Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. 1998.
2. Diederik P Kingma and Max Welling. Stochastic gradient vb and the variational auto-encoder. 2014.
3. Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. 2014.
4. Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. 2016.