Learning to Infer

Joseph Marino[1], Yisong Yue[1], Stephan Mandt[2]

[1]California Institute of Technology (Caltech), [2]Disney Research

abstract

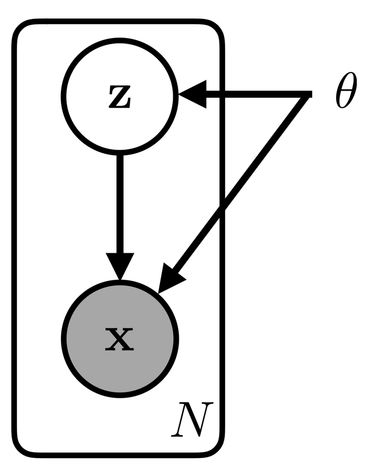
We introduce ***iterative inference models*** for deep latent variable models, which ***learn to infer*** the approximate posterior by *iteratively encoding approximate posterior gradients*.

- Generalize amortized inference models to iterative estimation.
- Theoretical justification for "top-down" inference in hierarchical latent variable models.
- Empirical results on image and text data.

background

Latent Variable Model

$$p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z}) p_\theta(\mathbf{z})$$



Latent Gaussian Model

Prior $p_\theta(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mu_p, \sigma_p^2)$

Conditional Likelihood e.g. $p_\theta(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}; \mu_\mathbf{x}, \sigma_\mathbf{x}^2)$

Variational Inference

Approximate Posterior e.g. $q(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}; \mu_q, \sigma_q^2)$ $\quad \lambda \equiv \{\mu_q, \sigma_q^2\}$

ELBO $\mathcal{L} = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - D_{KL}(q(\mathbf{z}|\mathbf{x}) || p_\theta(\mathbf{z}))$
$\leq \log p_\theta(\mathbf{x})$

<u>**Variational EM Algorithm**</u> [1]:

Variational **E-Step** (*Inference*): $\lambda = \text{argmax}_\lambda \mathcal{L}$
Variational **M-Step** (*Learning*): $\theta = \text{argmax}_\theta \mathcal{L}$

Conventional inference optimization, (e.g. SVI [2]):

$$\lambda = \lambda + \alpha \nabla_\lambda \mathcal{L}$$

Standard Inference Models, (e.g. VAE [3, 4]):

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

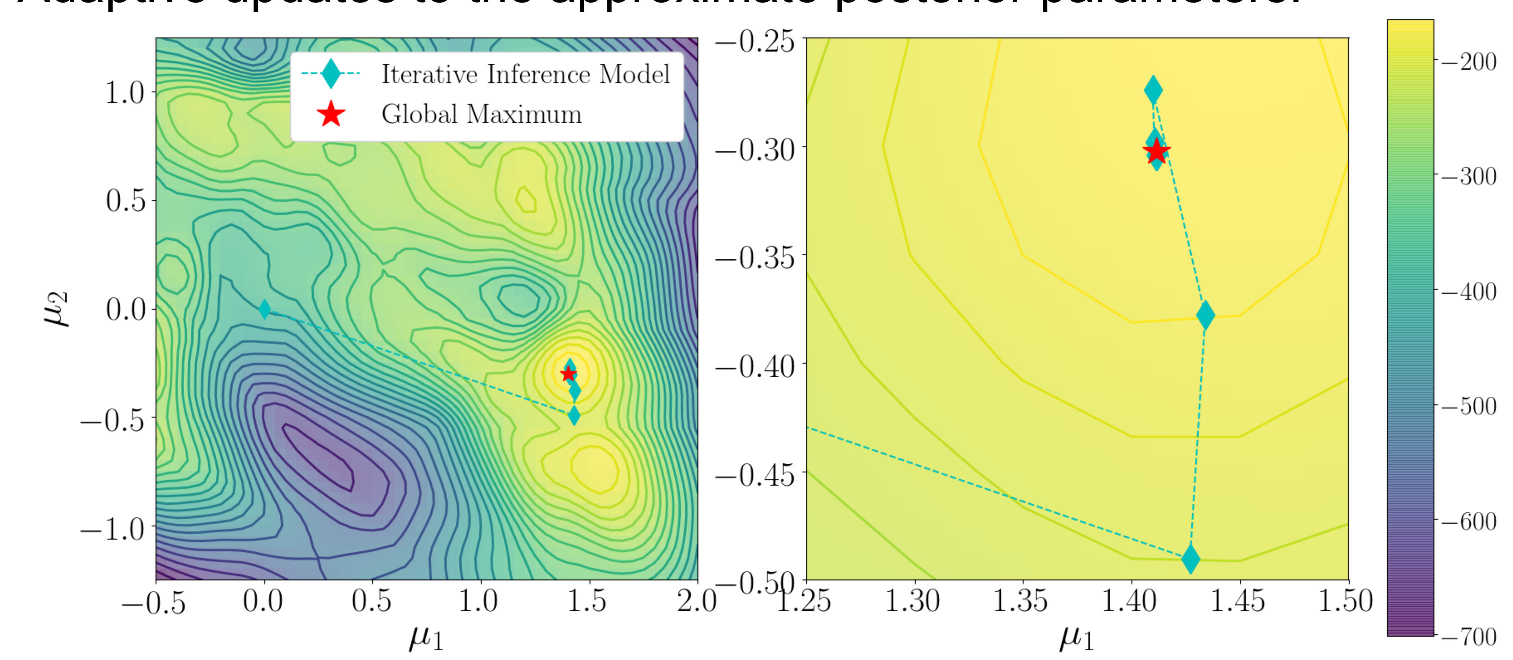
Iterative Inference Models:

$$\lambda = f_\phi(\lambda, \nabla_\lambda \mathcal{L})$$

results

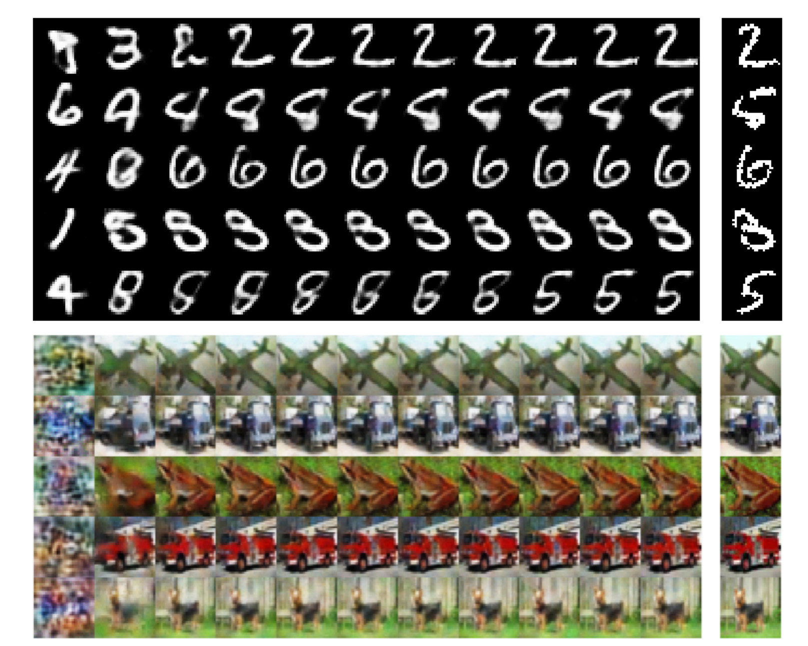
Visualizing Optimization in 2D

Adaptive updates to the approximate posterior parameters.



Reconstructions

Inference Iterations → Data



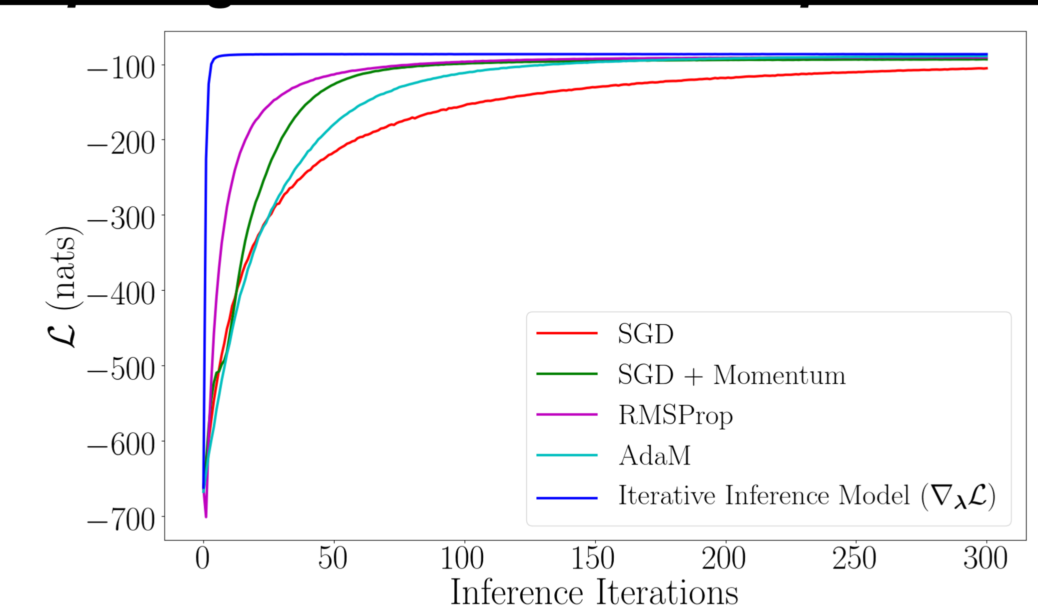
Comparing with Standard Inference Models

	$-\log p(\mathbf{x})$
MNIST	
Single-Level	
Standard	84.14 ± 0.02
Iterative	$\mathbf{83.84 \pm 0.05}$
Hierarchical	
Standard	82.63 ± 0.01
Iterative	$\mathbf{82.457 \pm 0.001}$
CIFAR-10	
Single-Level	
Standard	5.823 ± 0.001
Iterative	$\mathbf{5.64 \pm 0.03}$
Hierarchical	
Standard	5.565 ± 0.002
Iterative	$\mathbf{5.456 \pm 0.005}$

	Perplexity
RCV1	
Standard	323 ± 3
Iterative	$\mathbf{285.0 \pm 0.1}$

Iterative inference models outperform comparable standard inference models across data sets and model architectures.

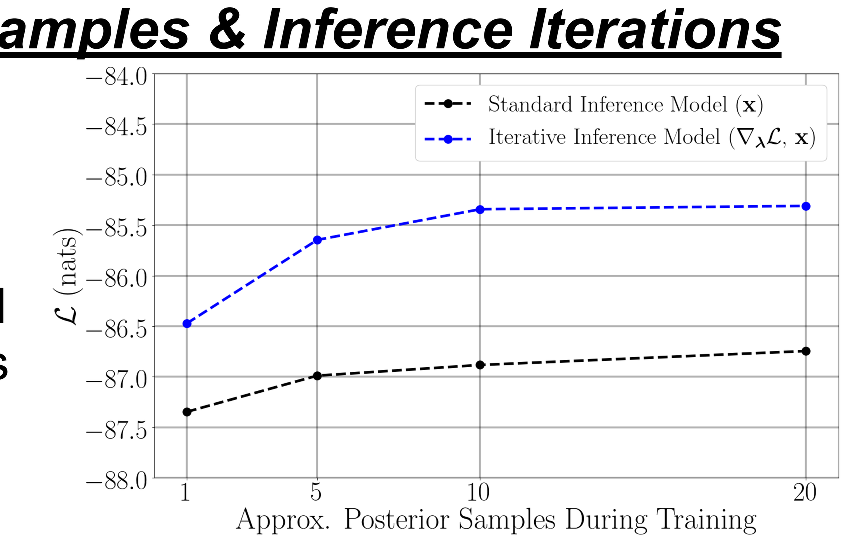
Comparing with Conventional Optimization



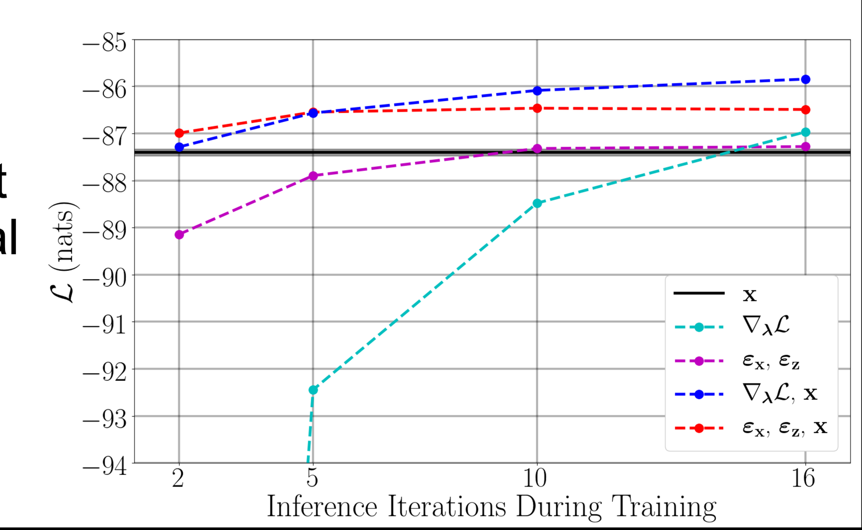
Iterative inference models outperform conventional optimizers in both speed and performance.

Increasing Samples & Inference Iterations

Improvement with additional latent samples

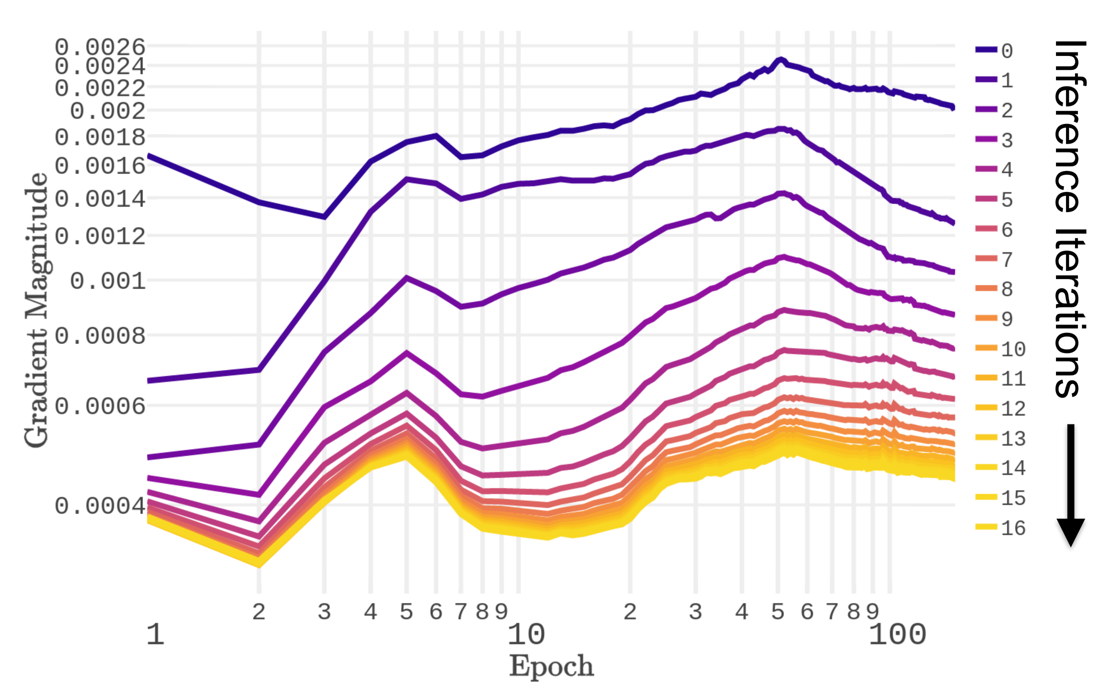


Improvement with additional inference iterations



Approximate Posterior Gradients

Throughout training, gradient magnitudes empirically decrease across inference iterations.



discussion

Generalizing Standard Inference Models

The approximate posterior gradients are <u>stochastic affine transformations</u> of the data.

E.g., approximate posterior mean gradient:

$$\nabla_{\mu_q} \mathcal{L} = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \left[\frac{\partial \mu_\mathbf{x}}{\partial \mu_q}^\intercal \frac{\mathbf{x} - \mu_\mathbf{x}}{\sigma_\mathbf{x}^2} - \frac{\mathbf{z} - \mu_p}{\sigma_p^2} \right]$$

$$\nabla_{\mu_q} \mathcal{L} = \mathbf{A}\mathbf{x} + \mathbf{b}$$

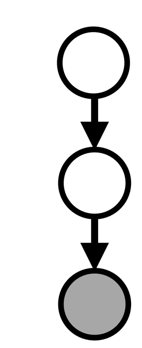
where

$$\mathbf{A} \equiv \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \left[\frac{\partial \mu_\mathbf{x}}{\partial \mu_q}^\intercal (\text{diag } \sigma_\mathbf{x}^2)^{-1} \right] \qquad \mathbf{b} \equiv -\mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \left[\frac{\partial \mu_\mathbf{x}}{\partial \mu_q}^\intercal \frac{\mu_\mathbf{x}}{\sigma_\mathbf{x}^2} + \frac{\mathbf{z} - \mu_p}{\sigma_p^2} \right]$$

→ ***Equivalent to initially encode the <u>data</u> or the <u>gradient</u>.***

Standard inference models are restricted to a single step. Iterative inference models can take multiple steps.

Justifying "Top-Down" Inference



Hierarchical models contain levels of latent variables, providing *empirical priors* on lower variables. These priors vary across data examples, adding flexibility.

The approximate posterior gradients optimally combine terms from **"top-down"** priors and **"bottom-up"** latent variables or data.

E.g., at intermediate levels, the approximate posterior mean gradient:

$$\nabla_{\mu_q^\ell} \mathcal{L} = \mathbb{E}_{q(\mathbf{z}|\cdot)} \left[\underbrace{\frac{\partial \mu_p^{\ell-1}}{\partial \mu_q^\ell}^\intercal \frac{\mathbf{z}^{\ell-1} - \mu_p^{\ell-1}}{(\sigma_p^{\ell-1})^2}}_{\text{bottom-up}} - \underbrace{\frac{\mathbf{z}^\ell - \mu_p^\ell}{(\sigma_p^\ell)^2}}_{\text{top-down}} \right]$$

<u>Standard inference models do not include top-down terms.</u> They were later proposed in [5]. *We provide the first theoretical justification for top-down inference.*

1. Radford M Neal and Geoffrey E Hinton. *A view of the em algorithm that justifies incremental, sparse, and other variants.* 1998.
2. Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. *Stochastic variational inference.* 2013.
3. Diederik P Kingma and Max Welling. *Stochastic gradient vb and the variational auto-encoder.* 2014.
4. Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. *Stochastic backpropagation and approximate inference in deep generative models.* 2014.
5. Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. *Ladder variational autoencoders.* 2016.