Iterative Amortized Inference

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

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

Summary

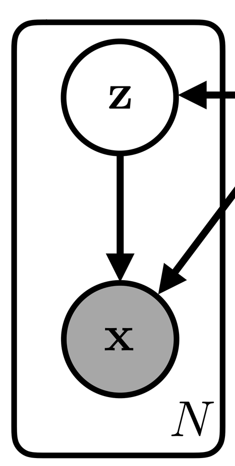
Iterative amortized inference models efficiently and accurately perform variational inference optimization by iteratively encoding *approximate posterior gradients* or *errors*, rather than directly encoding data examples**.** They

- Extend amortized inference models to iterative estimation.
- Provide a principled method of explicitly including latent priors in inference optimization.
- Outperform standard inference models on image and text data sets.

Background

Latent Variable Model

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



Latent Gaussian Model

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

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

Variational Inference

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

Variational EM Algorithm [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]):

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

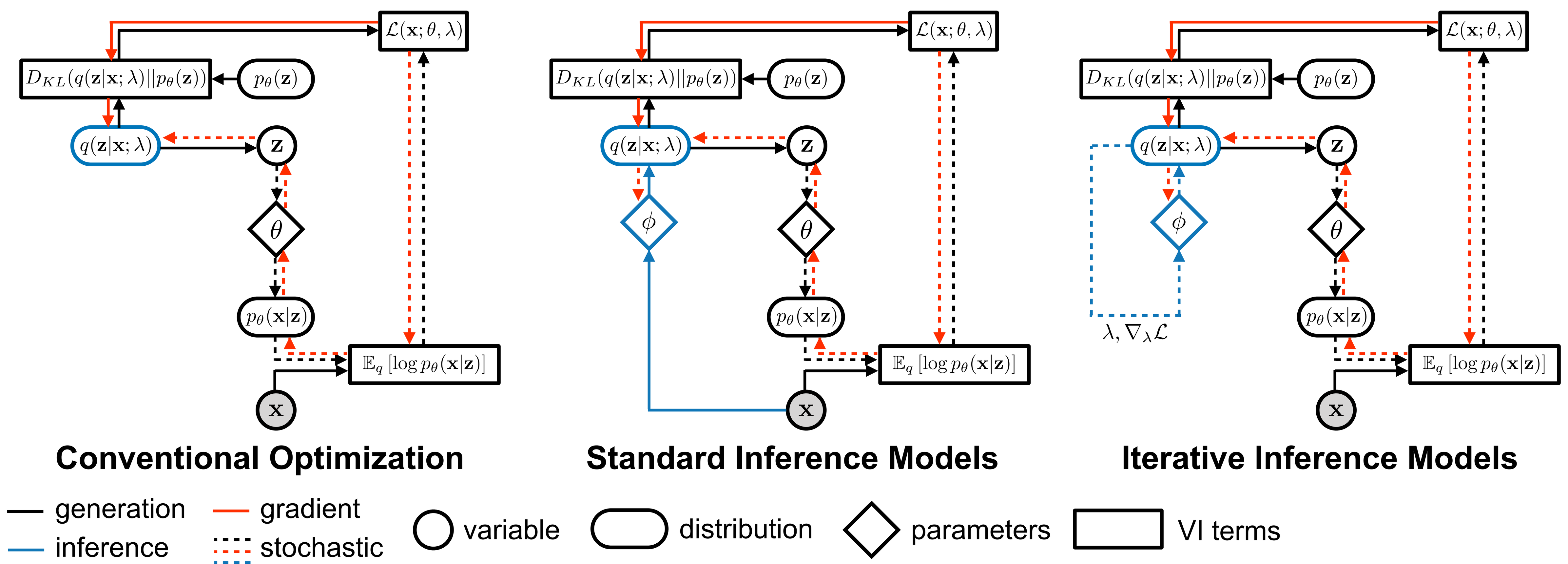
Standard Inference Models, (e.g. VAE [3, 4]):

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

Iterative Inference Models:

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

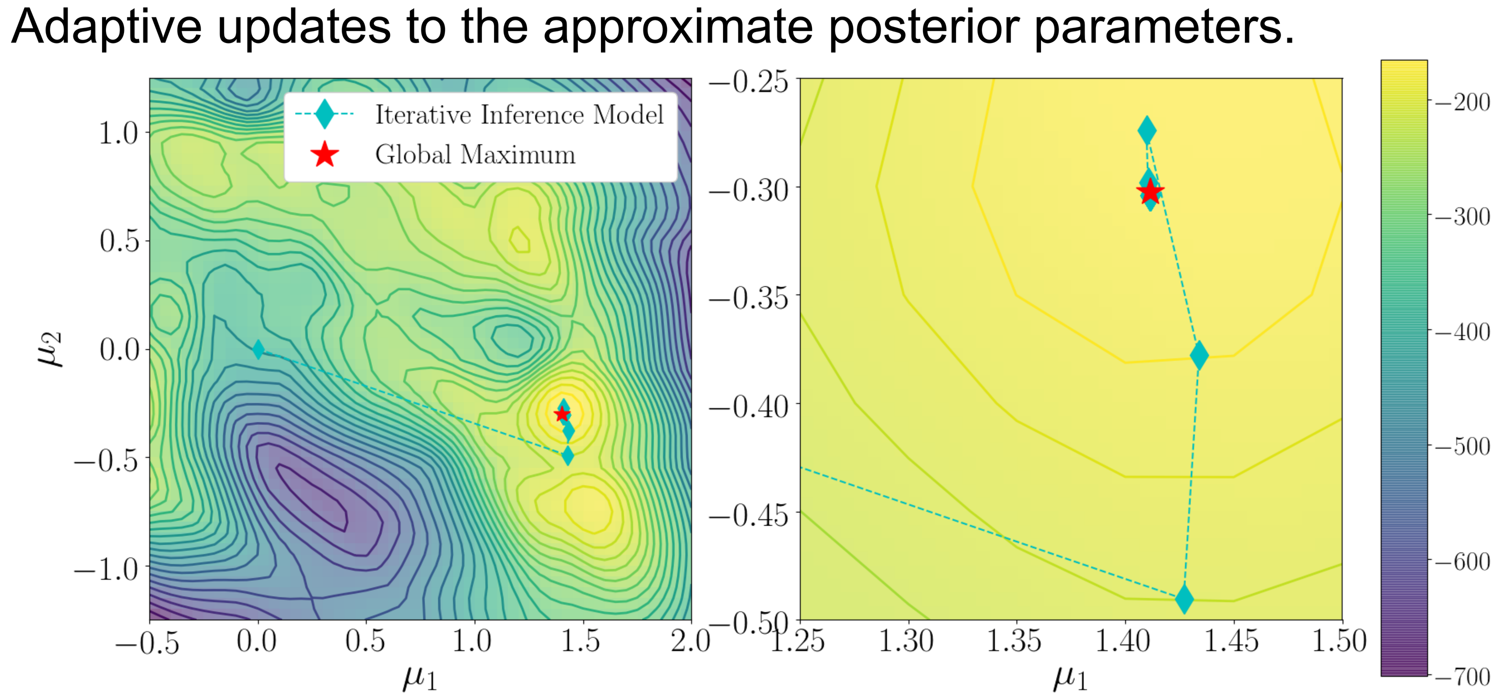
Computation Graphs



Results

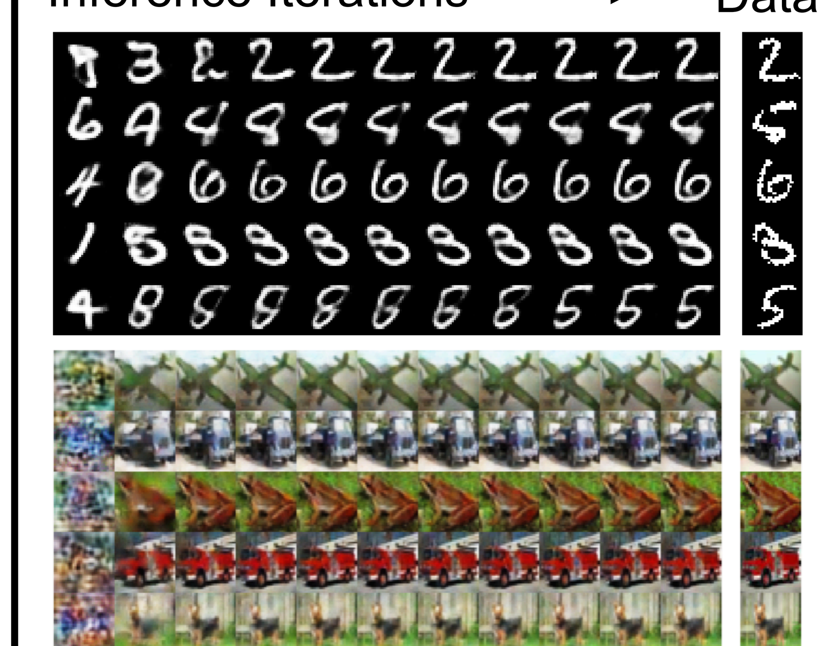
Visualizing Optimization

Adaptive updates to the approximate posterior parameters.

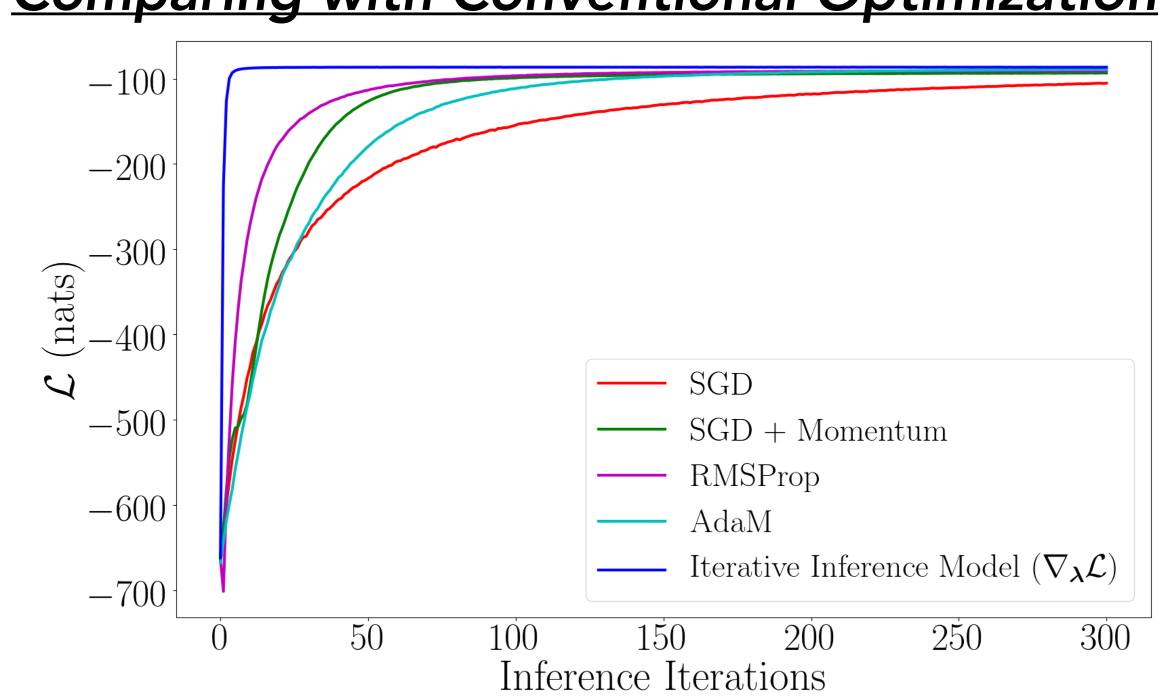


Reconstructions

Inference Iterations → Data



Comparing with Conventional Optimization



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

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.

Discussion

Encoding Errors

In latent Gaussian models, the gradients for the approximate posterior parameters include Jacobians and errors. E.g.:

$$\nabla_{\mu_q} \mathcal{L} = \mathbf{J}^\intercal \varepsilon_\mathbf{x} - \varepsilon_\mathbf{z}$$

where

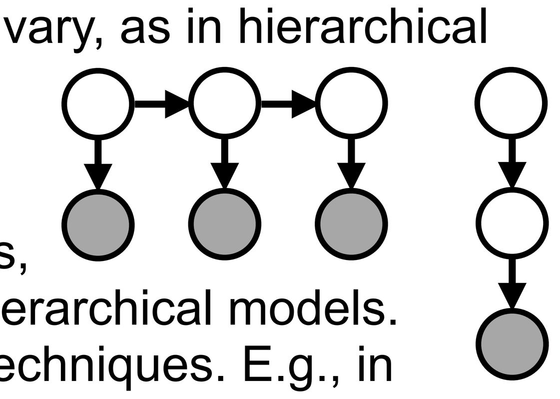
$$\mathbf{J} \equiv \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z}|\mathbf{x})}\left[\frac{\partial \mu_\mathbf{x}}{\partial \mu_q}\right]$$

$$\varepsilon_\mathbf{x} \equiv \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z}|\mathbf{x})}\left[\frac{\mathbf{x} - \mu_\mathbf{x}}{\sigma_\mathbf{x}^2}\right] \qquad \varepsilon_\mathbf{z} \equiv \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z}|\mathbf{x})}\left[\frac{\mathbf{z} - \mu_p}{\sigma_p^2}\right]$$

We propose letting the inference model *learn* the Jacobian, encoding the error terms. This allows us to **avoid computing gradients during inference**, and since the errors contain general curvature information, **models of this form can converge to better estimates in fewer iterations.**

Incorporating Latent Priors

Iterative inference models encode gradients or errors, which **explicitly account for latent priors during optimization**. This is important when these priors vary, as in hierarchical and dynamical models.



Previous works have proposed heuristics to account for these priors, such as top-down inference [5] in hierarchical models. The gradients help to justify these techniques. E.g., in hierarchical models:

$$\nabla_{\mu_q^\ell} \mathcal{L} = \mathbf{J}^{\ell\intercal} \varepsilon_\mathbf{z}^{\ell-1} - \varepsilon_\mathbf{z}^\ell$$

where $\varepsilon_\mathbf{z}^\ell$ is the "top-down" error from the prior. Without access to these terms, a bottom-up standard inference model must implicitly estimate the prior.

Similar arguments apply to dynamical latent variable models, where iterative inference models can explicitly account for dynamical priors.

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.