Iterative Amortized Inference

2018/11/02

OUTLINE

1. Inference Suboptimality

- Variational Inference
- Standard Inference Models
- Approximation Gap & Amortization Gap

2. Meta Learning

- Gradient

3. Iterative Amortized Inference

- Model
- Experiment

Variational Inference

Variation

The extension of differentials in a function space

Inference

Similar to encoding process : $x \rightarrow z$

Variational Inference

$P(z|x)\!: \ x \rightarrow z$

$KL(\ q(z|x)\ ||\ P(z|x)\ )$

$logP(x) = ELBO + KL(\ q(z|x)\ ||\ P(z|x)\ )$

$\min_{q(z|x)} KL(\ q(z|x)\ ||\ P(z|x)\ ) \leftrightarrow \max_{q(z|x)} ELBO$

Variational Inference

$$q(\mathbf{z}^{(i)}|\mathbf{x}^{(i)}) = \mathcal{N}(\mathbf{z}^{(i)}; \boldsymbol{\mu}_q^{(i)}, \mathrm{diag}\, \boldsymbol{\sigma}_q^{2(i)})$$

$$\boldsymbol{\lambda}^{(i)} = \{\boldsymbol{\mu}_q^{(i)}, \boldsymbol{\sigma}_q^{2(i)}\}$$

is not shared for each examples.

$$\boldsymbol{\lambda}^{(i)} \leftarrow \boldsymbol{\lambda}^{(i)} + \alpha \nabla_{\boldsymbol{\lambda}} \mathcal{L}(\mathbf{x}^{(i)}, \boldsymbol{\lambda}^{(i)}; \theta),$$

where L is the ELBO,
and θ is the global parameters.

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

Standard Inference Models

$$q(\mathbf{z}^{(i)}|\mathbf{x}^{(i)}) = \mathcal{N}(\mathbf{z}^{(i)}; \boldsymbol{\mu}_q^{(i)}, \operatorname{diag} \boldsymbol{\sigma}_q^{2(i)})$$

$$\boldsymbol{\lambda}^{(i)} = \{\boldsymbol{\mu}_q^{(i)}, \boldsymbol{\sigma}_q^{2(i)}\}$$

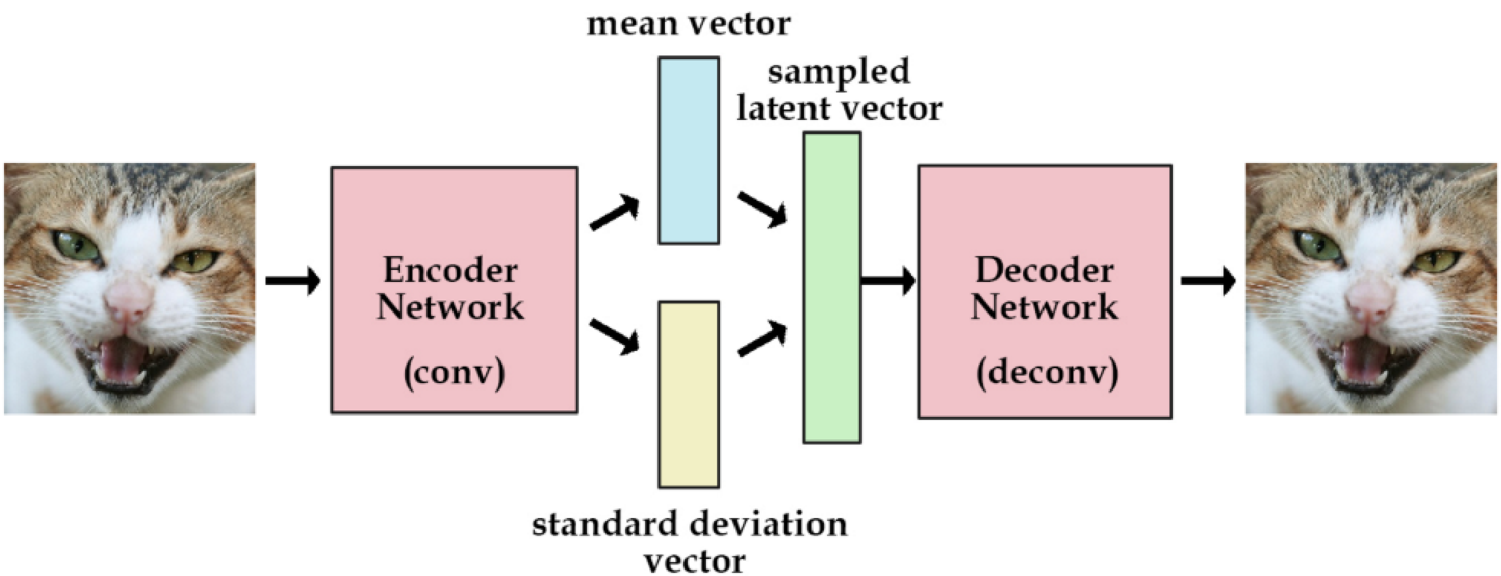
$$\boldsymbol{\lambda}^{(i)} \leftarrow f(\mathbf{x}^{(i)}; \phi).$$

ϕ is a global ***shared*** parameter which does not vary across data examples.

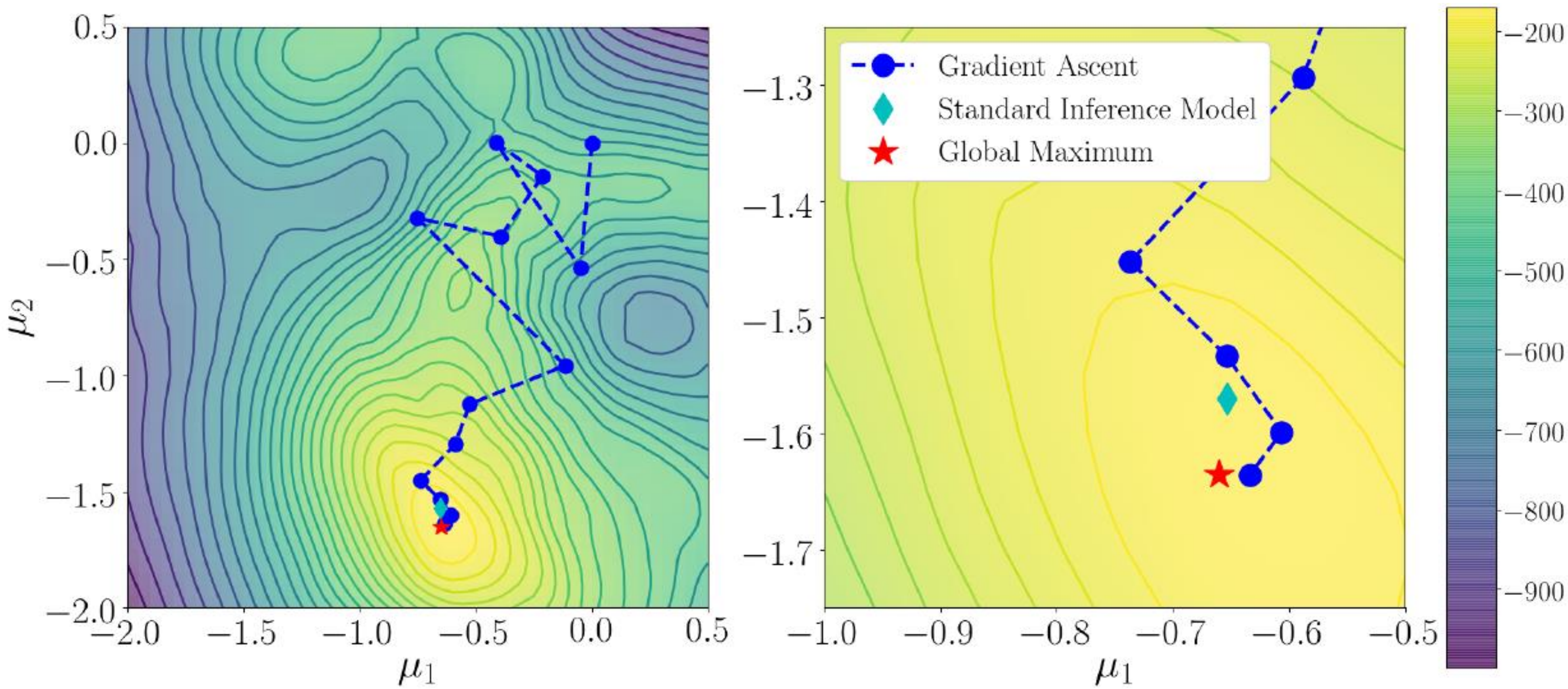
amortized

Standard Inference Models

VAE



Approximation Gap & Amortization Gap



Approximation Gap & Amortization Gap

Approximation Gap

The approximation gap comes from the inability of the variational distribution family to exactly match the true posterior.

Amortization Gap

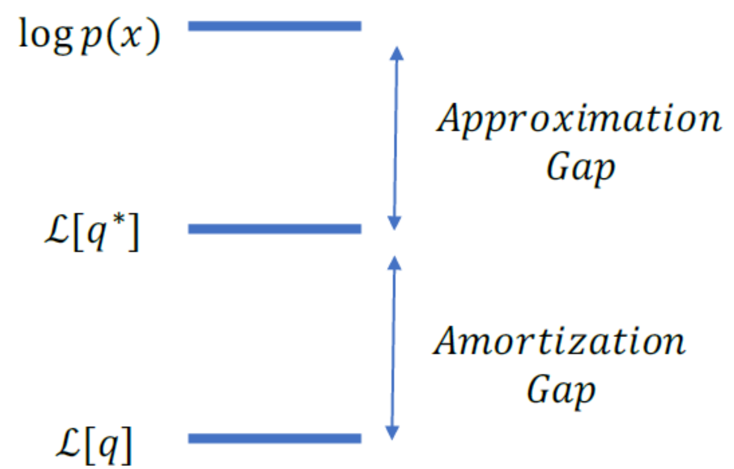
The amortization gap refers to the difference caused by amortizing the variational parameters over the entire training set, instead of optimizing for each training example individually.

Cremer C, Li X, Duvenaud D. Inference Suboptimality in Variational Autoencoders[J]. 2018.

Approximation Gap & Amortization Gap

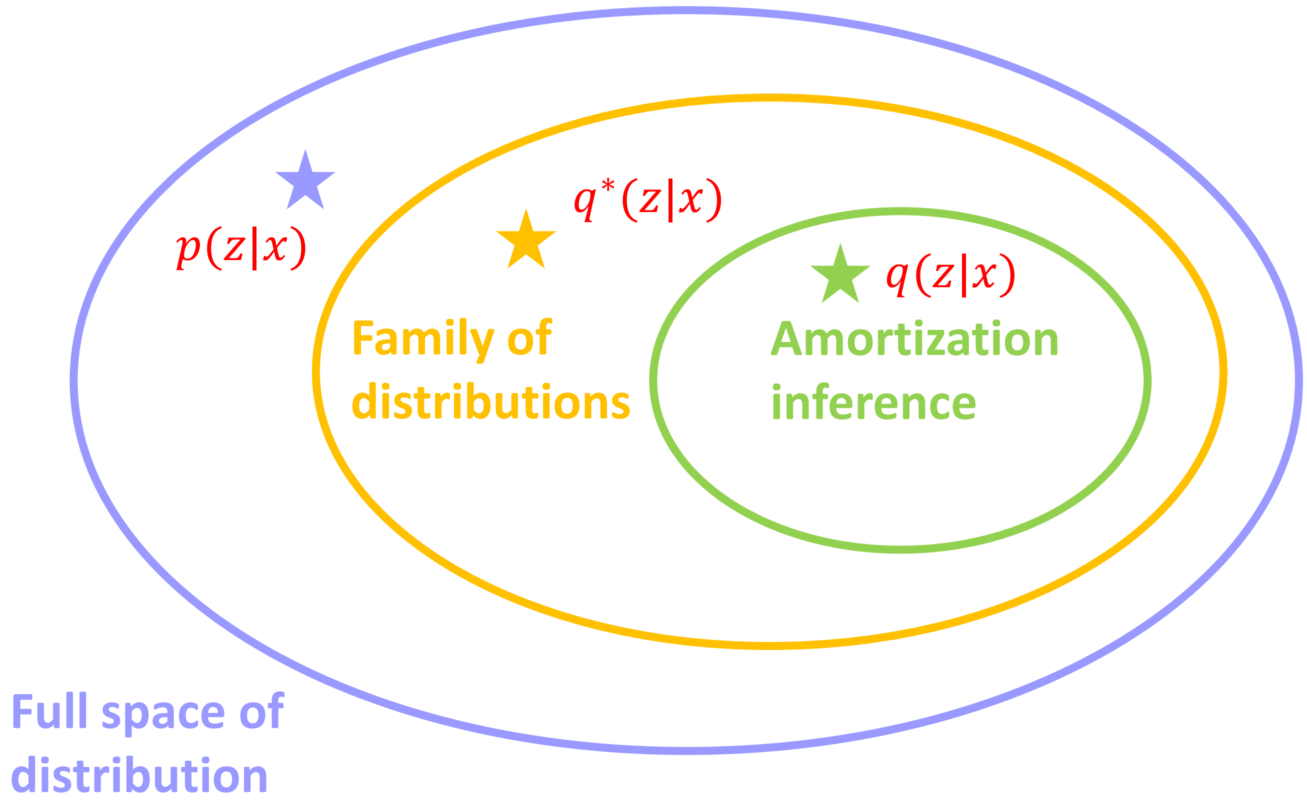
The ELBO evaluated using the optimal approximation within its variational family.

The ELBO evaluated using an amortized distribution q, as is typical of VAE training.



Cremer C, Li X, Duvenaud D. Inference Suboptimality in Variational Autoencoders[J]. 2018.

Approximation Gap & Amortization Gap



Approximation Gap & Amortization Gap

$$\mathcal{G} = \log p(x) - \mathcal{L}[q] = \underbrace{\log p(x) - \mathcal{L}[q^*]}_{\text{Approximation}} + \underbrace{\mathcal{L}[q^*] - \mathcal{L}[q]}_{\text{Amortization}}.$$

The inference gap $\mathcal{G}$ is the difference between the marginal log-likelihood and a lower bound 。

$$\begin{aligned}\mathcal{G}_{\text{VAE}} = &\underbrace{\text{KL}\big(q^*(z|x)||p(z|x)\big)}_{\text{Approximation}} \\ &+ \underbrace{\text{KL}\big(q(z|x)||p(z|x)\big) - \text{KL}\big(q^*(z|x)||p(z|x)\big)}_{\text{Amortization}}.\end{aligned}$$

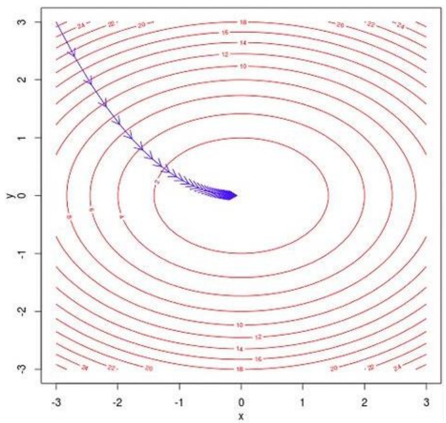
Cremer C, Li X, Duvenaud D. Inference Suboptimality in Variational Autoencoders[J]. 2018.

Learning to learn by gradient descent by gradient descent

Marcin Andrychowicz , Misha Denil
Sergio Gómez Colmenarejo , Matthew W. Hoffman
David Pfau , Tom Schaul , Brendan Shillingford
Nando de Freitas

*30th Conference on Neural Information Processing Systems (**NIPS 2016**), Barcelona, Spain*

Gradient Descent



$$\theta_{t+1} = \theta_t - \alpha_t \nabla f(\theta_t)$$

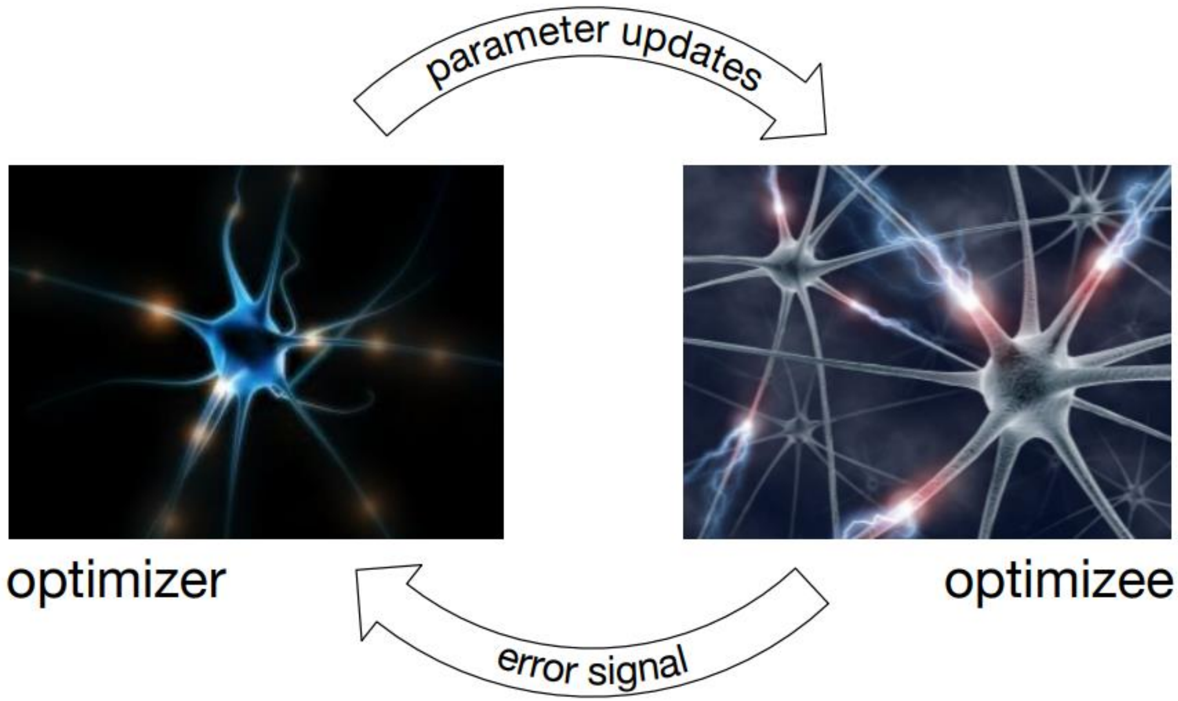
Second-order Information:

Hessian matrix , Gauss-Newton matrix , Fisher information matrix

Non-convex Optimization:

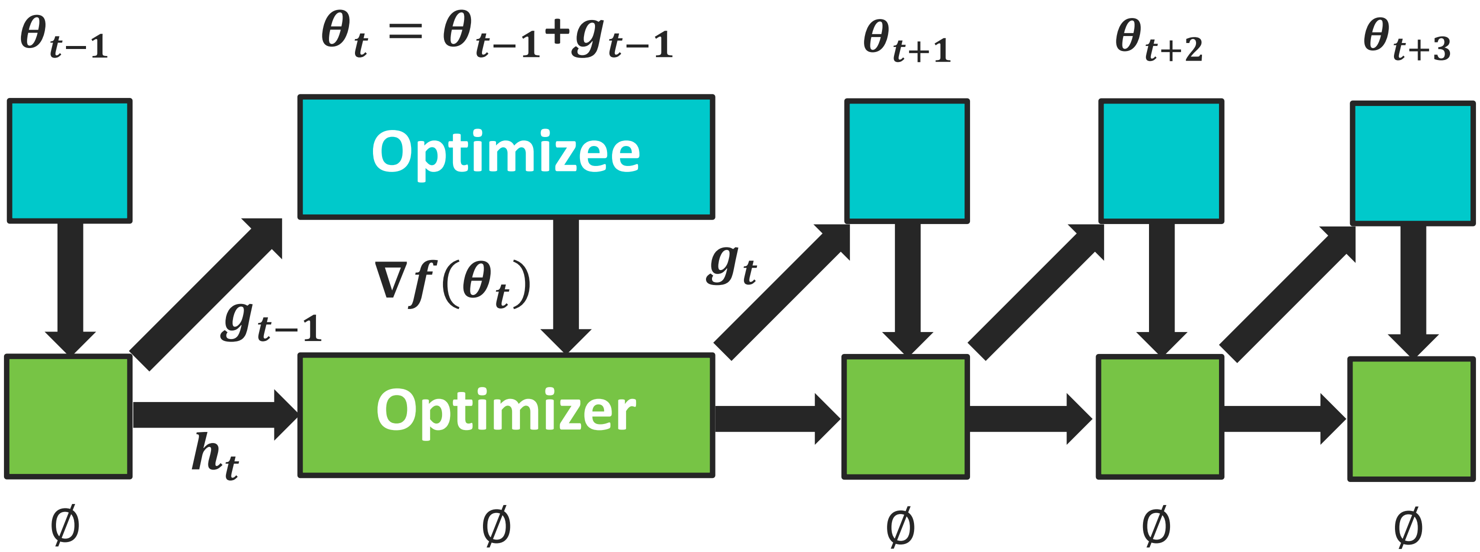
Momentum , Rprop , Adagrad , RMSprop , ADAM

Optimizer



$$\theta_{t+1} = \theta_t + g_t(\nabla f(\theta_t), \phi)$$

Optimizer



θ_{t-1}
$\theta_t = \theta_{t-1} + g_{t-1}$
θ_{t+1}
θ_{t+2}
θ_{t+3}
Optimizee
$\nabla f(\theta_t)$
g_t
g_{t-1}
Optimizer
h_t
$\emptyset$
$\emptyset$
$\emptyset$
$\emptyset$
$\emptyset$

Caltech × Disney Research

Iterative Amortized Inference

Joseph Marino, Yisong Yue, Stephan Mandt

Proceedings of the 35th International Conference on Machine Learning
***(ICML 2018)**, Stockholm, Sweden*

Iterative Amortized Inference

Gradient ascent

$$x_i \xrightarrow{\text{optimize}} q(z|x_i)$$

For each examples

Standard inference models

$$x_i \longrightarrow \boxed{\text{inference models}} \longrightarrow q(z|x_i)$$

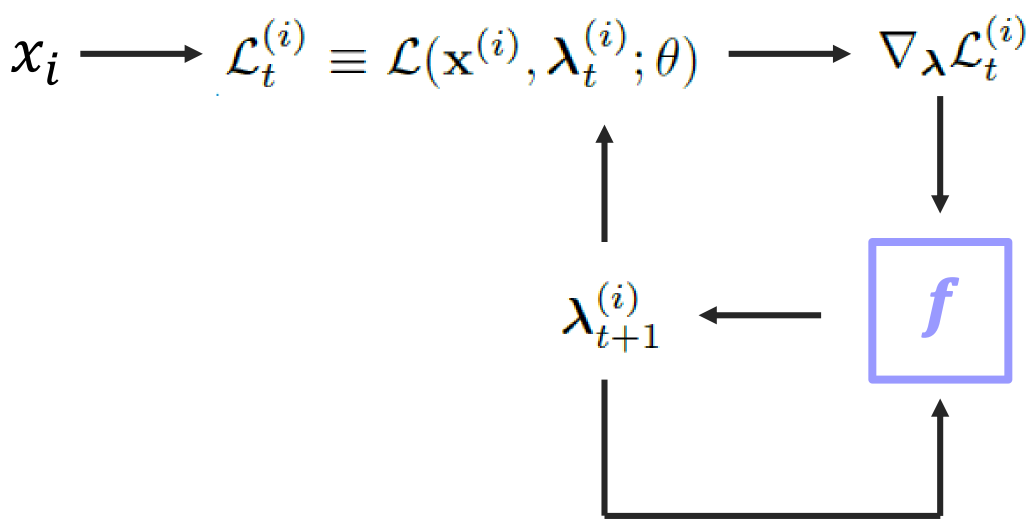
Shared & Amortized

Iterative Amortized Inference

Iterative Inference Models

$$\boldsymbol{\lambda}_{t+1}^{(i)} \leftarrow f_t(\nabla_{\boldsymbol{\lambda}}\mathcal{L}_t^{(i)}, \boldsymbol{\lambda}_t^{(i)}; \phi),$$

f*: *An optimizer



Iterative Amortized Inference

Algorithm 1 Iterative Amortized Inference

Input: data $\mathbf{x}$, generative model $p_\theta(\mathbf{x}, \mathbf{z})$, inference model f

Initialize $t = 0$

Initialize $\nabla_\phi = 0$

Initialize $q(\mathbf{z}|\mathbf{x})$ with $\boldsymbol{\lambda}_0$

repeat

  Sample $\mathbf{z} \sim q(\mathbf{z}|\mathbf{x})$

  Evaluate $\mathcal{L}_t = \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}_t; \theta)$

  Calculate $\nabla_{\boldsymbol{\lambda}} \mathcal{L}_t$ and $\nabla_\phi \mathcal{L}_t$

  Update $\boldsymbol{\lambda}_{t+1} = f_t(\nabla_{\boldsymbol{\lambda}} \mathcal{L}_t, \boldsymbol{\lambda}_t; \phi)$

  $t = t + 1$

  $\nabla_\phi = \nabla_\phi + \nabla_\phi \mathcal{L}_t$

until $\mathcal{L}$ converges

$\theta = \theta + \alpha_\theta \nabla_\theta \mathcal{L}$

$\phi = \phi + \alpha_\phi \nabla_\phi$

Iterative Amortized Inference

Latent Gaussian Models

$$q(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_q, \operatorname{diag} \boldsymbol{\sigma}_q^2) \qquad \boldsymbol{\lambda}^{(i)} : \{\boldsymbol{\mu}_q^{(i)}, \boldsymbol{\sigma}_q^{2(i)}\}$$

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_p, \operatorname{diag} \boldsymbol{\sigma}_p^2)$$

$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{\mathbf{x}}, \operatorname{diag} \boldsymbol{\sigma}_{\mathbf{x}}^2)$$

$$\boldsymbol{\mu}_{q,t+1} = f_t^{\boldsymbol{\mu}_q}(\nabla_{\boldsymbol{\mu}_q} \mathcal{L}_t, \boldsymbol{\mu}_{q,t}; \phi), \quad \boldsymbol{\sigma}_{q,t+1}^2 = f_t^{\boldsymbol{\sigma}_q^2}(\nabla_{\boldsymbol{\sigma}_q^2} \mathcal{L}_t, \boldsymbol{\sigma}_{q,t}^2; \phi),$$

Iterative Amortized Inference

Latent Gaussian Models

$$q(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_q, \operatorname{diag} \boldsymbol{\sigma}_q^2) \qquad p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_\mathbf{x}, \operatorname{diag} \boldsymbol{\sigma}_\mathbf{x}^2)$$

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_p, \operatorname{diag} \boldsymbol{\sigma}_p^2)$$

$$\nabla_{\boldsymbol{\mu}_q} \mathcal{L} = \mathbf{J}^\intercal \varepsilon_\mathbf{x} - \varepsilon_\mathbf{z}, \qquad \mathbf{J} \equiv \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z}|\mathbf{x})} \left[\frac{\partial \boldsymbol{\mu}_\mathbf{x}}{\partial \boldsymbol{\mu}_q} \right]$$

$$\varepsilon_\mathbf{x} \equiv \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z}|\mathbf{x})} [(\mathbf{x} - \boldsymbol{\mu}_\mathbf{x}) / \boldsymbol{\sigma}_\mathbf{x}^2],$$

$$\varepsilon_\mathbf{z} \equiv \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z}|\mathbf{x})} [(\mathbf{z} - \boldsymbol{\mu}_p) / \boldsymbol{\sigma}_p^2].$$

assume μ_x is a function of z and σ_x is a global parameter.

Iterative Amortized Inference

Latent Gaussian Models

$$\nabla_{\boldsymbol{\mu}_q}\mathcal{L} = \mathbf{J}^{\intercal}\boldsymbol{\varepsilon}_{\mathbf{x}} - \boldsymbol{\varepsilon}_{\mathbf{z}}, \qquad \mathbf{J} \equiv \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|\mathbf{x})}\left[\frac{\partial \boldsymbol{\mu}_{\mathbf{x}}}{\partial \boldsymbol{\mu}_q}\right]$$

$$\boldsymbol{\varepsilon}_{\mathbf{x}} \equiv \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|\mathbf{x})}[(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})/\boldsymbol{\sigma}_{\mathbf{x}}^2],$$

$$\boldsymbol{\varepsilon}_{\mathbf{z}} \equiv \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|\mathbf{x})}[(\mathbf{z} - \boldsymbol{\mu}_p)/\boldsymbol{\sigma}_p^2].$$

Inspecting and understanding the composition of the gradients reveals the forces pushing the approximate posterior toward **agreement with the data, through ε_x, and agreement with the prior, through ε_z**. In other words, inference is as much a top-down process as it is a bottom-up process, and the optimal combination of these terms is given by the approximate posterior gradients.

Iterative Amortized Inference

Latent Gaussian Models

$$\nabla_{\boldsymbol{\mu}_q} \mathcal{L} = \mathbf{J}^\intercal \boldsymbol{\varepsilon}_{\mathbf{x}} - \boldsymbol{\varepsilon}_{\mathbf{z}}, \qquad \mathbf{J} \equiv \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z}|\mathbf{x})} \left[\frac{\partial \boldsymbol{\mu}_{\mathbf{x}}}{\partial \boldsymbol{\mu}_q} \right]$$

$$\boldsymbol{\varepsilon}_{\mathbf{x}} \equiv \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z}|\mathbf{x})} [(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}}) / \boldsymbol{\sigma}_{\mathbf{x}}^2],$$

$$\boldsymbol{\varepsilon}_{\mathbf{z}} \equiv \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z}|\mathbf{x})} [(\mathbf{z} - \boldsymbol{\mu}_p) / \boldsymbol{\sigma}_p^2].$$

$$\boldsymbol{\mu}_{q,t+1} = f_t^{\boldsymbol{\mu}_q}(\boldsymbol{\varepsilon}_{\mathbf{x},t}, \boldsymbol{\varepsilon}_{\mathbf{z},t}, \boldsymbol{\mu}_{q,t}; \phi),$$

$$\boldsymbol{\sigma}_{q,t+1}^2 = f_t^{\boldsymbol{\sigma}_q^2}(\boldsymbol{\varepsilon}_{\mathbf{x},t}, \boldsymbol{\varepsilon}_{\mathbf{z},t}, \boldsymbol{\sigma}_{q,t}^2; \phi),$$

Iterative Amortized Inference

Generalization

$$\varepsilon_{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{b}, \qquad \mathbf{A} \equiv \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}\left[(\operatorname{diag}\boldsymbol{\sigma}_{\mathbf{x}}^2)^{-1}\right],$$

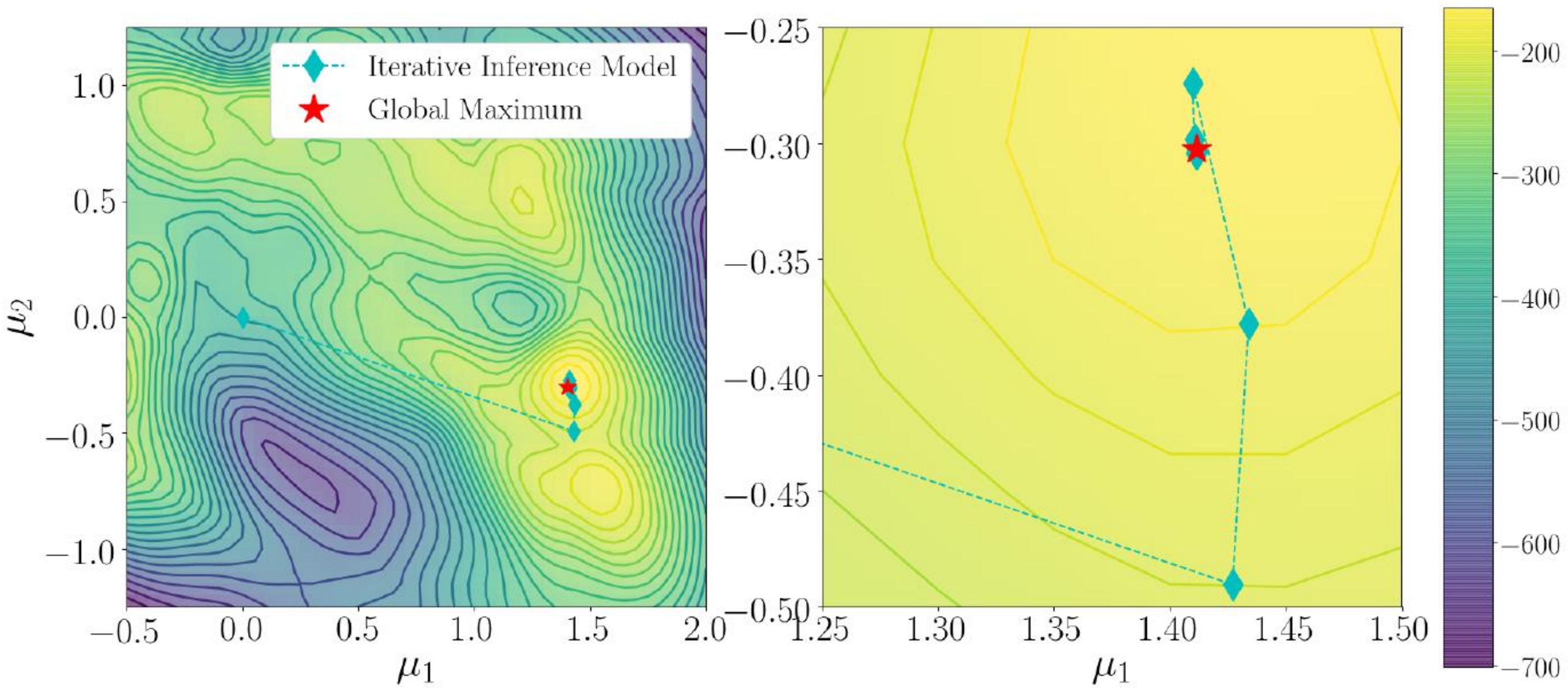
$$\mathbf{b} \equiv -\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}\left[\frac{\boldsymbol{\mu}_{\mathbf{x}}}{\boldsymbol{\sigma}_{\mathbf{x}}^2}\right].$$

$$\boldsymbol{\mu}_{q,t+1} = f_t^{\boldsymbol{\mu}_q}(\varepsilon_{\mathbf{x},t}, \varepsilon_{\mathbf{z},t}, \boldsymbol{\mu}_{q,t}; \phi),$$

$$\boldsymbol{\sigma}_{q,t+1}^2 = f_t^{\boldsymbol{\sigma}_q^2}(\varepsilon_{\mathbf{x},t}, \varepsilon_{\mathbf{z},t}, \boldsymbol{\sigma}_{q,t}^2; \phi),$$

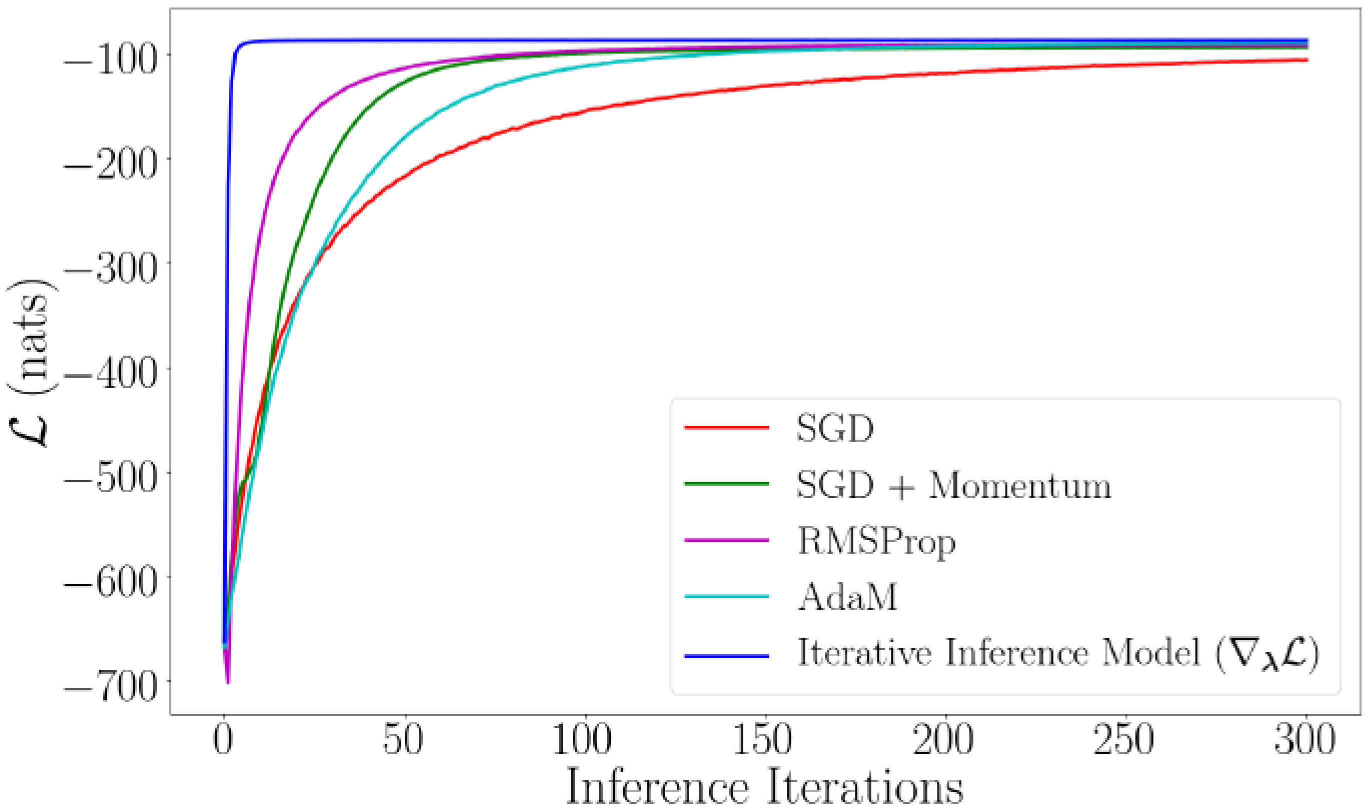
Reasonably assuming that the initial approximate ***posterior and prior are both constant***, standard inference models are equivalent to the special case of a one-step iterative inference model.

Experiments



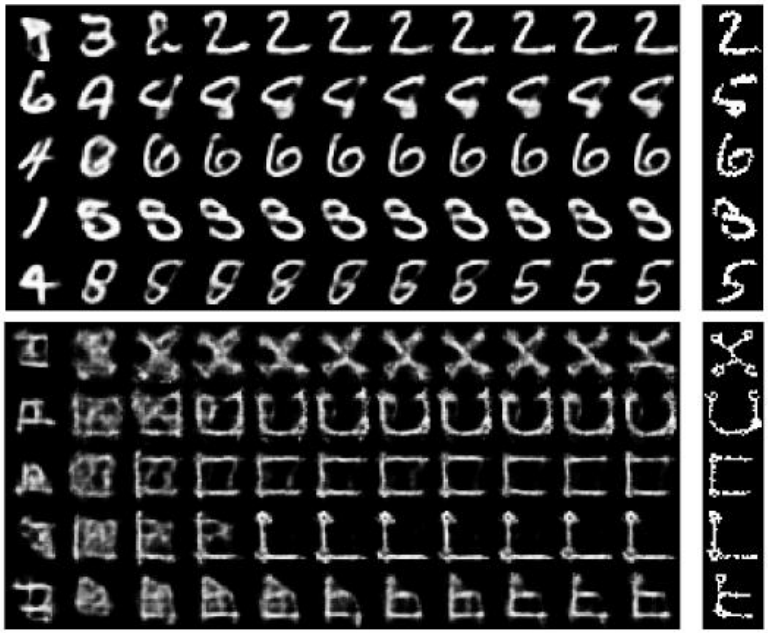
MNIST

Experiments

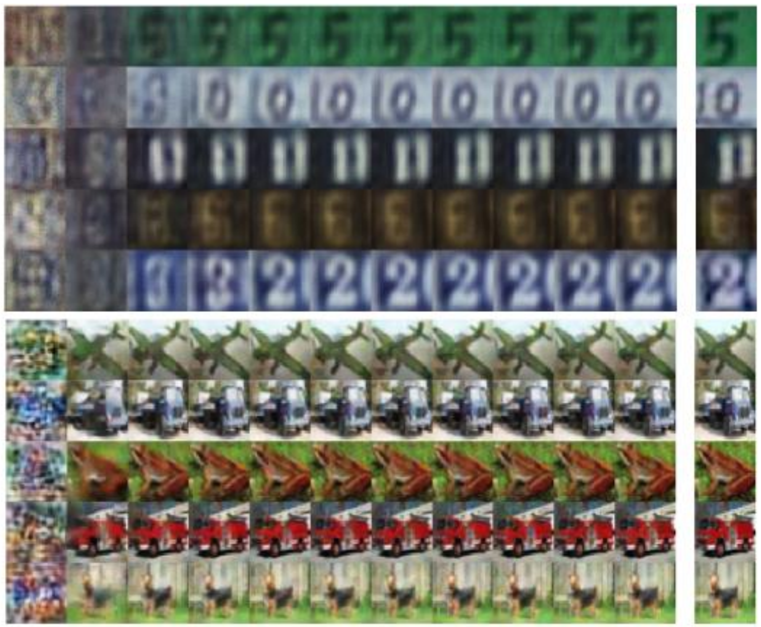


MNIST

Experiments

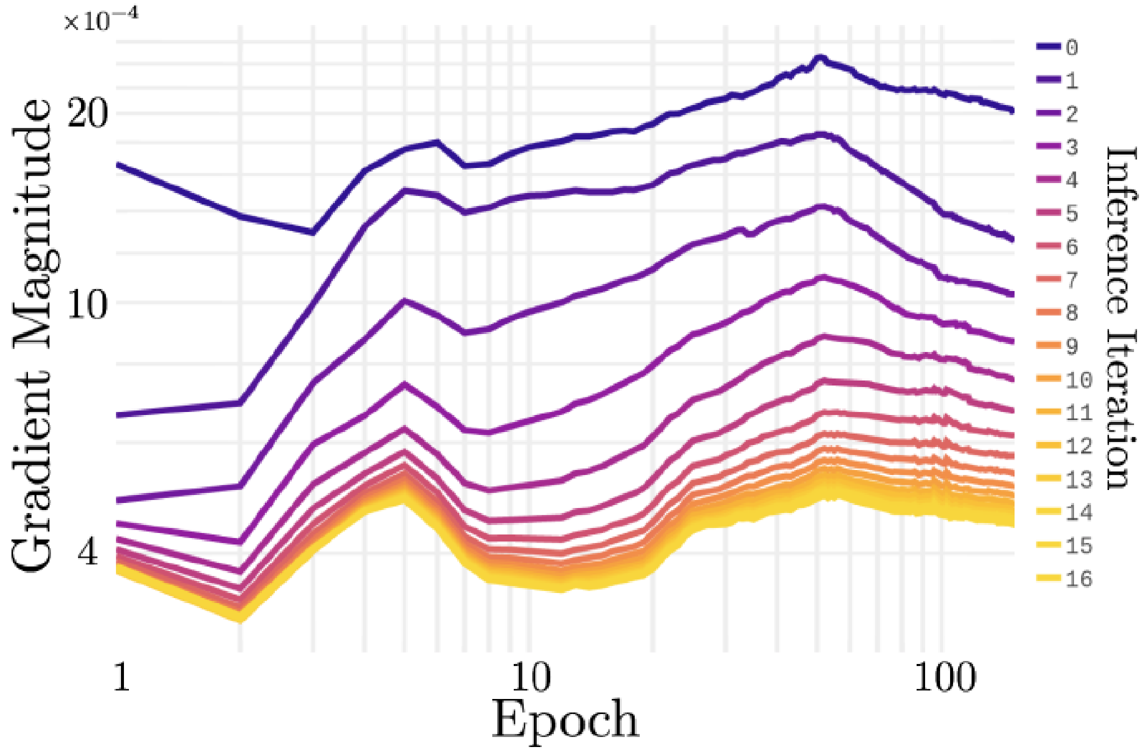


MNIST, Omniglot



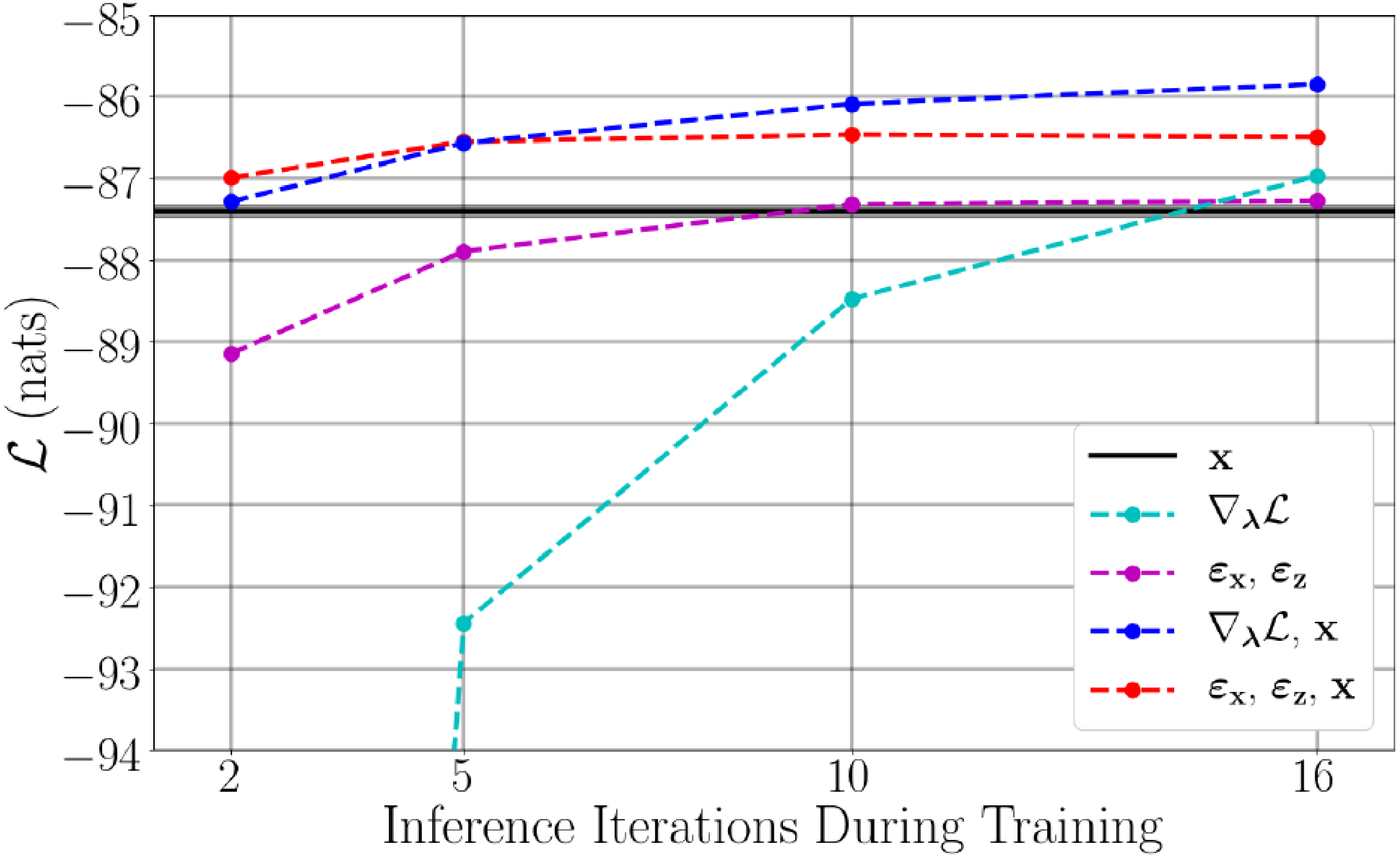
SVHN, CIFAR-10

Experiments

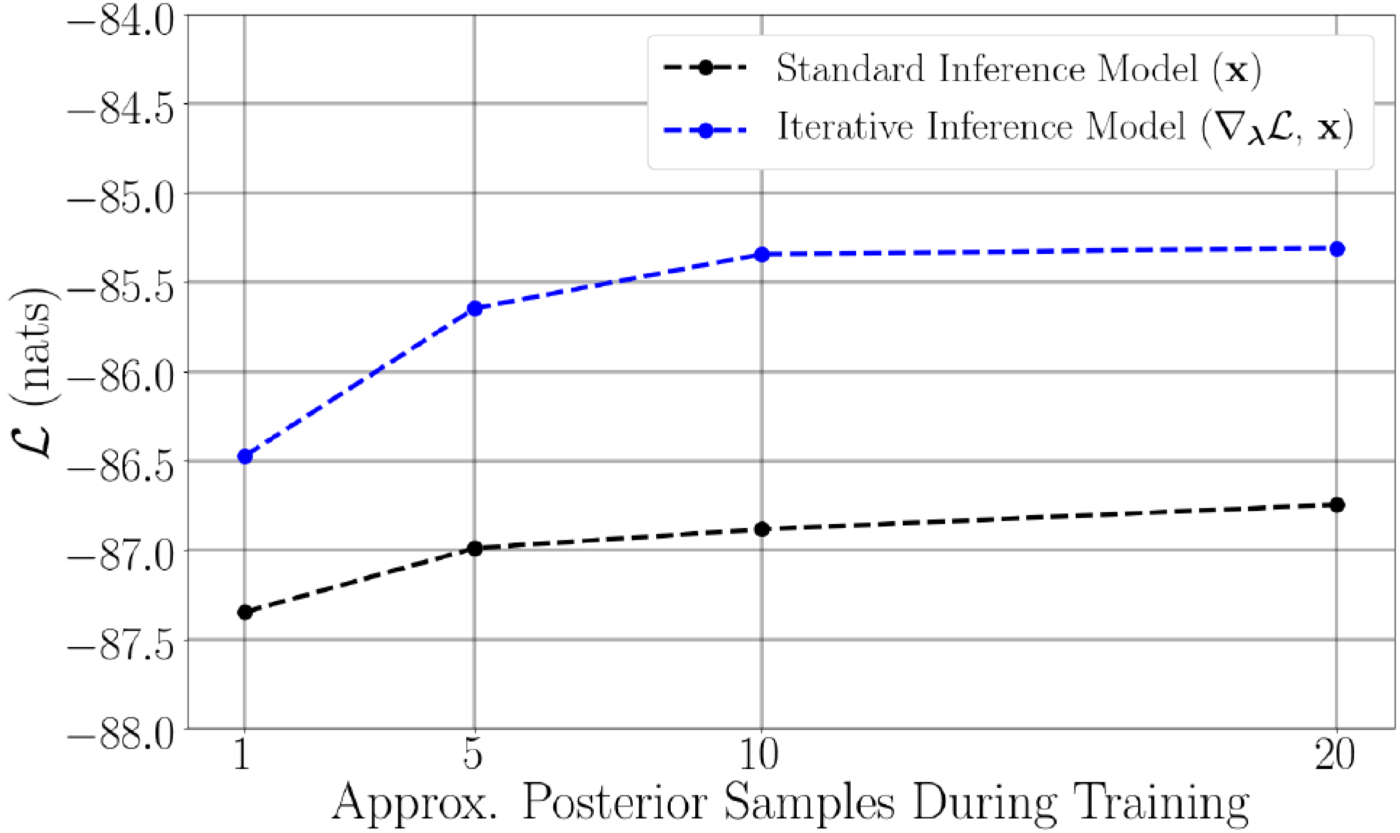


RCV1

Experiments



Experiments



Experiments

	$-\log p(\mathbf{x})$
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	$\leq$
RCV1		
Krishnan et al. (2018)		331
Standard	323 ± 3	377.4 ± 0.5
Iterative	$\mathbf{285.0 \pm 0.1}$	$\mathbf{314 \pm 1}$

$$P \equiv \exp\left(-\frac{1}{N}\sum_i \frac{1}{N_i} \log p(\mathbf{x}^{(i)})\right),$$

Iterative Amortized Inference

Pros

- Retain the advantages of varitional inference and standard inference models.
- Employ meta-learning to guide the update.
- Generalize the standard inference models.

Cons

- Require additional computation.
- Introduce a hype-parameter, the number of iteration.

Thank you

2018/11/02