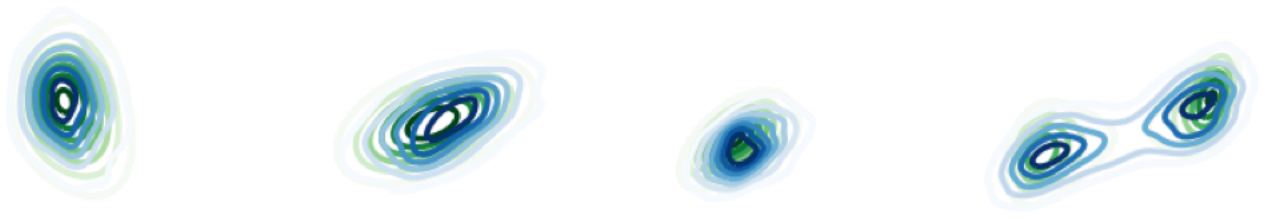
Inference Suboptimality in Variational Autoencoders



Chris Cremer, Xuechen Li, David Duvenaud

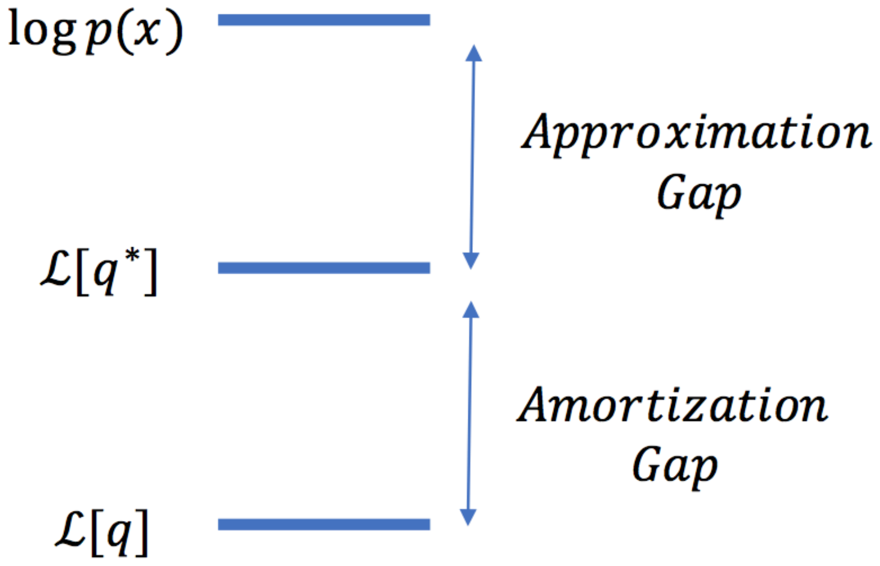
VAE Objective

$$\log p(x) = \underbrace{\mathbb{E}_{z\sim q(z|x)}\left[\log\left(\frac{p(x,z)}{q(z|x)}\right)\right]}_{\text{ELBO}} + \underbrace{\mathbf{KL}\left(q(z|x)||p(z|x)\right)}_{\text{Inference Gap}}$$

Inference Gaps

- Approximation Gap
  - Inability of the variational distribution to model the true posterior

- Amortization Gap
  - Limited capacity of the recognition network to generalize inference over all datapoints



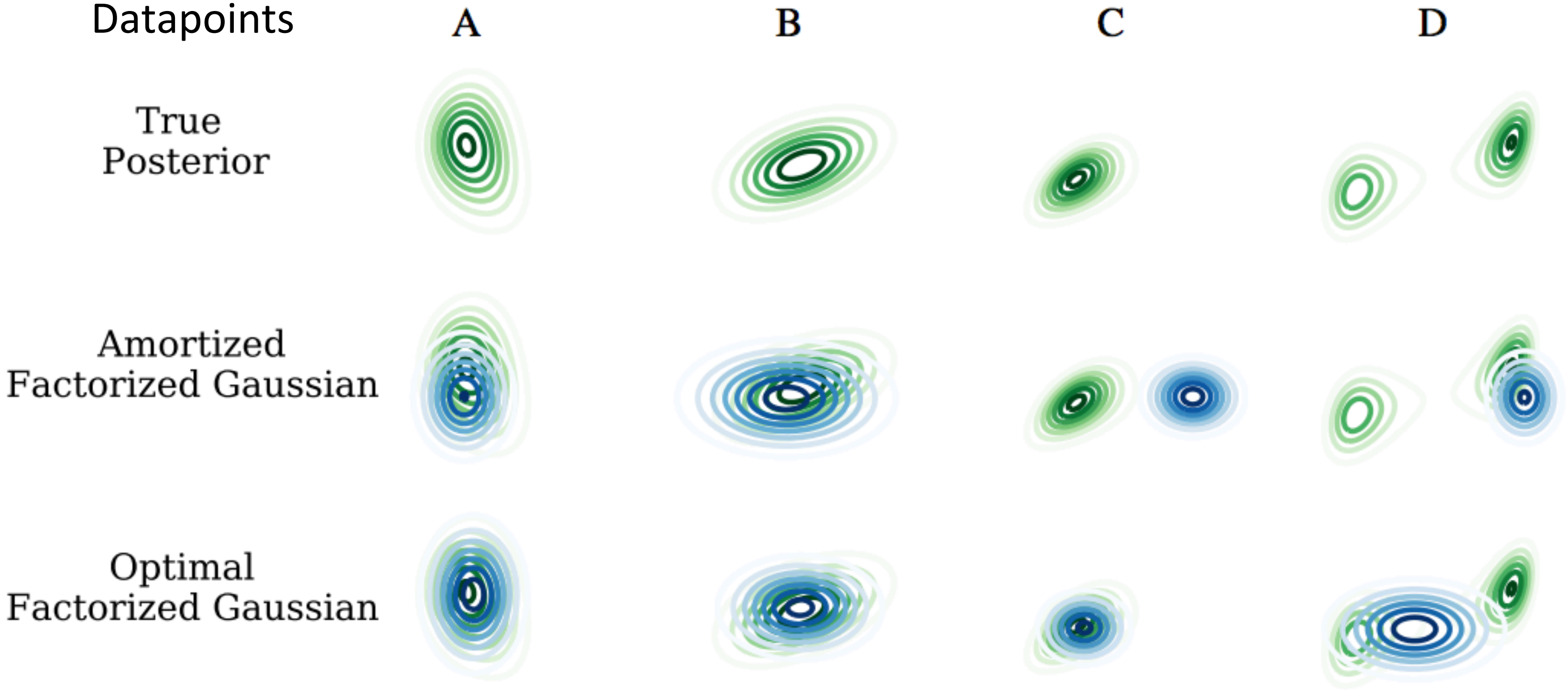
Posterior Visualizations

Datapoints

$p(z|x)$

q_{FFG}
$(z|x)$

$q_{FFG}^{*}(z|$

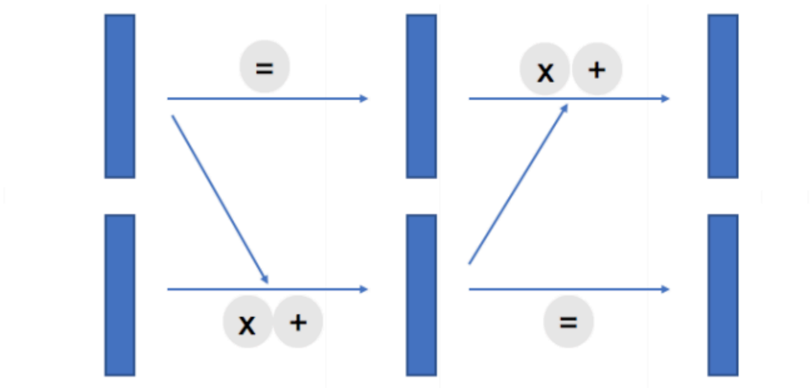


Amortization Gap

Approximation Gap

Flexible Approximations

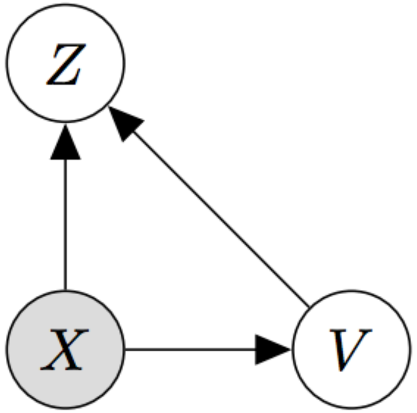
Flow Transformation



$$v' = v \circ \sigma_1(z) + \mu_1(z)$$
$$z' = z \circ \sigma_2(v') + \mu_2(v')$$

$q_{Flow}(z|x)$

Auxiliary Variable



Inference Model

$q_{AF}(z|x)$

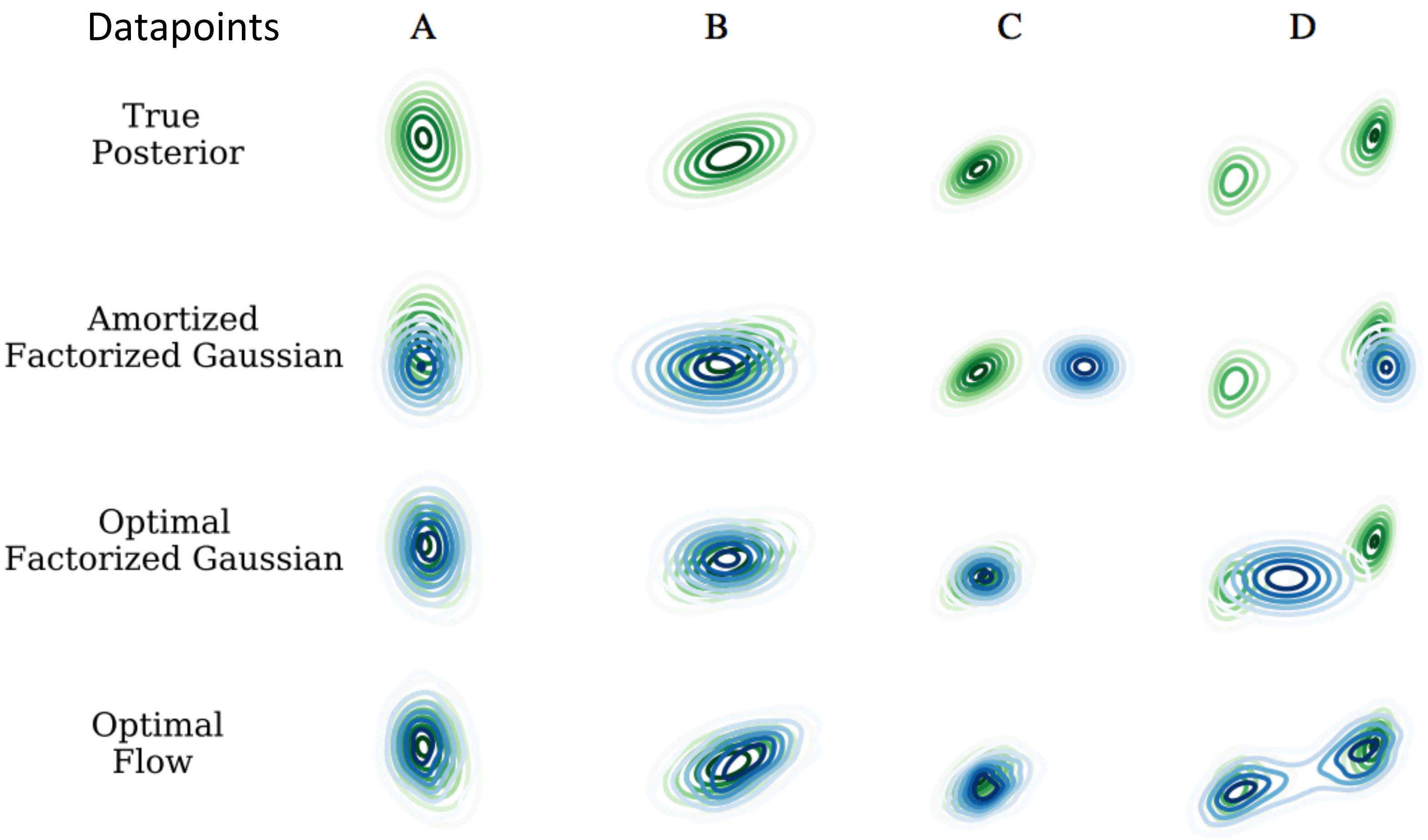
Posterior Visualizations

$p(z|x)$

$q_{FFG}(z|x)$

$q_{FFG}^{*}(z|x)$

$q_{Flow}^{*}(z|x)$



Estimating the Gaps

$\max(\mathcal{L}_{IWAE}[q^*], \mathcal{L}_{AIS})$ $\approx$ $\log p(x)$

Approximation Gap

Lower bound with optimal q within its variational family
For every datapoint, optimize its variational parameters $\approx$ $\mathcal{L}[q^*]$

Amortization Gap

Lower bound with the amortized q $\approx$ $\mathcal{L}[q]$

Amortization vs Approximation

	MNIST		Fashion-MNIST		3-BIT CIFAR	
	q_{FFG}	q_{AF}	q_{FFG}	q_{AF}	q_{FFG}	q_{AF}
$\log \hat{p}(x)$	-89.80	-88.94	-97.47	-97.41	-816.9	-820.56
$\mathcal{L}_{\text{VAE}}[q^*_{AF}]$	-90.80	-90.38	-98.92	-99.10	-820.19	-822.16
$\mathcal{L}_{\text{VAE}}[q^*_{FFG}]$	-91.23	-113.54	-100.53	-132.46	-831.65	-861.62
$\mathcal{L}_{\text{VAE}}[q]$	-92.57	-91.79	-104.75	-103.76	-869.12	-864.28
Approximation	1.43	1.44	3.06	1.69	14.75	1.60
Amortization	1.34	1.41	4.22	4.66	37.47	42.12
Inference	2.77	2.85	7.28	6.35	52.22	43.72

For these model choices:

- Amortization is generally larger than approximation gap

Can we reduce the amortization gap by increasing encoder capacity?

Larger Encoder Reduces Amortization Error

	MNIST		Fashion-MNIST	
	q_{FFG}	q_{AF}	q_{FFG}	q_{AF}
Regular Encoder	1.34	1.41	4.22	4.66
Larger Encoder	1.11	0.75	3.09	3.76

Parameters of Flow Reduce Amortization Gap

- Common reasoning for flow: reduce approximation gap
  - Could improvements also be due to reduction in amortization gap?

- Experiment:
  - Trained a VAE on MNIST
  - Retrained new encoders on the fixed decoder
    - Encoders differ only in their variational distribution

	q_{FFG}	q_{Flow}
Approximation	1.91	0.43
Amortization	43.22	12.86

Generator Learns to Accommodate the Approximation

How much does $p(z|x)$ fit to $q(z|x)$?

How Gaussian is $p(z|x)$ when trained with q_{FFG} vs q_{AF}?

	Generator Trained With	
	q_{FFG}	q_{AF}
$KL(q_{FFG}^{*}(z|x)\,||\,p(z|x))$	1.43	24.60
$KL(q_{AF}^{*}(z|x)\,||\,p(z|x))$	1.00	1.44

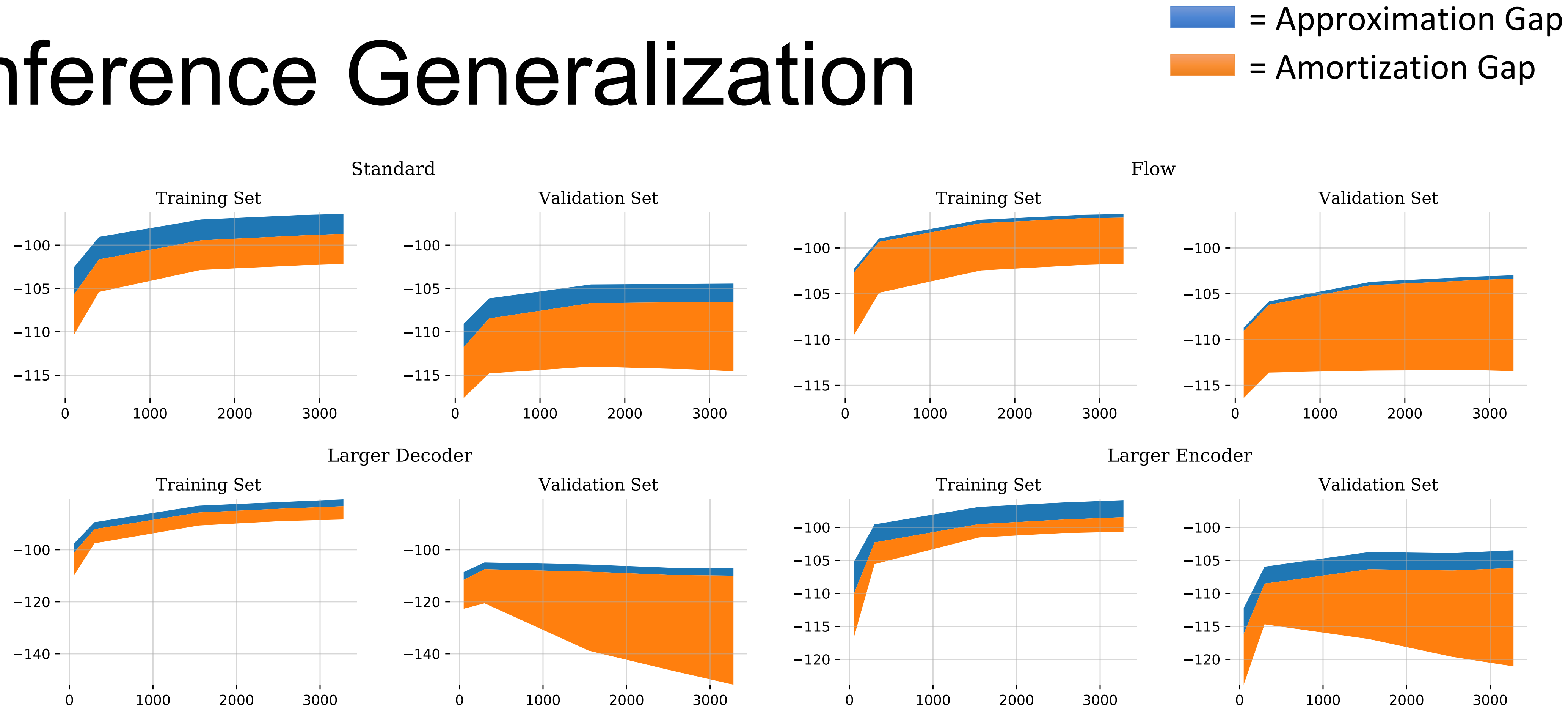
Larger Decoder Capacity Reduces Approximation Gap

- Does increasing decoder capacity decrease the approximation gap?
  - Does a more powerful decoder make the true posterior easier to model with the choice of approximation?

- Experiment:
  - Train VAEs with decoders that have 0, 2, 4 hidden layers
  - Compute the approximation gaps (ie. How Gaussian is $pz\square x$?)

$Generator\ Hidden\ Layers$	0	2	4
$Approximation\ Gap$	3.90	1.83	1.63

Inference Generalization



- From training to validation set:
  - amortization gap increases
  - approximation gap constant
- Increased capacity: more prone to overfitting but better inference
- Flow improves model while overfitting less

Summary

- Inference Gap: Amortization vs Approximation
- Amortization > Approximation
- Generator accommodates approximation
- Inform model design choices

Poster: Hall B #176

Thanks

Experiments

- Encoder Capacity
- Decoder Capacity
- Variational Distribution

Dataset Models

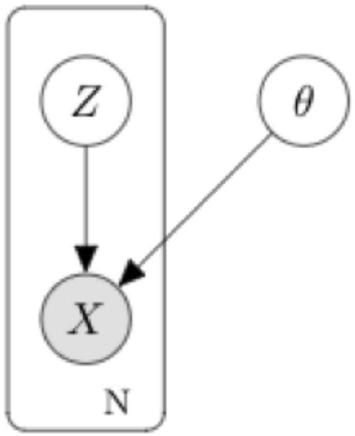
MNIST/Fashion	**3-BIT CIFAR-10**
784-200-200-50	Conv-Conv-Conv-FC
50-200-200-784	FC-ConvT-ConvT-ConvT

VAE

- Latent Variable Model
- Amortized Inference

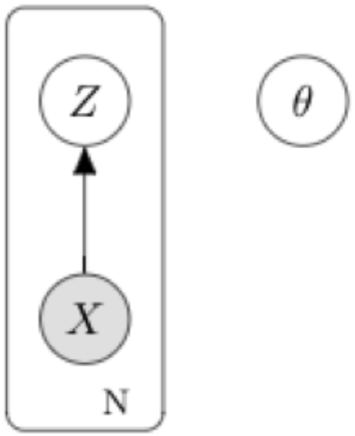
- Inference Suboptimality

Generative Model



$p_\theta(x|z)\,p(z)$
Generator
(Decoder)

Inference Model



$q_\phi(z|x)$
Recognition Net
(Encoder)

$$\log p(x) = \mathbb{E}_{z\sim q(z|x)}\left[\log\left(\frac{p(x,z)}{q(z|x)}\right)\right] + \mathbf{KL}\left(q(z|x)||p(z|x)\right)$$

ELBO

Inference Gap