AMORTIZED INFERENCE IN DEEP GENERATIVE MODELS

JOSEPH MARINO

APRIL 30TH, 2020 - CALTECH CS 159

OUTLINE

Sec. 1: background

Sec. 2: variational autoencoders

Sec. 3: iterative amortized inference

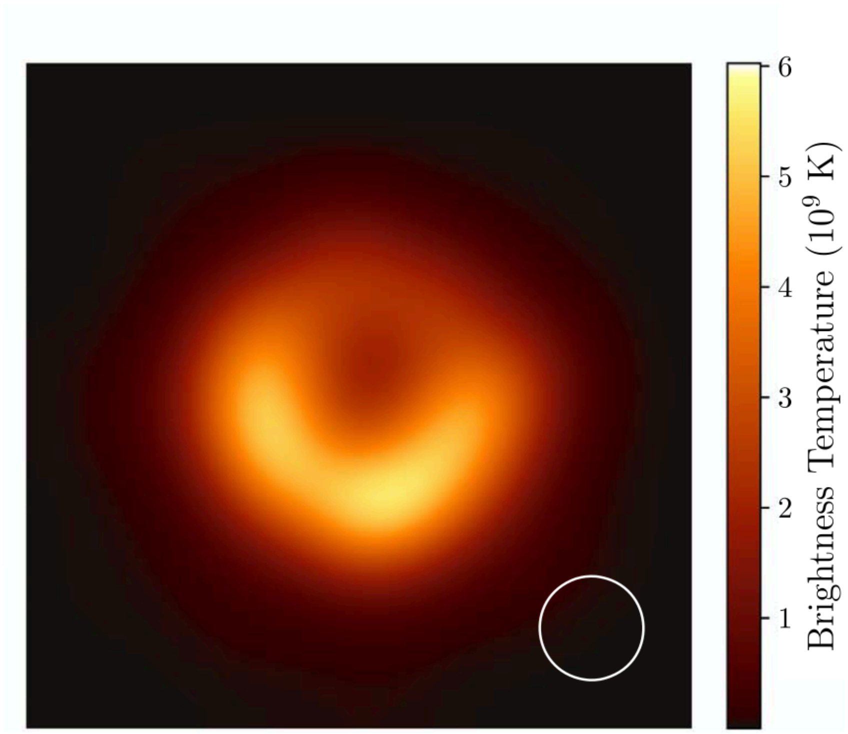
Sec. 4: closing remarks

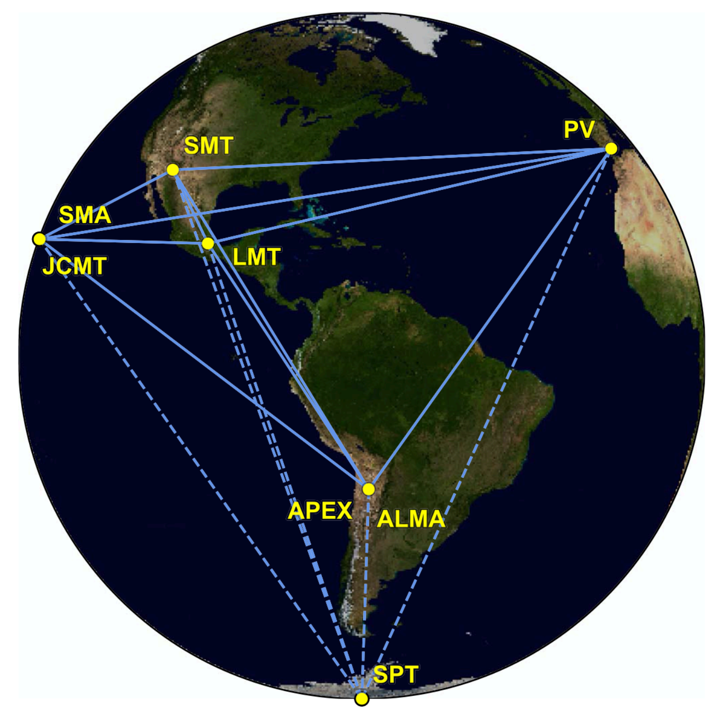
BACKGROUND

FORWARD MODELS & INVERSE PROBLEMS

forward model: model of how an observation is generated

black hole brightness → radio Fourier frequencies



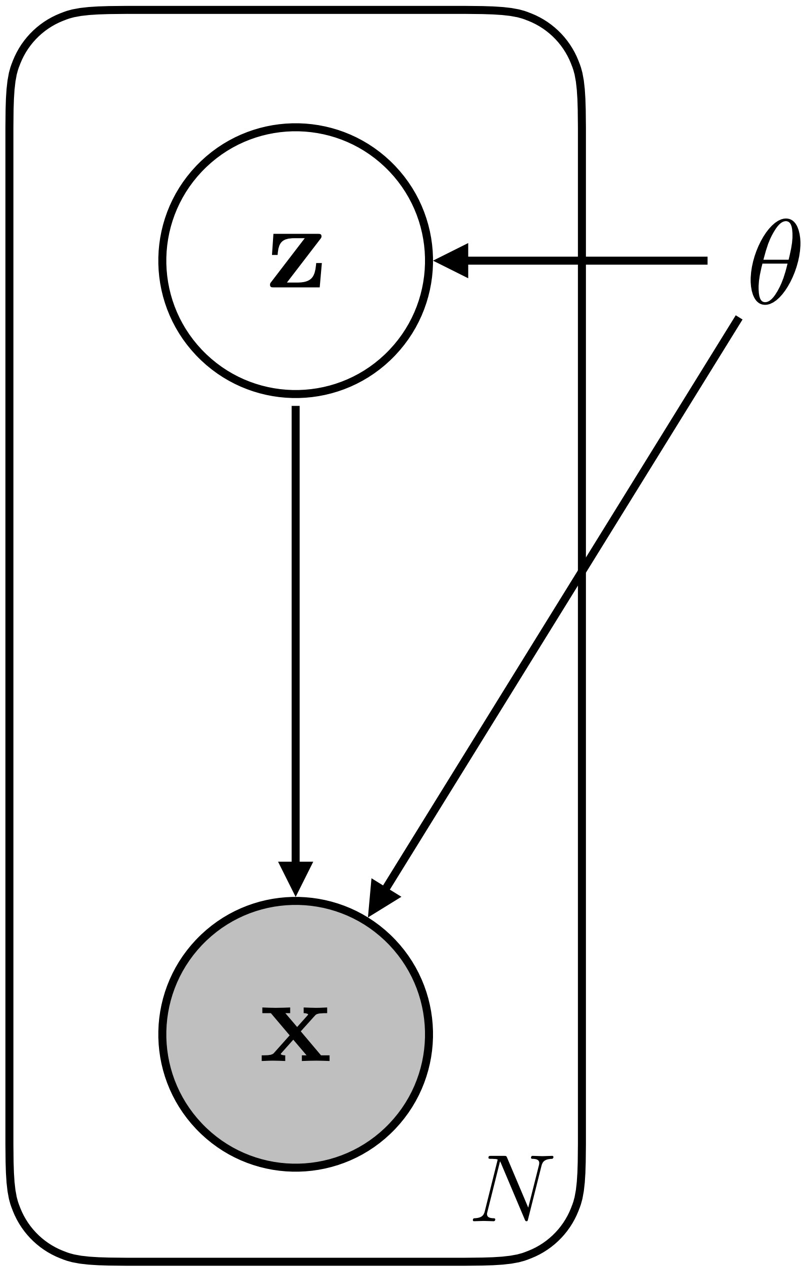


black hole brightness ← radio Fourier frequencies

inverse problem: inverting a forward model to obtain the underlying state

Event Horizon Telescope Collaboration, 2019

LATENT VARIABLE MODELS



model:

$$\underbrace{p_\theta(\mathbf{x}, \mathbf{z})}_{\textit{joint}} = \underbrace{p_\theta(\mathbf{x}|\mathbf{z})}_{\substack{\textit{conditional}\\ \textit{likelihood}}}\underbrace{p_\theta(\mathbf{z})}_{\textit{prior}}$$

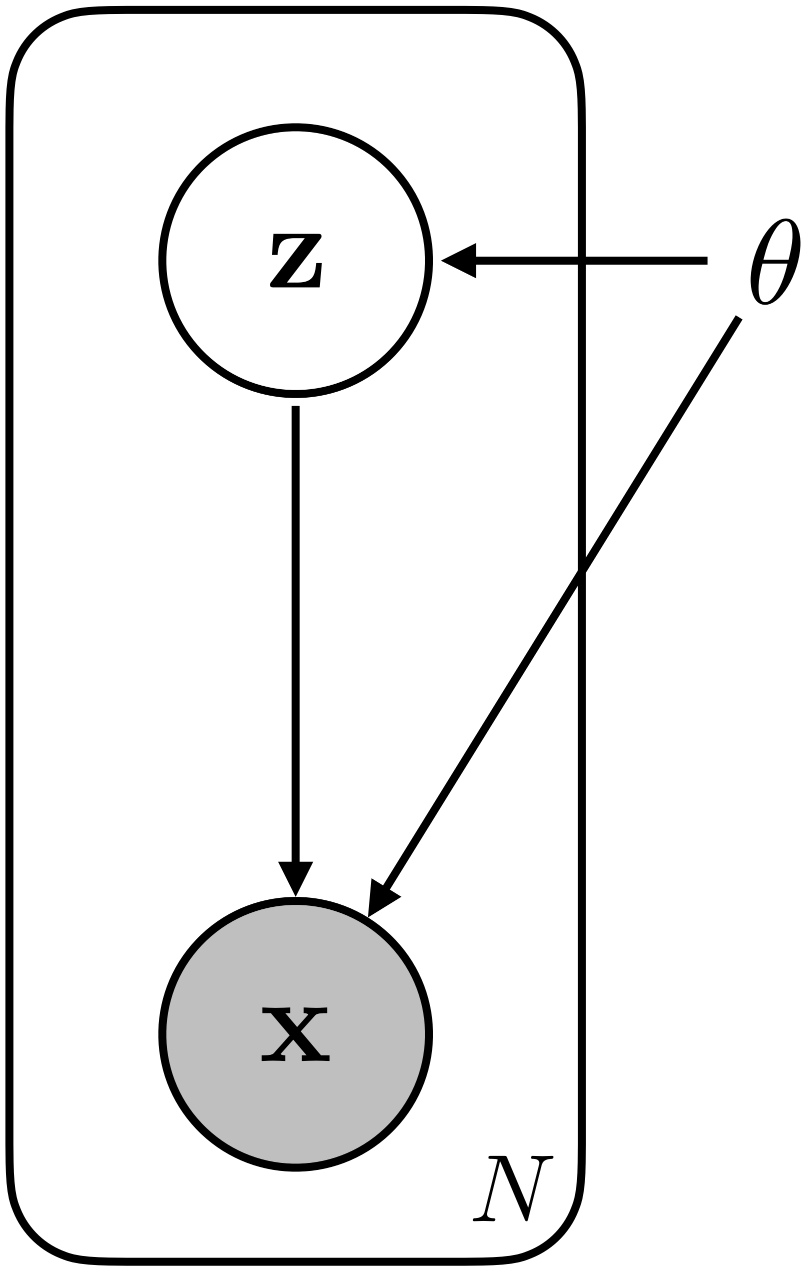
marginalization:

$$\underbrace{p_\theta(\mathbf{x})}_{\substack{\textit{marginal}\\ \textit{likelihood}}} = \int p_\theta(\mathbf{x}, \mathbf{z}) d\mathbf{z}$$

inference:

$$\underbrace{p_\theta(\mathbf{z}|\mathbf{x})}_{\textit{posterior}} = \frac{p_\theta(\mathbf{x}, \mathbf{z})}{p_\theta(\mathbf{x})}$$

LATENT VARIABLE MODELS



maximum likelihood is typically intractable

$$\theta^* = \arg\max_\theta \mathbb{E}_{p_{\text{data}}(\mathbf{x})} \left[\log p_\theta(\mathbf{x})\right]$$

$$\approx \arg\max_\theta \frac{1}{N} \sum_{i=1}^{N} \log p_\theta(\mathbf{x}^{(i)})$$

$$\approx \arg\max_\theta \frac{1}{N} \sum_{i=1}^{N} \log \underbrace{\left[\int p_\theta(\mathbf{x}^{(i)}, \mathbf{z}) d\mathbf{z}\right]}_{\text{intractable integral}}$$

must resort to approximation techniques

VARIATIONAL INFERENCE

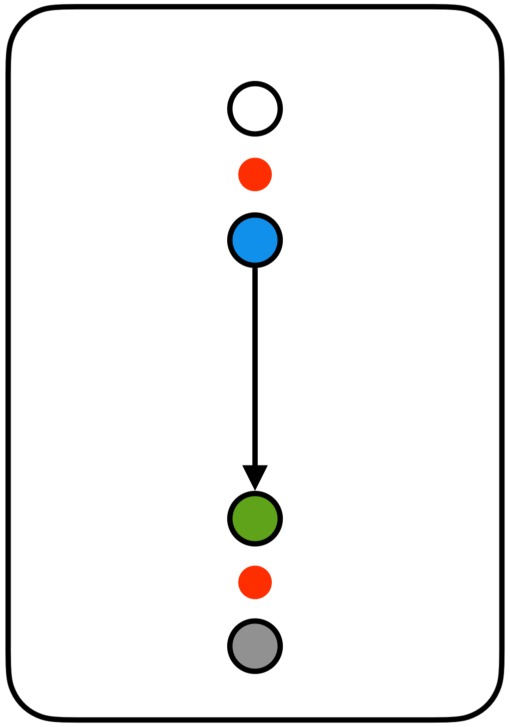
approximate posterior $q(\mathbf{z}|\mathbf{x})$

variational lower bound

$$\log p_\theta(\mathbf{x}) \geq \mathcal{L}(\mathbf{x}; q)$$

where

$$\mathcal{L}(\mathbf{x}; q) = \mathbb{E}_q \left[\underbrace{\log p_\theta(\mathbf{x}|\mathbf{z})}_{\text{"reconstruction"}} - \underbrace{\log \frac{q(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z})}}_{\text{"regularization"}} \right]$$



VARIATIONAL INFERENCE

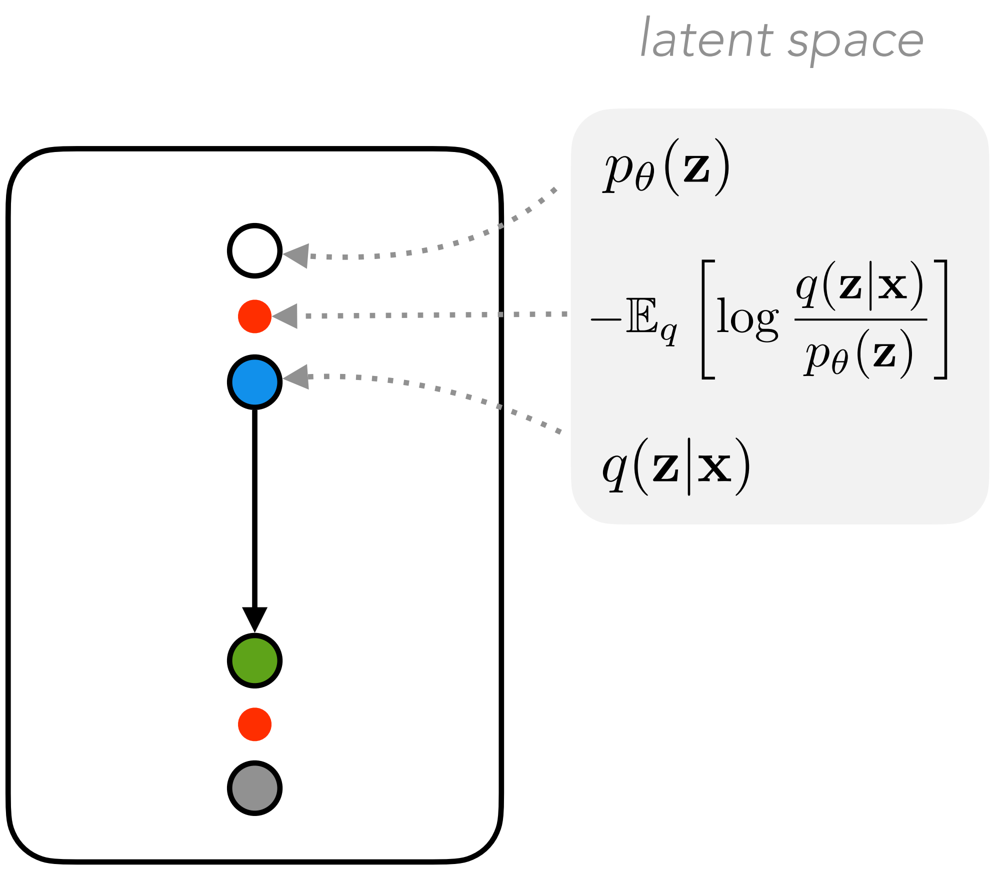
approximate posterior $q(\mathbf{z}|\mathbf{x})$

variational lower bound

$$\log p_\theta(\mathbf{x}) \geq \mathcal{L}(\mathbf{x}; q)$$

where

$$\mathcal{L}(\mathbf{x}; q) = \mathbb{E}_q \left[\underbrace{\log p_\theta(\mathbf{x}|\mathbf{z})}_{\text{"reconstruction"}} - \underbrace{\log \frac{q(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z})}}_{\text{"regularization"}} \right]$$



VARIATIONAL INFERENCE

approximate posterior $q(\mathbf{z}|\mathbf{x})$

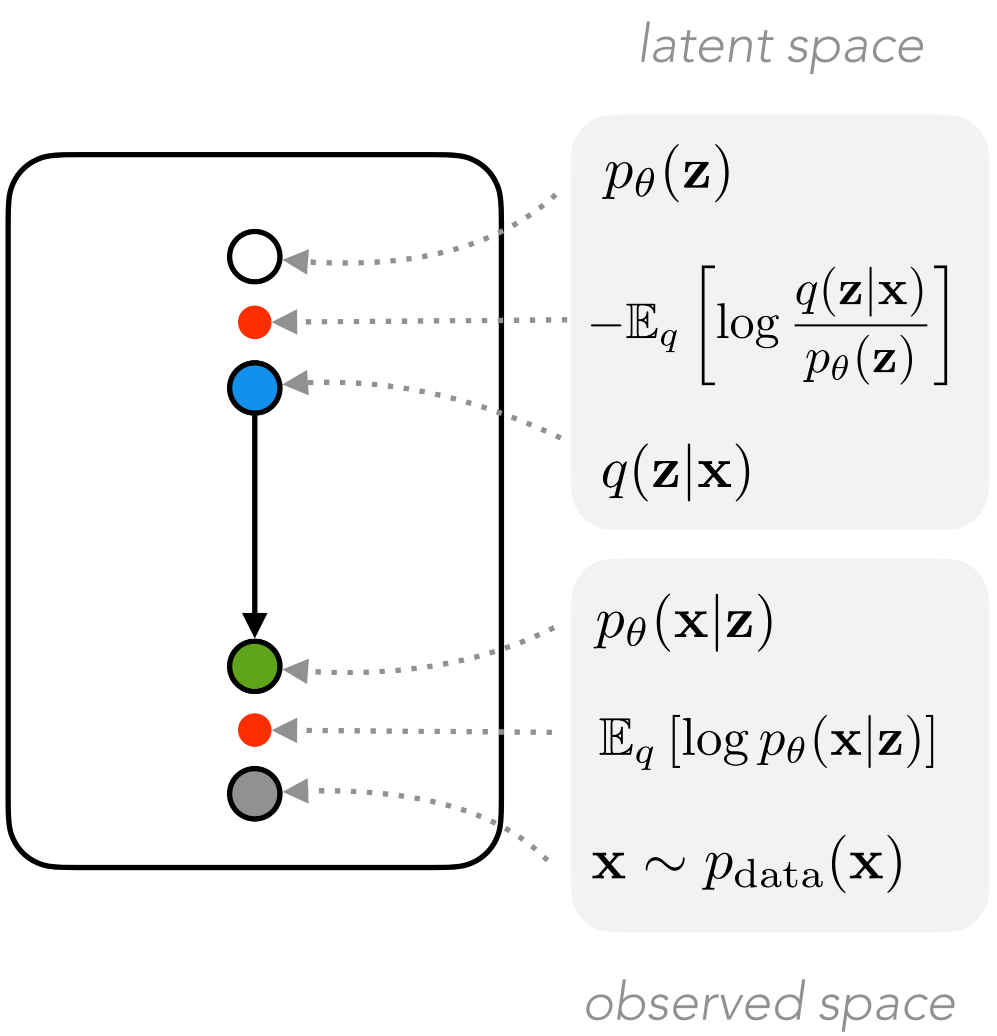
---

variational lower bound

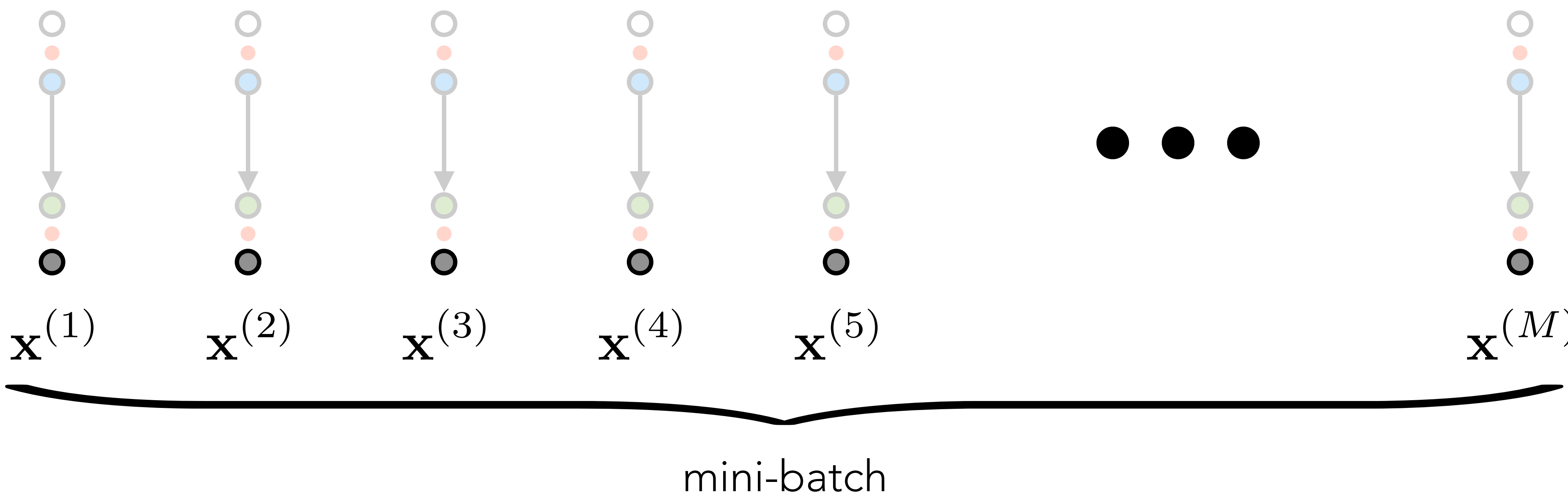
$$\log p_\theta(\mathbf{x}) \geq \mathcal{L}(\mathbf{x}; q)$$

where

$$\mathcal{L}(\mathbf{x}; q) = \mathbb{E}_q \left[\underbrace{\log p_\theta(\mathbf{x}|\mathbf{z})}_{\text{“reconstruction”}} - \underbrace{\log \frac{q(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z})}}_{\text{“regularization”}} \right]$$



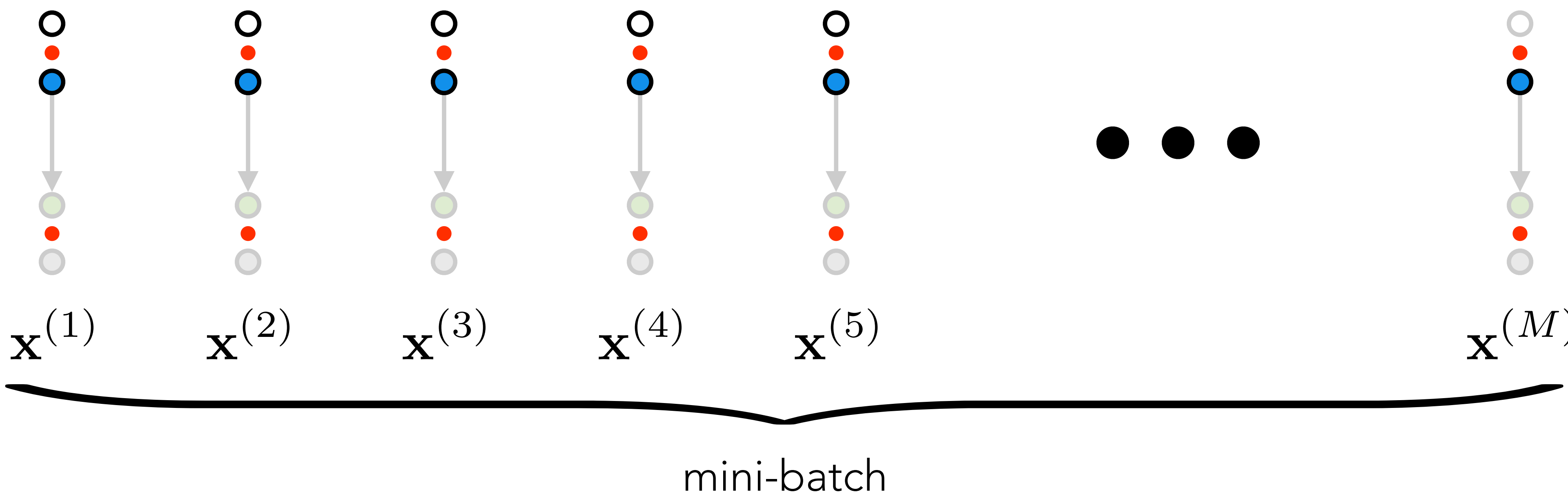
VARIATIONAL EXPECTATION MAXIMIZATION



Variational EM (single-step)

$\texttt{sample}\ \mathbf{x}^{(1:M)} \sim p_{\text{data}}(\mathbf{x})$

VARIATIONAL EXPECTATION MAXIMIZATION



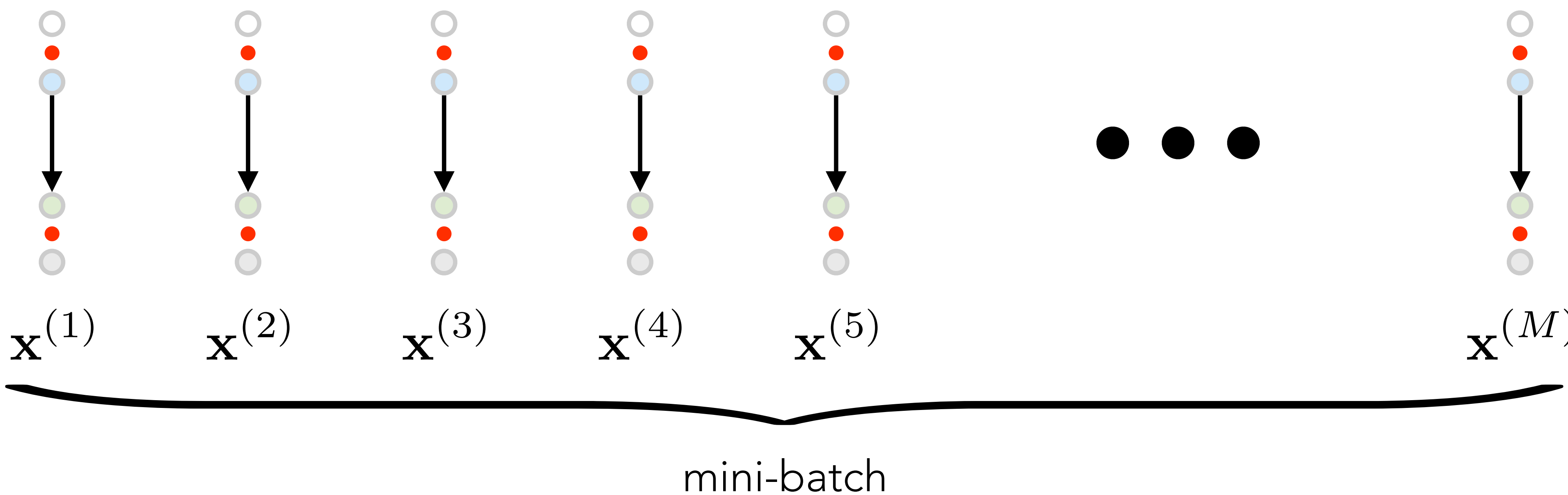
Variational EM (single-step)

`sample` $\mathbf{x}^{(1:M)} \sim p_{\text{data}}(\mathbf{x})$

`for` $\mathbf{x}^{(i)}$ `in` $\mathbf{x}^{(1:M)}$:

    `maximize` $\mathcal{L}(\mathbf{x}^{(i)}, q^{(i)})$ `w.r.t.` $q^{(i)}$ *# E-step*

VARIATIONAL EXPECTATION MAXIMIZATION



Variational EM (single-step)

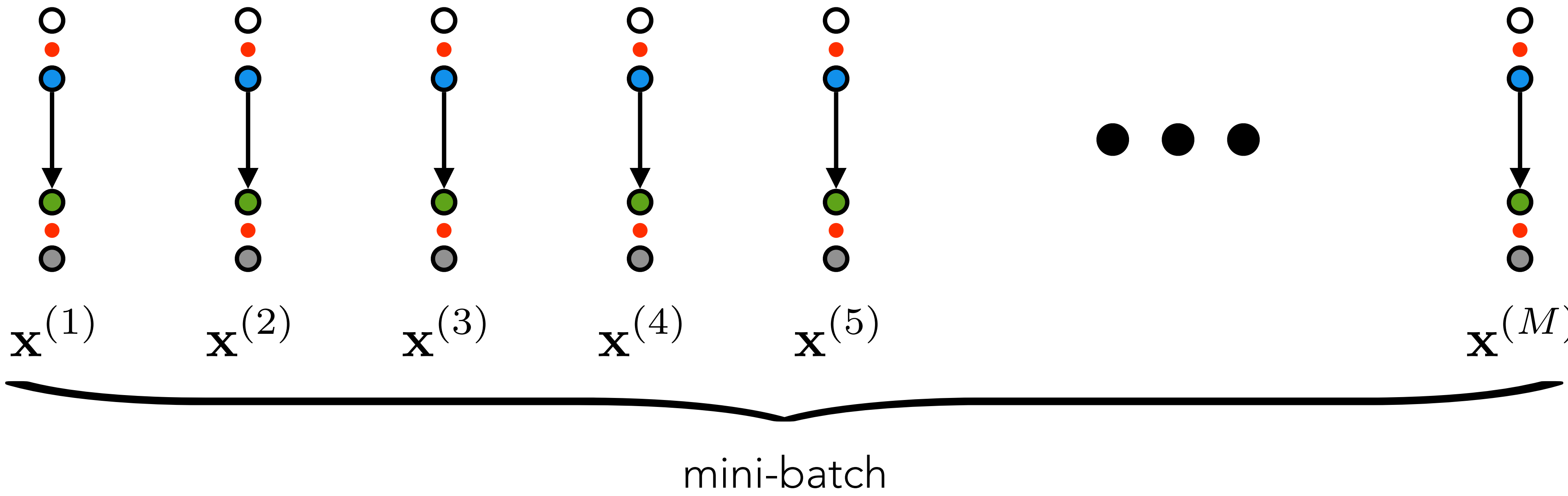
sample $\mathbf{x}^{(1:M)} \sim p_{\text{data}}(\mathbf{x})$

for $\mathbf{x}^{(i)}$ in $\mathbf{x}^{(1:M)}$:

maximize $\mathcal{L}(\mathbf{x}^{(i)}, q^{(i)})$ w.r.t. $q^{(i)}$ # E-step

maximize $\frac{1}{M}\sum_{i=1}^{M}\mathcal{L}(\mathbf{x}^{(i)}, q^{(i)})$ w.r.t. θ # M-step

VARIATIONAL EXPECTATION MAXIMIZATION



Variational EM (single-step)

sample $\mathbf{x}^{(1:M)} \sim p_{\text{data}}(\mathbf{x})$

for $\mathbf{x}^{(i)}$ in $\mathbf{x}^{(1:M)}$:

expensive

  maximize $\mathcal{L}(\mathbf{x}^{(i)}, q^{(i)})$ w.r.t. $q^{(i)}$ *# E-step*

maximize $\frac{1}{M} \sum_{i=1}^{M} \mathcal{L}(\mathbf{x}^{(i)}, q^{(i)})$ w.r.t. θ *# M-step*

AUTOENCODERS

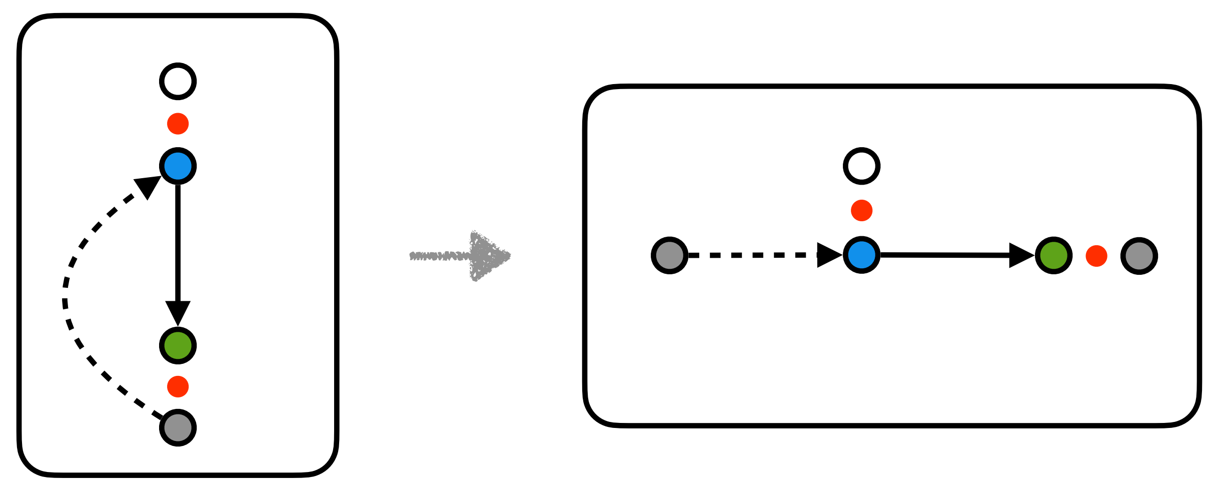
AMORTIZED OPTIMIZATION

<u>spreading</u> the cost of optimization over multiple runs using a learned model

AUTOENCODER *(self-encoder)*

<u>learn</u> to directly estimate $q(\mathbf{z}|\mathbf{x})$ as a conditional mapping from $\mathbf{x}$

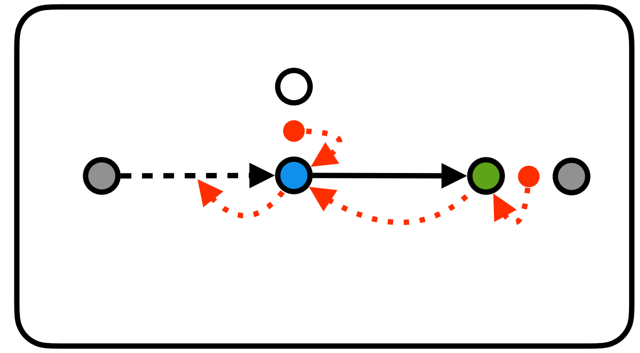
refer to this mapping as $q_\phi(\mathbf{z}|\mathbf{x})$



learning to infer/estimate $q(\mathbf{z}|\mathbf{x})$ *is a form of* $\underbrace{\textit{meta-optimization!}}_{\text{this class}}$

STOCHASTIC GRADIENT ESTIMATION

to learn an *encoder*, we need some way of estimating $\nabla_\phi \mathcal{L}$



variational lower bound: $\mathcal{L}(\mathbf{x}; q) = \mathbb{E}_q \left[\log p_\theta(\mathbf{x}|\mathbf{z}) - \log \frac{q(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z})} \right]$

variational inference optimization requires ***stochastic gradient estimation***

$$\nabla_\phi \mathbb{E}_{q_\phi(\mathbf{z})} [f(\mathbf{z})] \qquad \phi \longrightarrow \mathbf{z} \longrightarrow \boxed{f}$$

some options:

- **point estimate**: let $q_\phi(\mathbf{z}) = \delta(\mathbf{z} = \hat{\mathbf{z}}_\phi)$, yielding $\nabla_\phi \mathbb{E}_{q_\phi(\mathbf{z})} [f(\mathbf{z})] = \nabla_\phi f(\hat{\mathbf{z}}_\phi)$

  inexpressive

- **score function/REINFORCE**: use any distribution, $\nabla_\phi \mathbb{E}_{q_\phi(\mathbf{z})} [f(\mathbf{z})] = \mathbb{E}_{q_\phi(\mathbf{z})} [f(\mathbf{z}) \nabla_\phi \log q_\phi(\mathbf{z})]$

  high variance, *but see NVIL, DARN, MuProp, VIMCO, etc.*

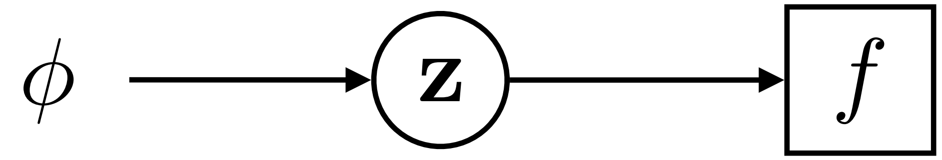
Schulman *et al.*, 2016

VARIATIONAL AUTOENCODERS

REPARAMETERIZATION TRICK

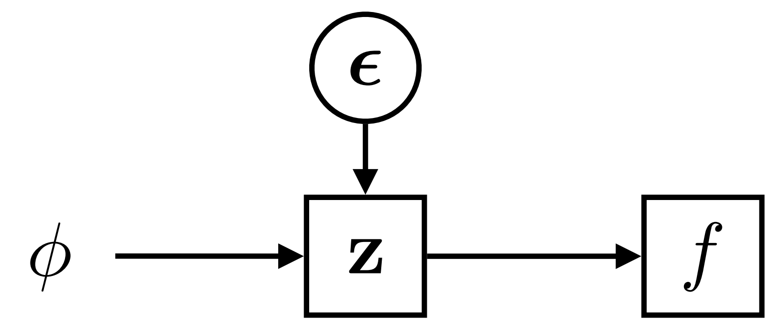
variational inference optimization requires ***stochastic gradient estimation***

$$\nabla_\phi \mathbb{E}_{q_\phi(\mathbf{z})}\left[f(\mathbf{z})\right]$$



reparameterization trick / pathwise derivative:

reparameterize $q_\phi(\mathbf{z})$ as a deterministic function of an auxiliary variable $\boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon})$, i.e. $\mathbf{z} = g_\phi(\boldsymbol{\epsilon})$



the gradient can then be expressed as $\nabla_\phi \mathbb{E}_{q_\phi(\mathbf{z})}\left[f(\mathbf{z})\right] = \mathbb{E}_{p(\boldsymbol{\epsilon})}\left[\nabla_\phi f(g_\phi(\boldsymbol{\epsilon}))\right]$

canonical example:

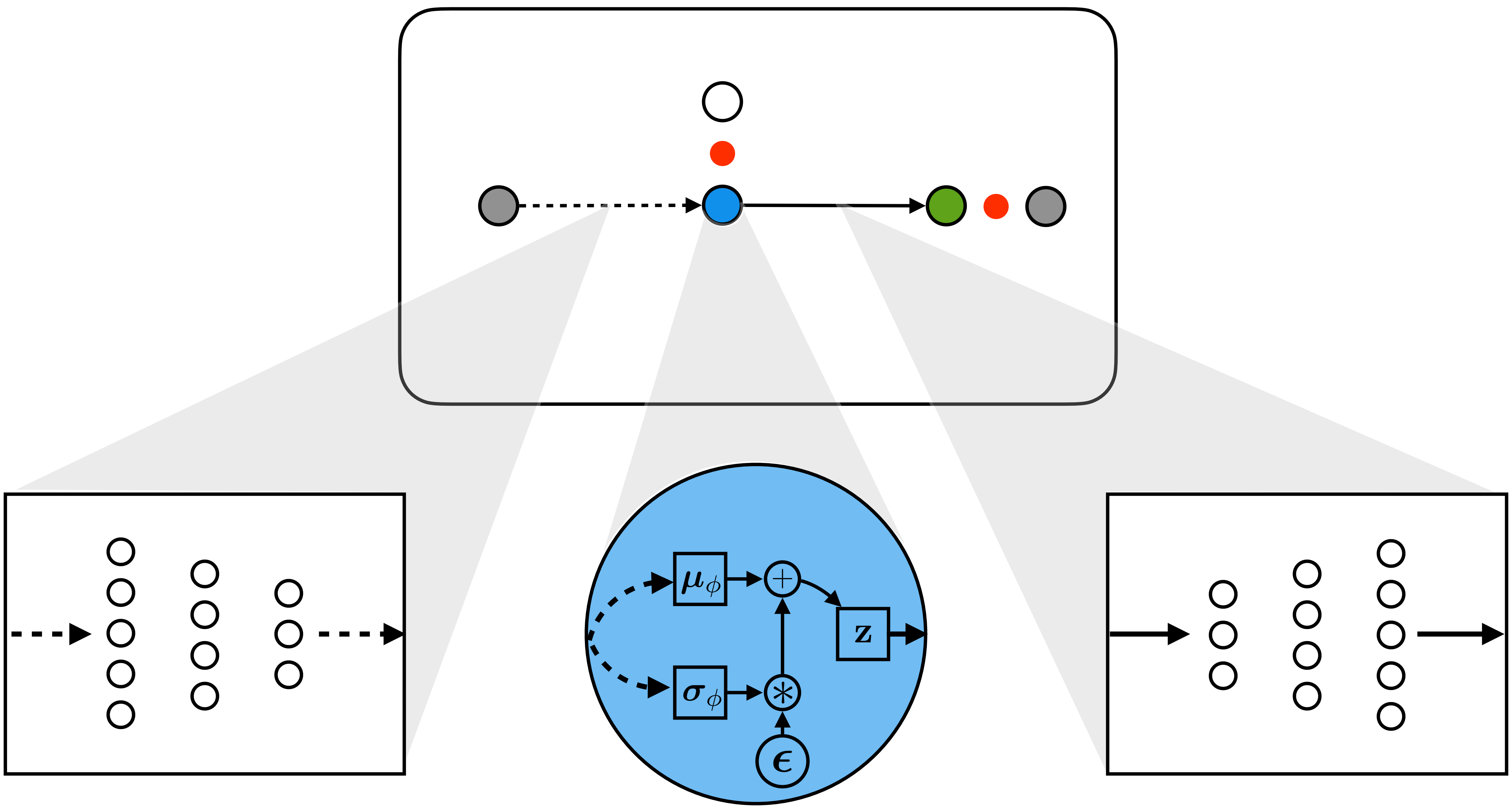
$$\mathbf{z} \sim \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\sigma}^2)) \xrightarrow{\textit{reparameterize}} \begin{array}{c} \mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}; \mathbf{0}, \mathbf{I}) \end{array}$$

Schulman *et al.*, 2016

VARIATIONAL AUTOENCODERS

Variational Autoencoder (VAE):

deep latent variable model + variational inference + direct encoder + *reparameterized* Gaussian



Kingma & Welling, 2014

Rezende et al., 2014

VARIATIONAL AUTOENCODERS

improving sample quality:

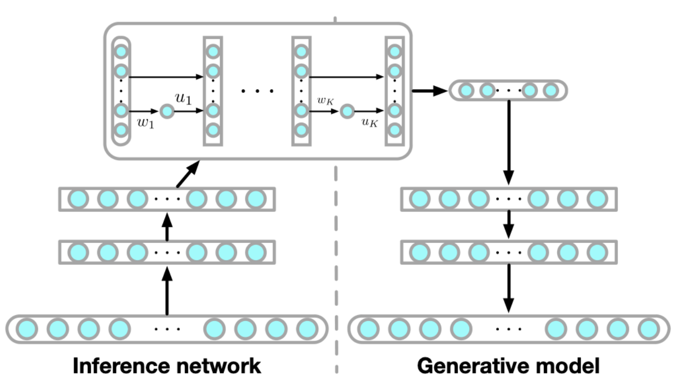
Rezende et al., 2014

Kingma et al., 2016

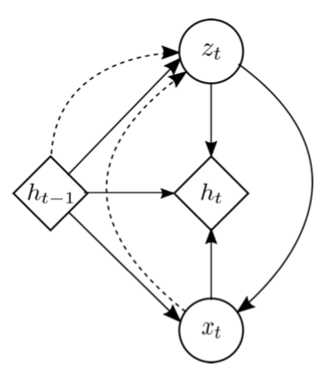
Maaløe et al., 2019

Razavi et al., 2019

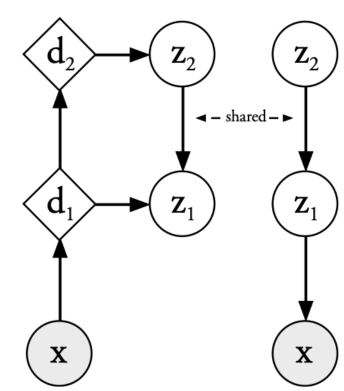
technical innovations:



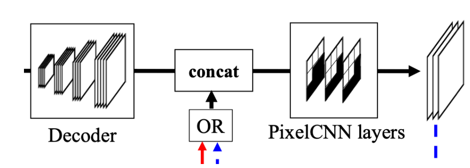
Normalizing Flows
Rezende & Mohamed, 2015



Dynamical Models
Chung et al., 2015



Hierarchy / Top-Down Inference
Sønderby et al., 2016



Autoregressive Decoder
Gulrajani et al., 2017

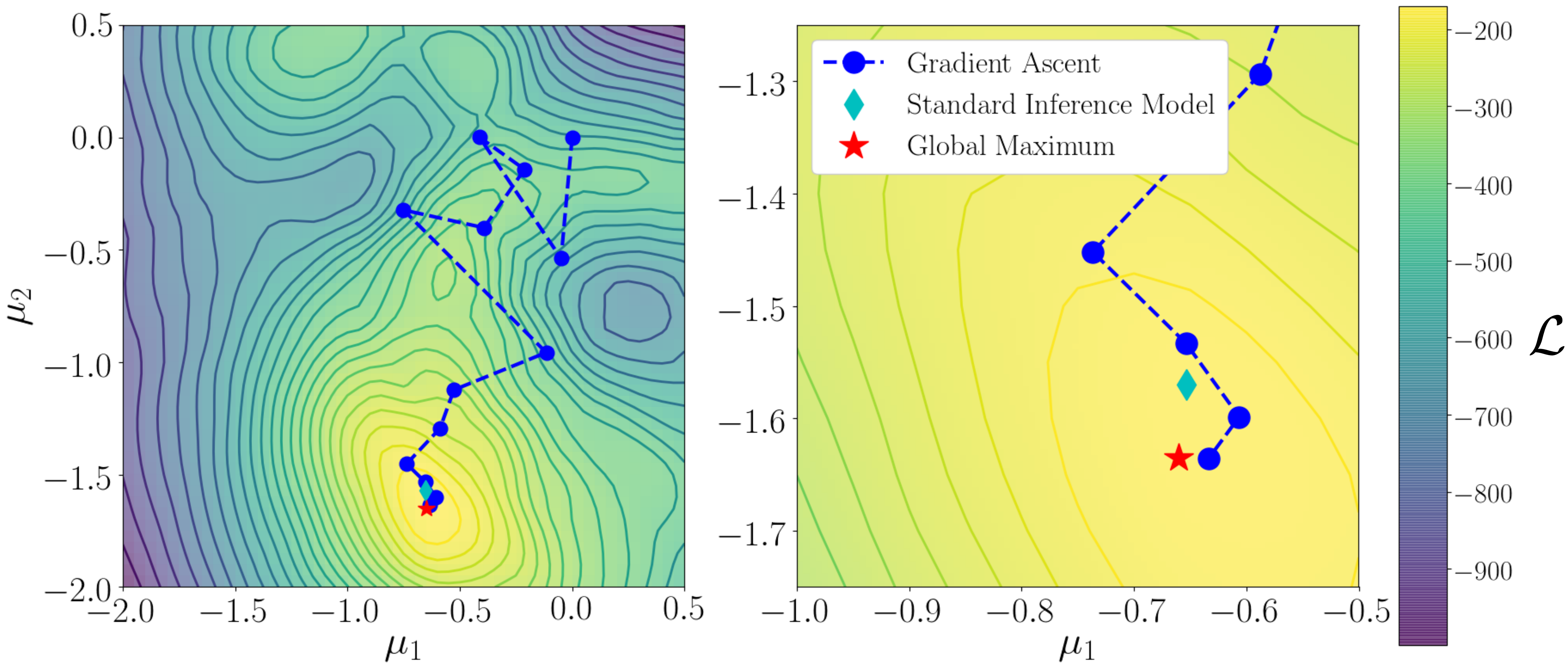
AMORTIZATION GAP

amortized inference may fail to estimate the optimal distribution

Cremer *et al.*, 2018

AMORTIZATION GAP

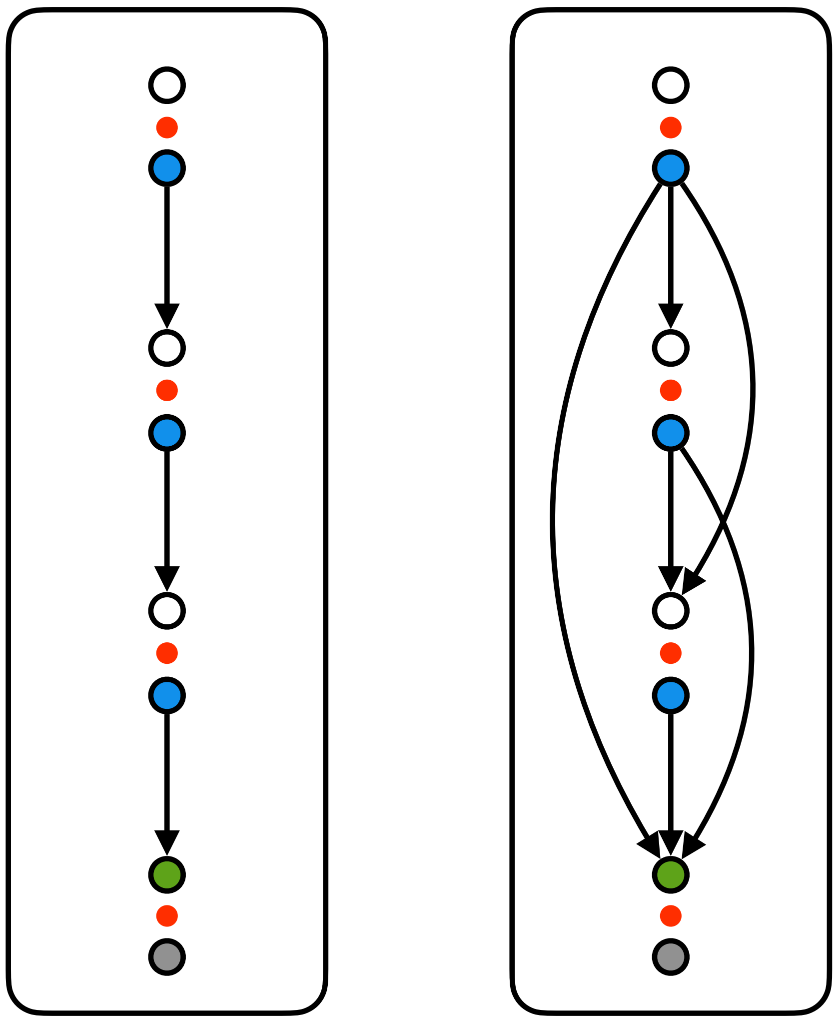
can also visualize the gap in terms of the variational optimization surface



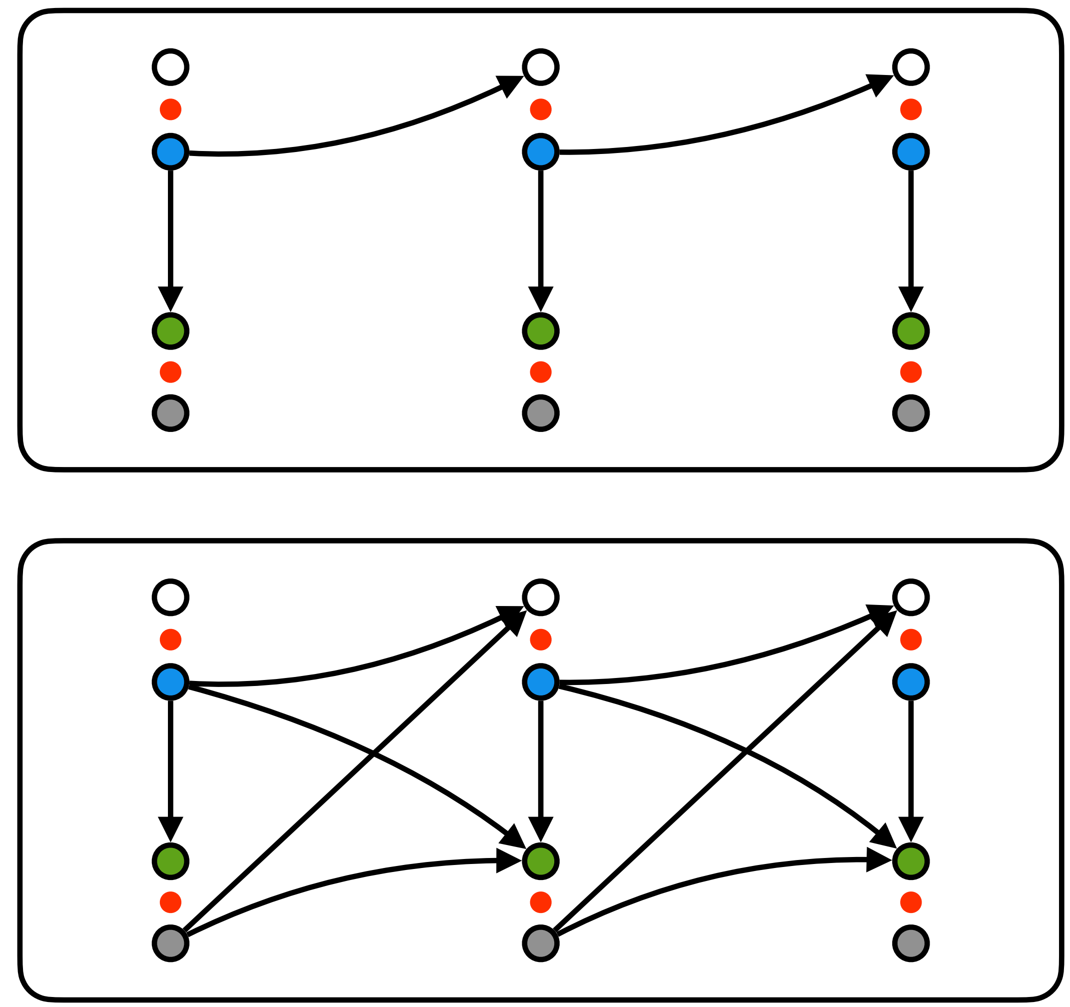
Marino *et al.*, 2018a

STRUCTURED APPROXIMATE POSTERIORS

structured models



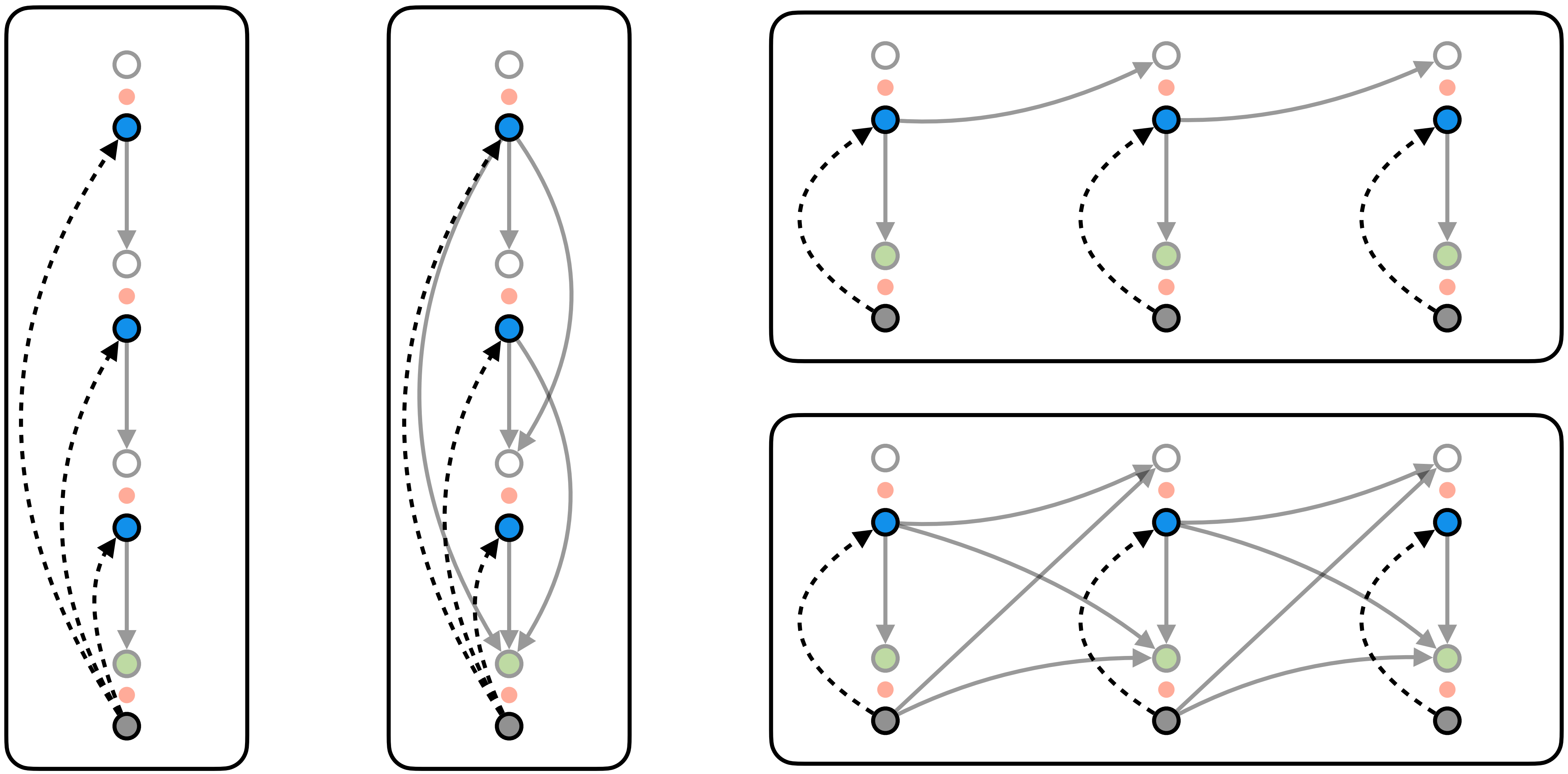
hierarchical



dynamical

STRUCTURED APPROXIMATE POSTERIORS

structured models



hierarchical

dynamical

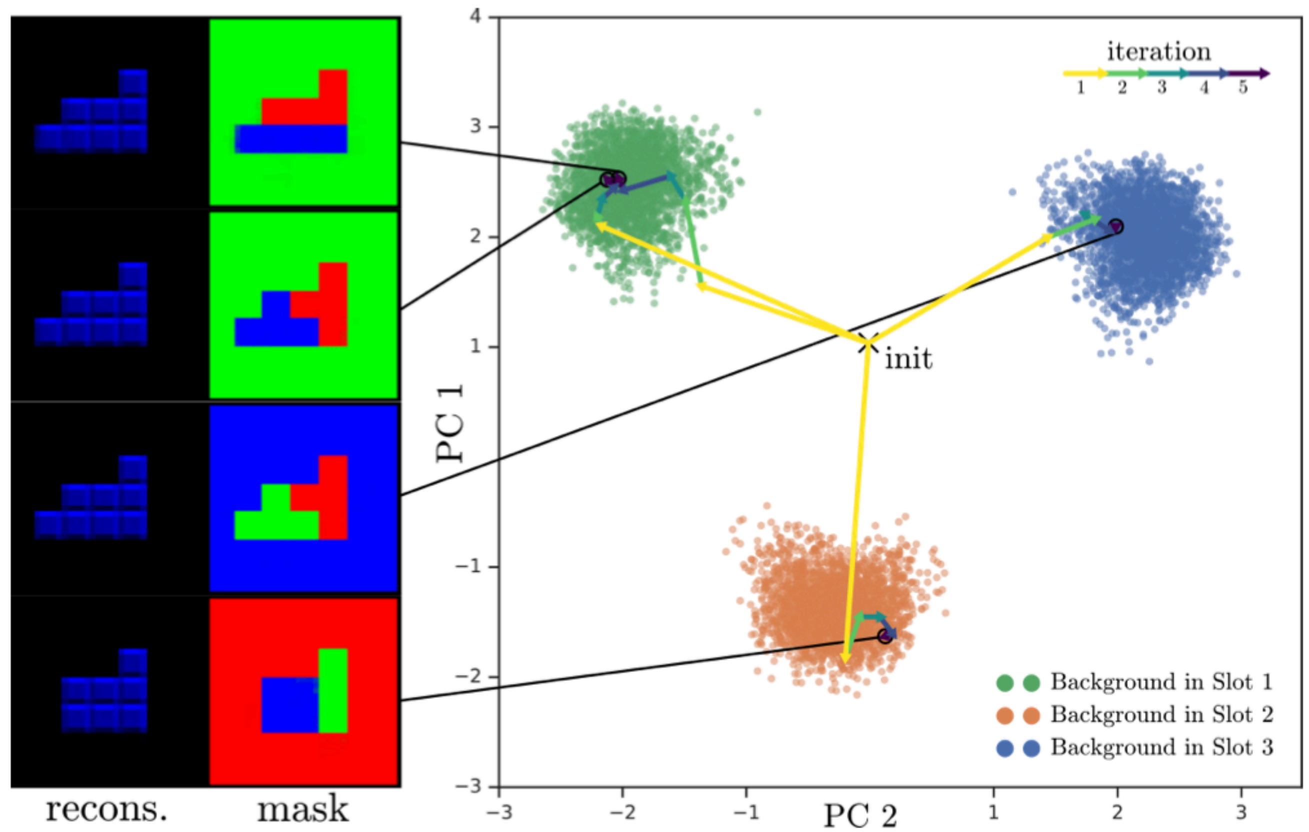
(naïve) direct encoders cannot account for structured estimates

$\mathbf{z}_k$ depends on $\mathbf{z}_{<k}$, but $q_\phi(\mathbf{z}_k|\mathbf{x})$ does not have access to this information

SYMMETRY BREAKING / MULTIPLE OPTIMA

there may be multiple equally valid estimates,

however, a direct mapping from $\mathbf{x} \rightarrow q_\phi(\mathbf{z}|\mathbf{x})$ can only provide a single estimate



example: multiple ways to parse an image into separate objects

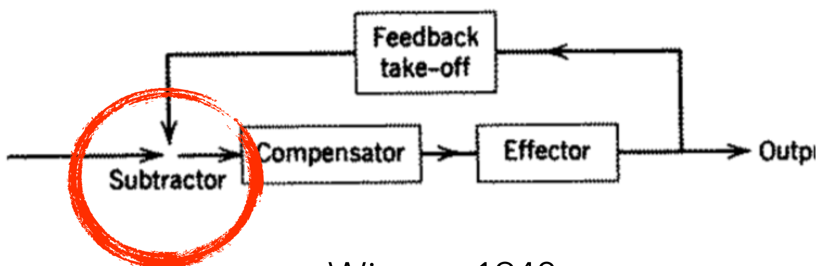
Greff et al., 2019

ITERATIVE AMORTIZED INFERENCE

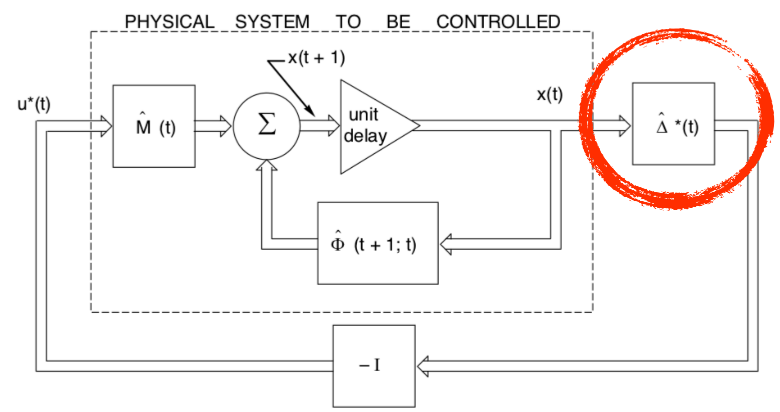
motivation

can we improve inference optimization while retaining the benefits of amortization?

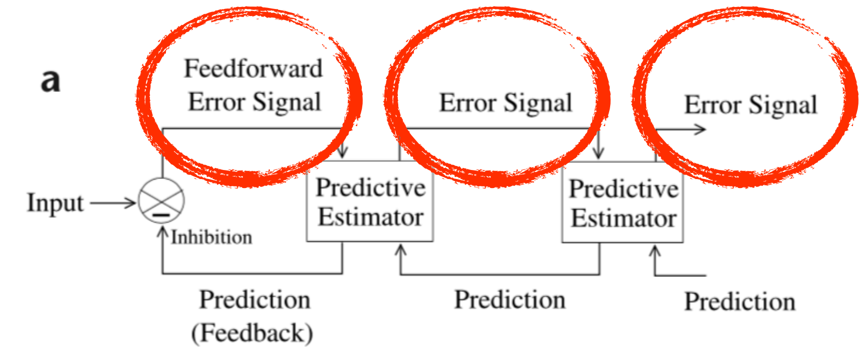
STATE ESTIMATION USING ERRORS



Wiener, 1948



Kalman, 1960



Rao & Ballard, 1999

control

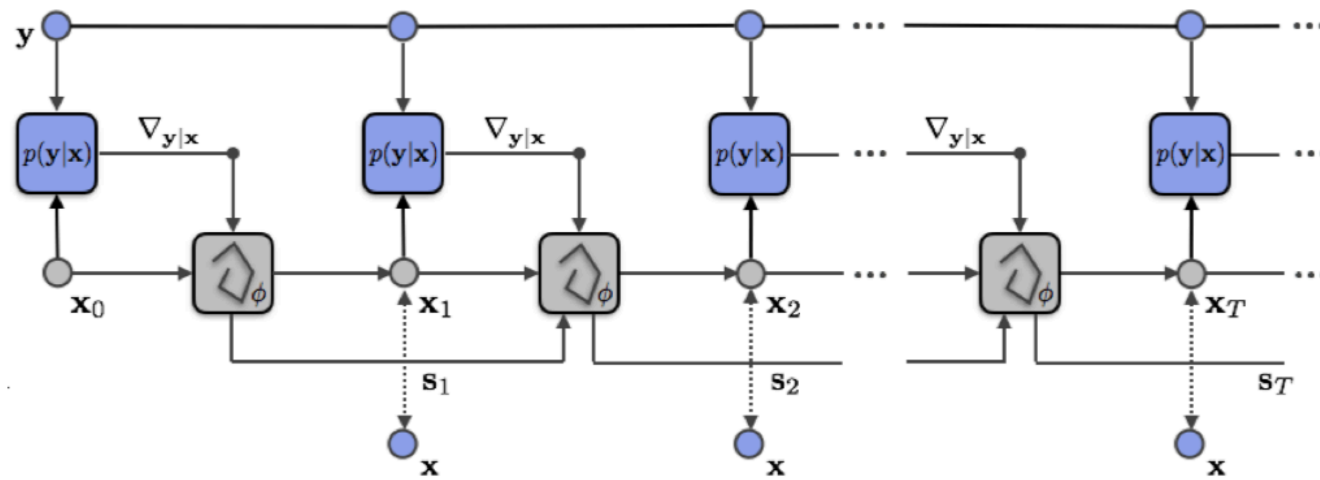
control/state-estimation

state-estimation

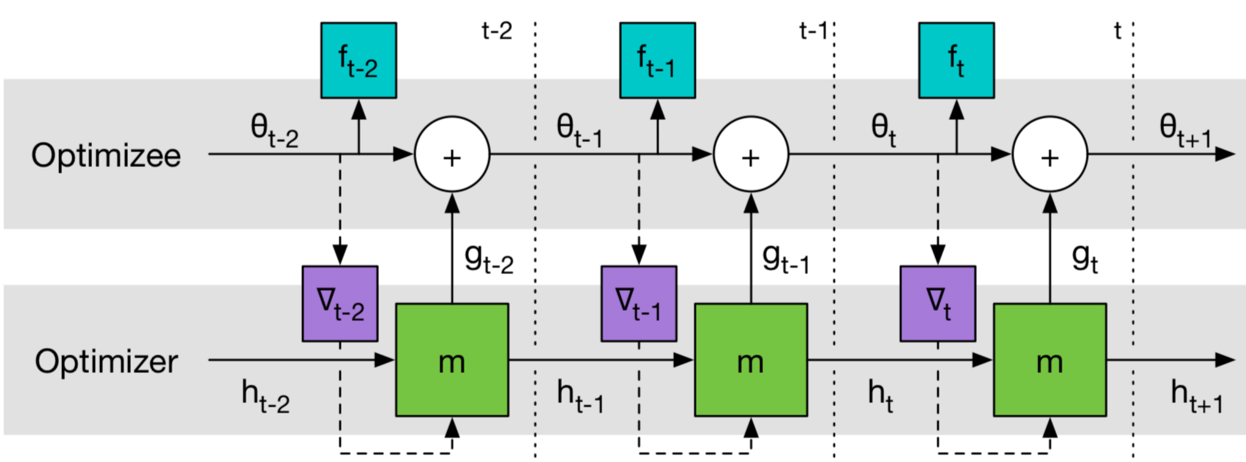
classical control/state estimation techniques

perform *iterative estimation* using **error** signals

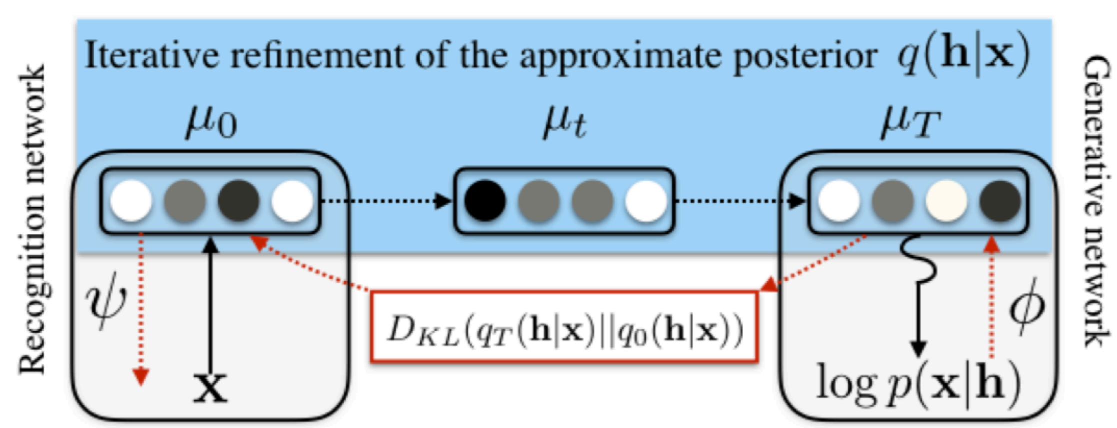
RELATED WORK



Recurrent Inference Machines
Putzky & Welling, 2017



Learning to Learn GD by GD
Andrychowicz *et al.*, 2016

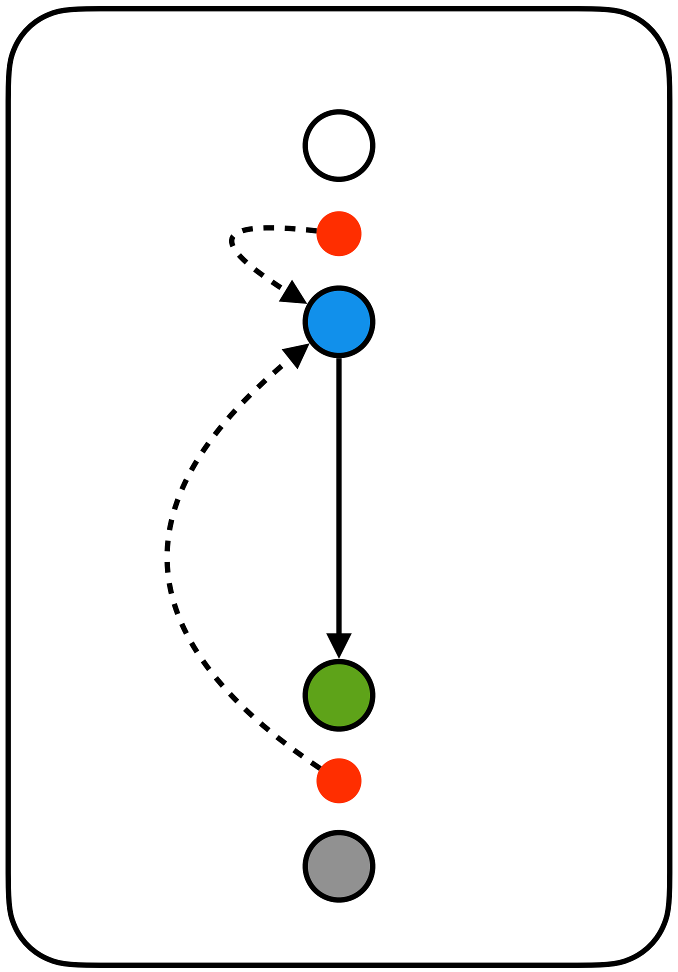


Initial Encoding, Iterative Refinement
Krishnan *et al.*, 2018
Hjelm *et al.*, 2016
Kim *et al.*, 2018

ITERATIVE AMORTIZED INFERENCE

let $\boldsymbol{\lambda}$ be the distribution parameters of $q_\phi(\mathbf{z}|\mathbf{x})$

i.e. $\boldsymbol{\lambda} \equiv \{\boldsymbol{\mu}, \boldsymbol{\sigma}\}$



iterative inference models:

gradient encoding: $\boldsymbol{\lambda} \leftarrow f_\phi(\boldsymbol{\lambda}, \nabla_{\boldsymbol{\lambda}} \mathcal{L})$

error encoding: $\boldsymbol{\lambda} \leftarrow f_\phi(\boldsymbol{\lambda}, \boldsymbol{\varepsilon}_{\mathbf{x}}, \boldsymbol{\varepsilon}_{\mathbf{z}})$

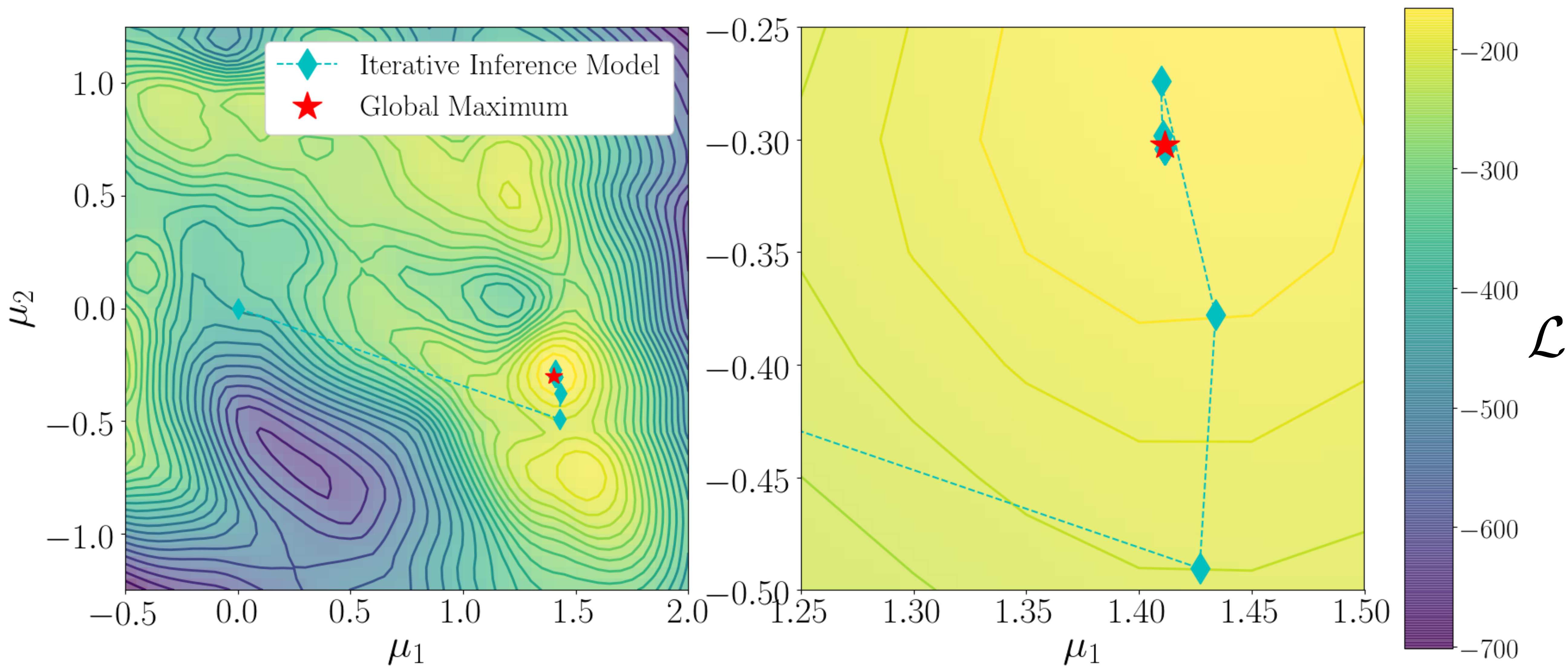
where $\boldsymbol{\varepsilon}_{\mathbf{x}} \equiv \dfrac{\boldsymbol{\mu}_{\mathbf{x}}(\mathbf{z}) - \mathbf{x}}{\boldsymbol{\sigma}_{\mathbf{x}}}$ $\qquad$ $\boldsymbol{\varepsilon}_{\mathbf{z}} \equiv \dfrac{\boldsymbol{\mu}_{\mathbf{z}} - \mathbf{z}}{\boldsymbol{\sigma}_{\mathbf{z}}}$

reconstruction error $\qquad$ latent error

INFERENCE OPTIMIZATION

directly visualize inference in the optimization landscape

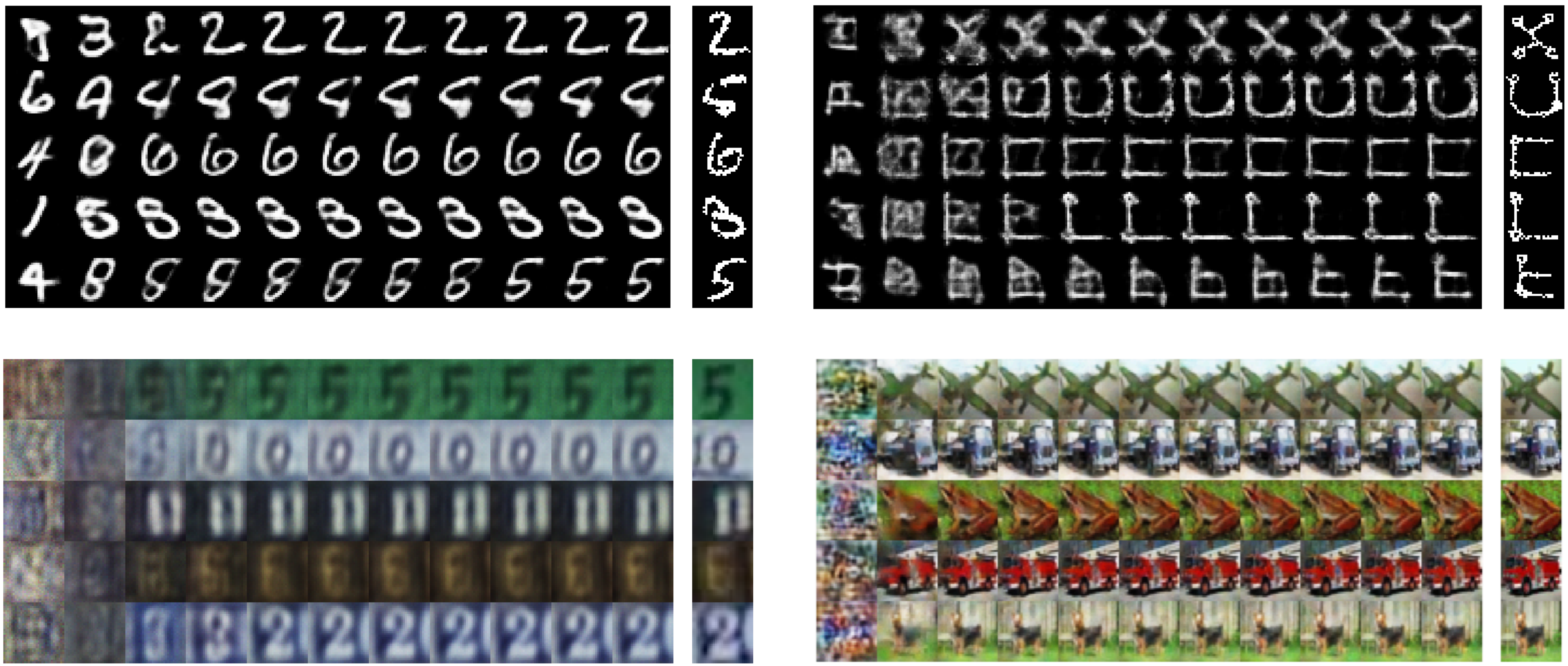
2D model, MNIST



Marino *et al.*, 2018a

INFERENCE OPTIMIZATION

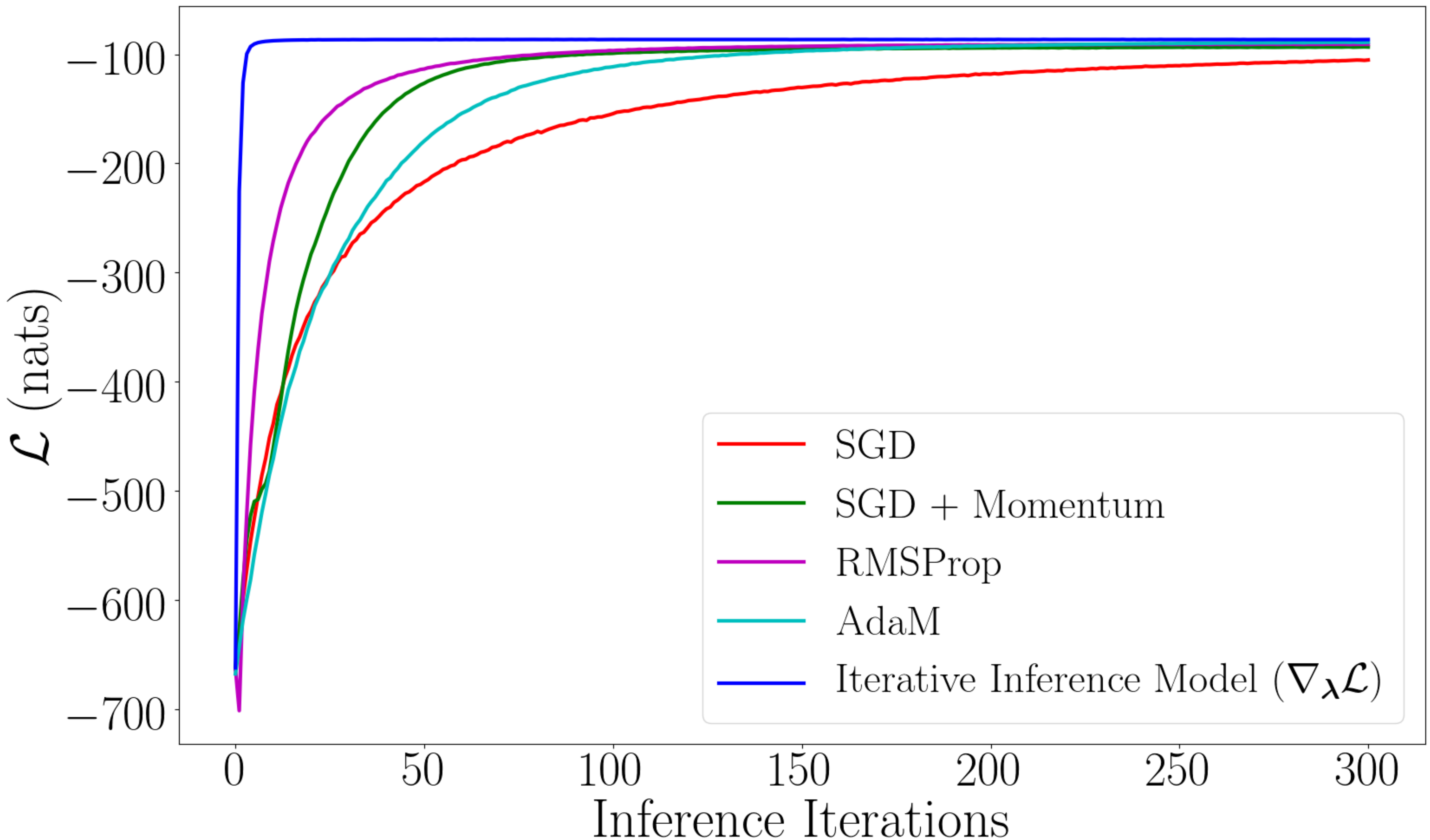
visualize data reconstructions over inference iterations



Marino *et al.*, 2018a

INFERENCE OPTIMIZATION

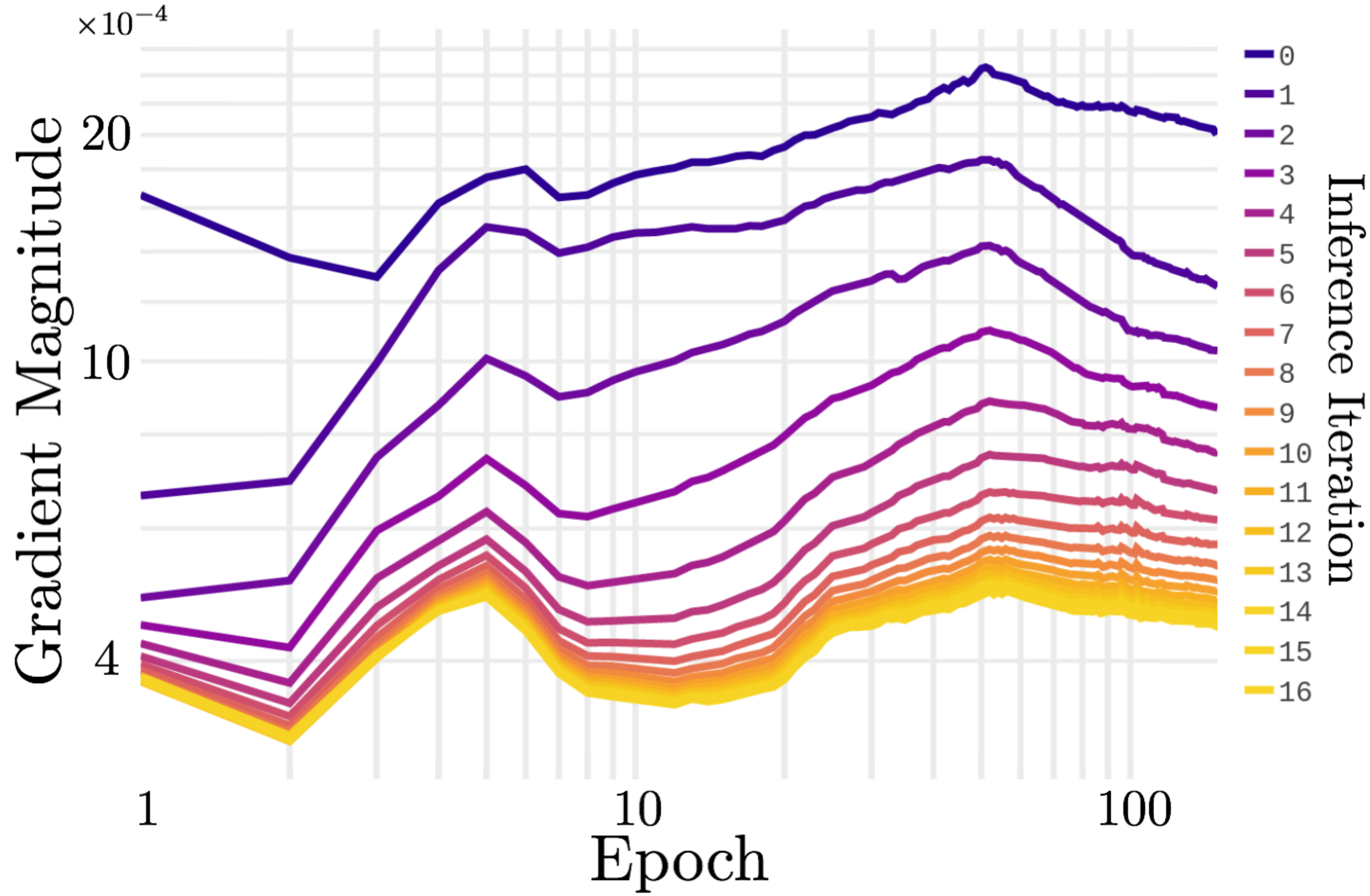
plot the ELBO over inference iterations



Marino *et al.*, 2018a

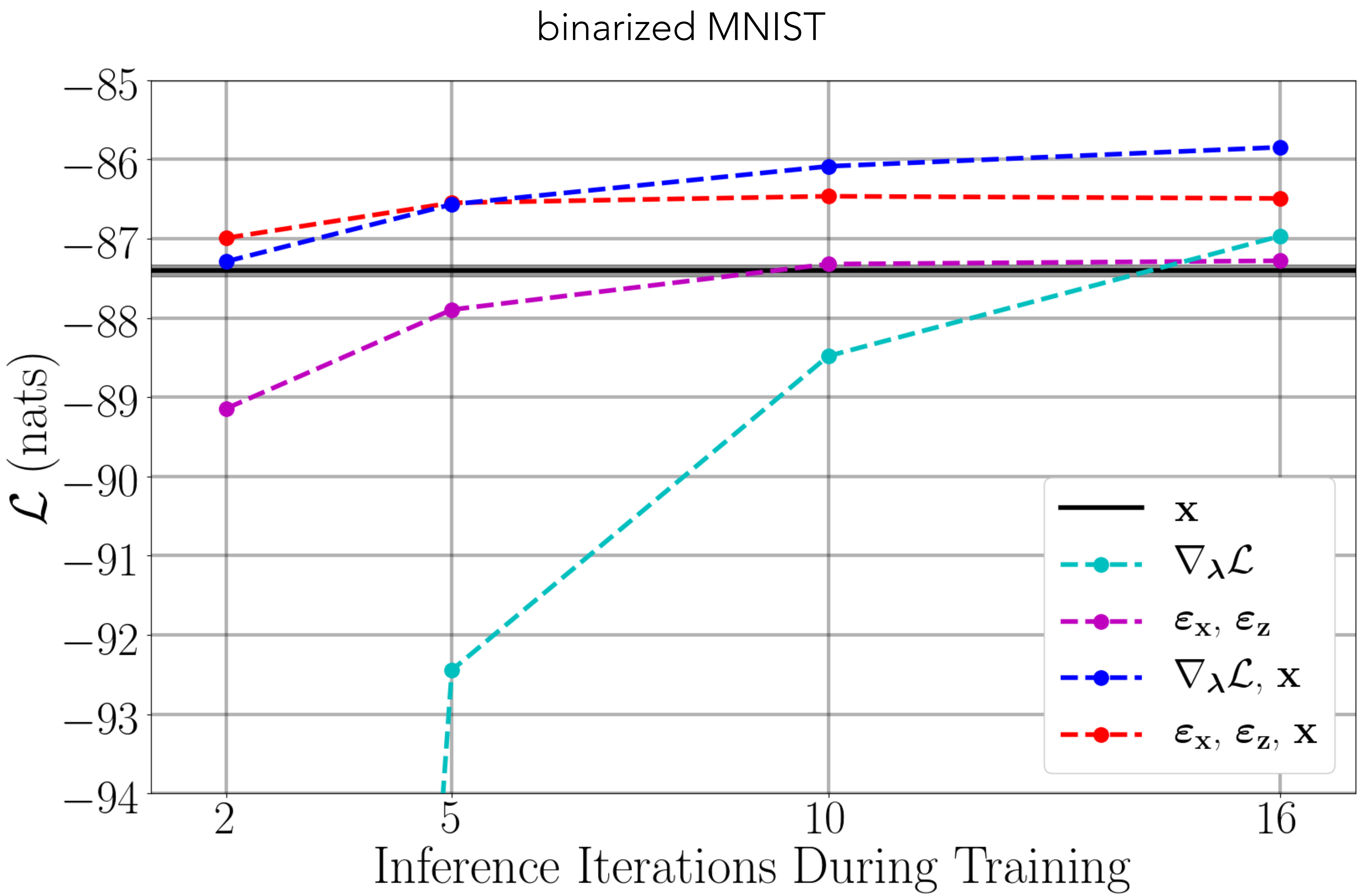
INFERENCE OPTIMIZATION

gradient magnitude decreases over inference iterations throughout training



Marino *et al.*, 2018a

QUANTITATIVE COMPARISON

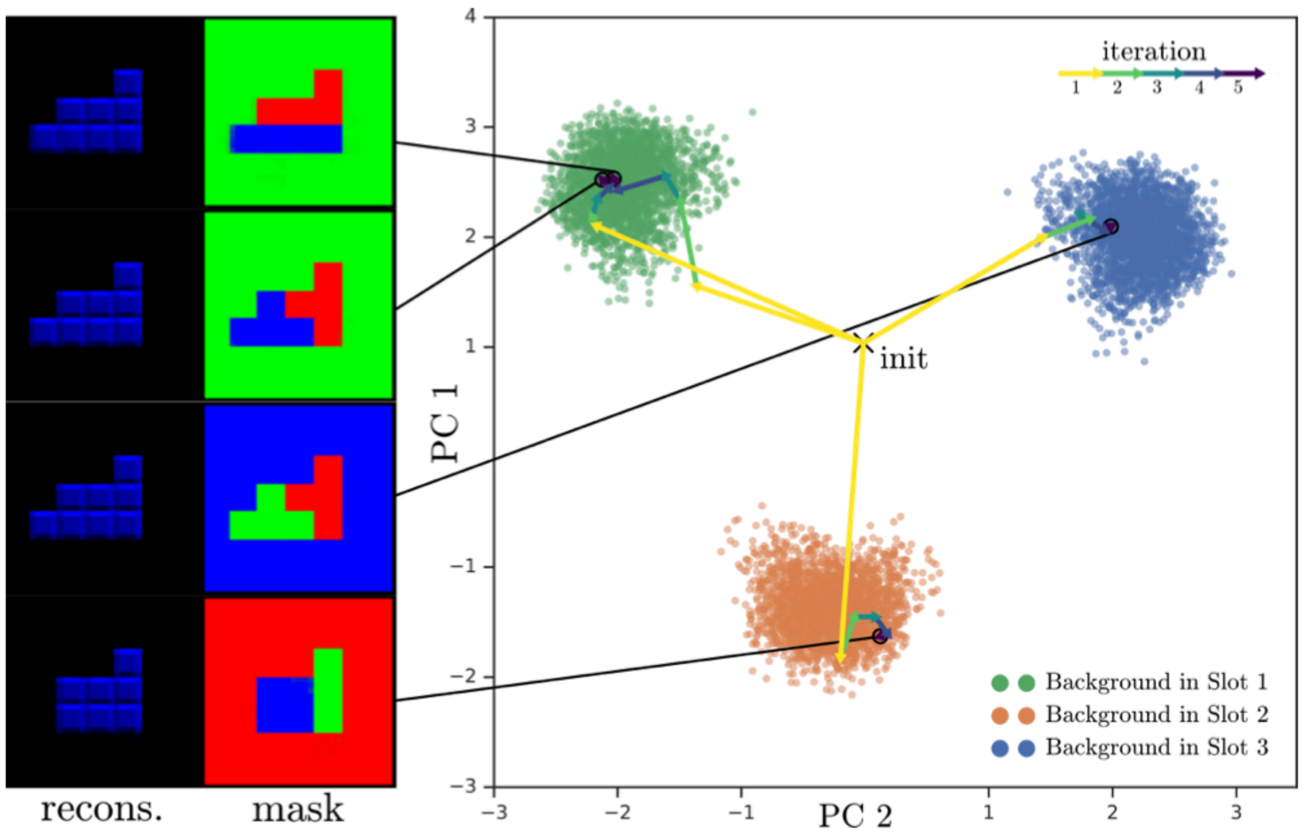


Marino *et al.*, 2018a

SYMMETRY BREAKING / MULTIPLE OPTIMA

there may be multiple equally valid estimates,

by using the gradient/errors, iterative inference can explore them

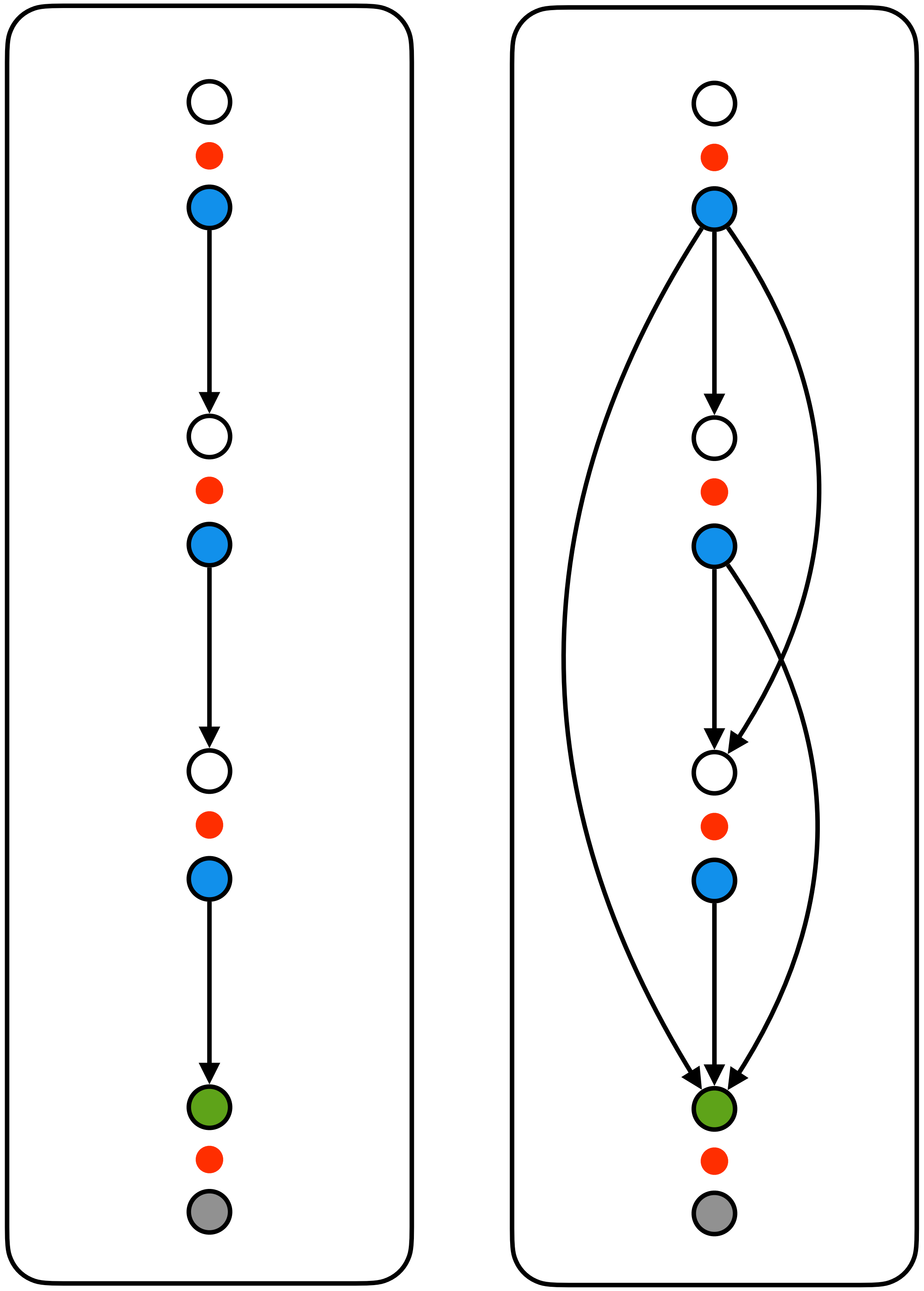


example: multiple ways to parse an image into separate objects

Greff et al., 2019

ITERATIVE AMORTIZED INFERENCE

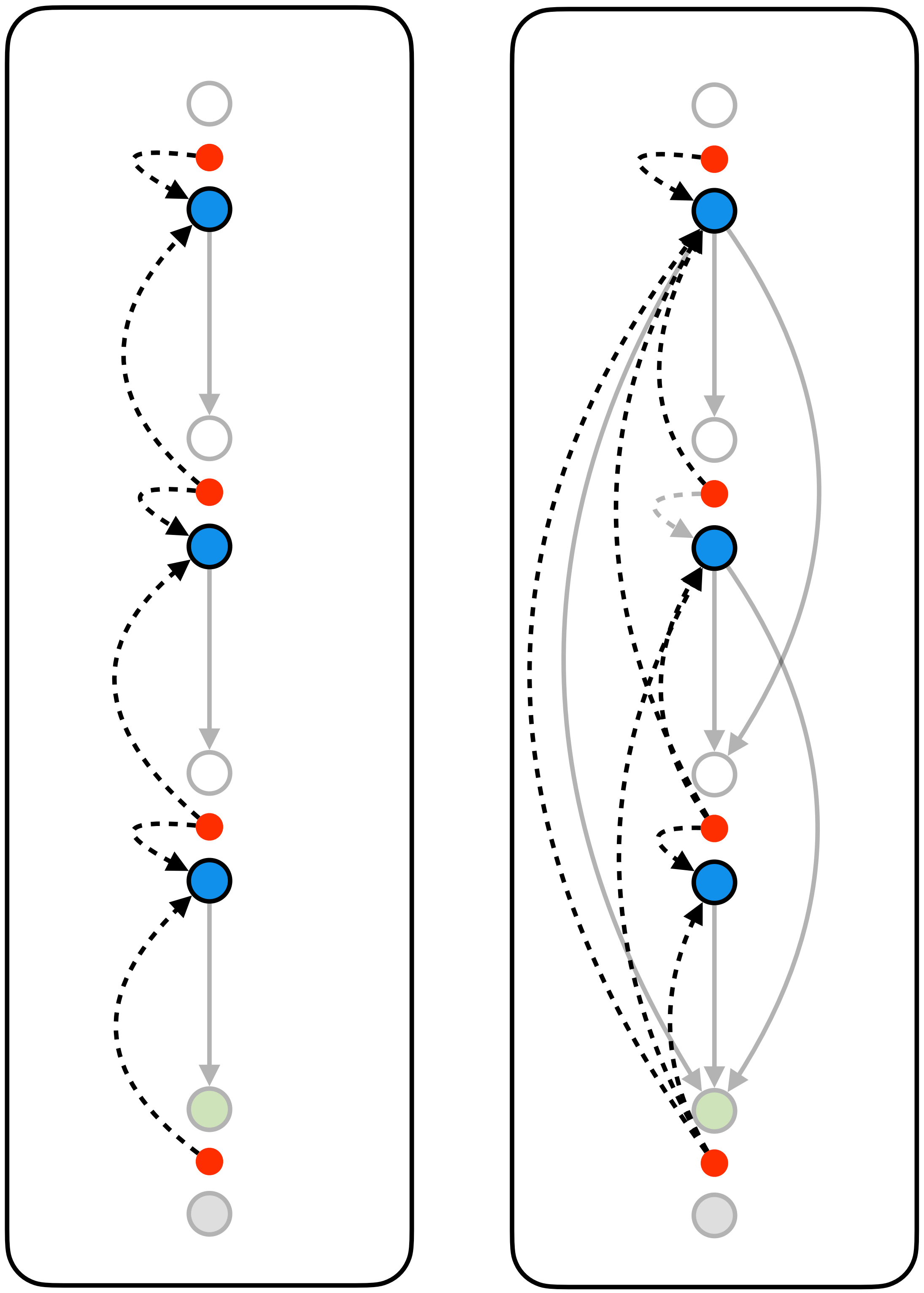
can be trivially extended
to structured models



ITERATIVE AMORTIZED INFERENCE

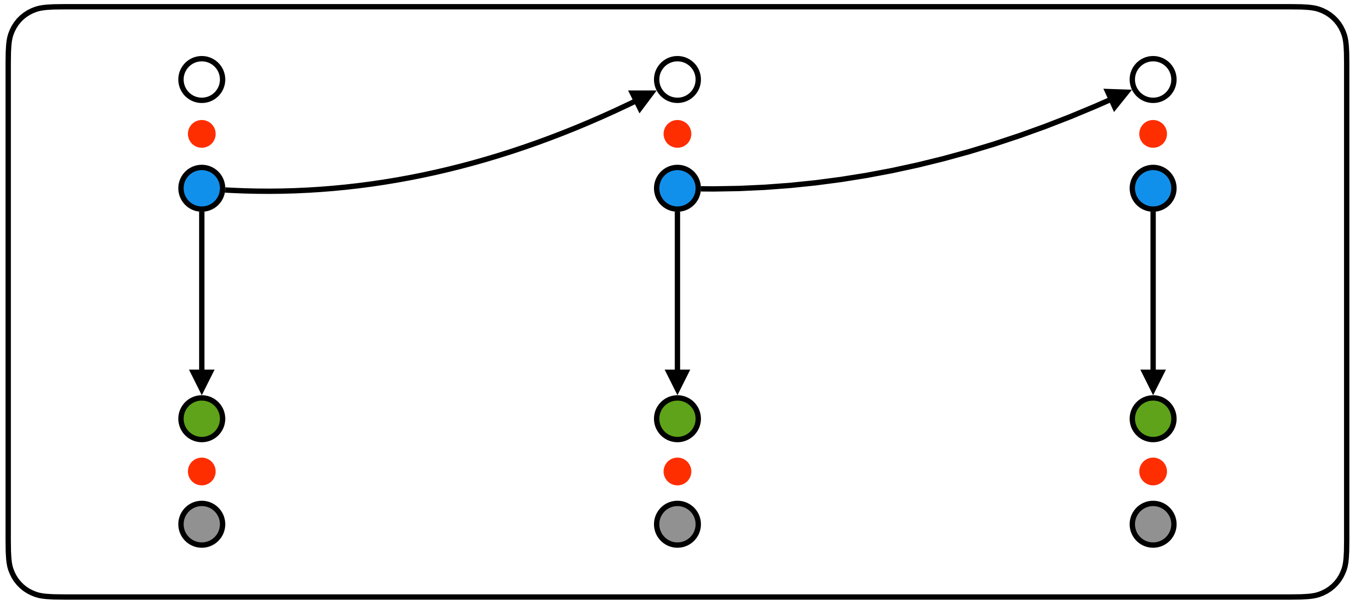
can be trivially extended
to structured models

structure defines gradients,
which define inference

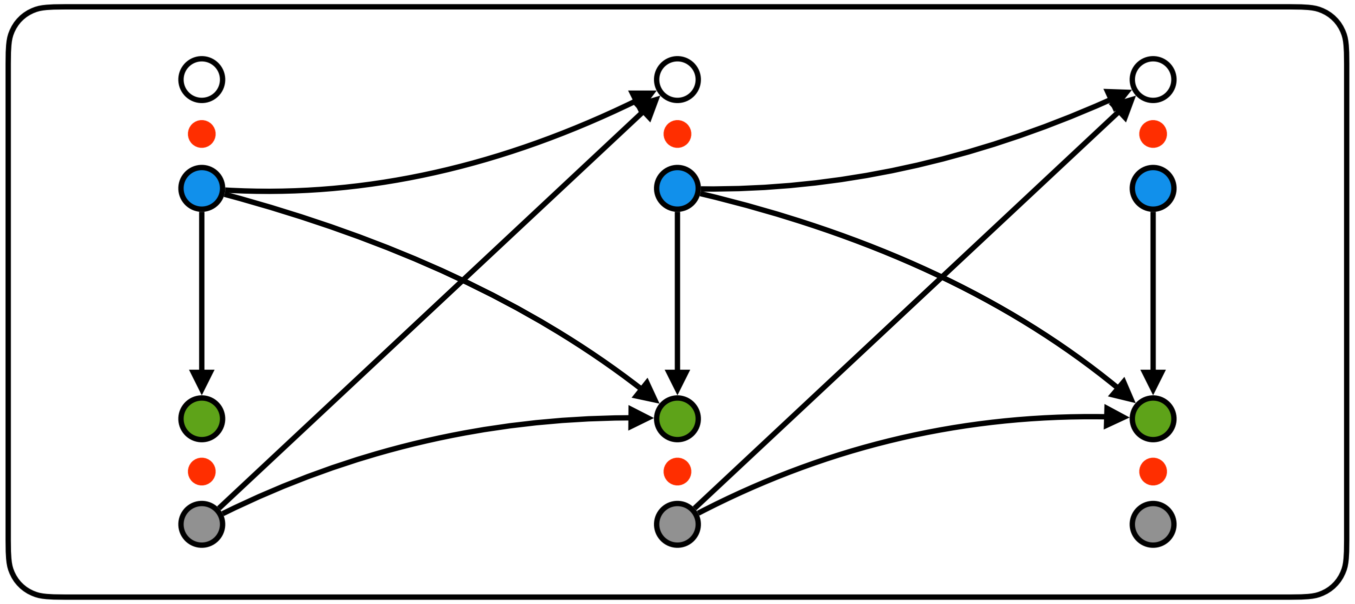


ITERATIVE AMORTIZED INFERENCE

can be trivially extended
to structured models

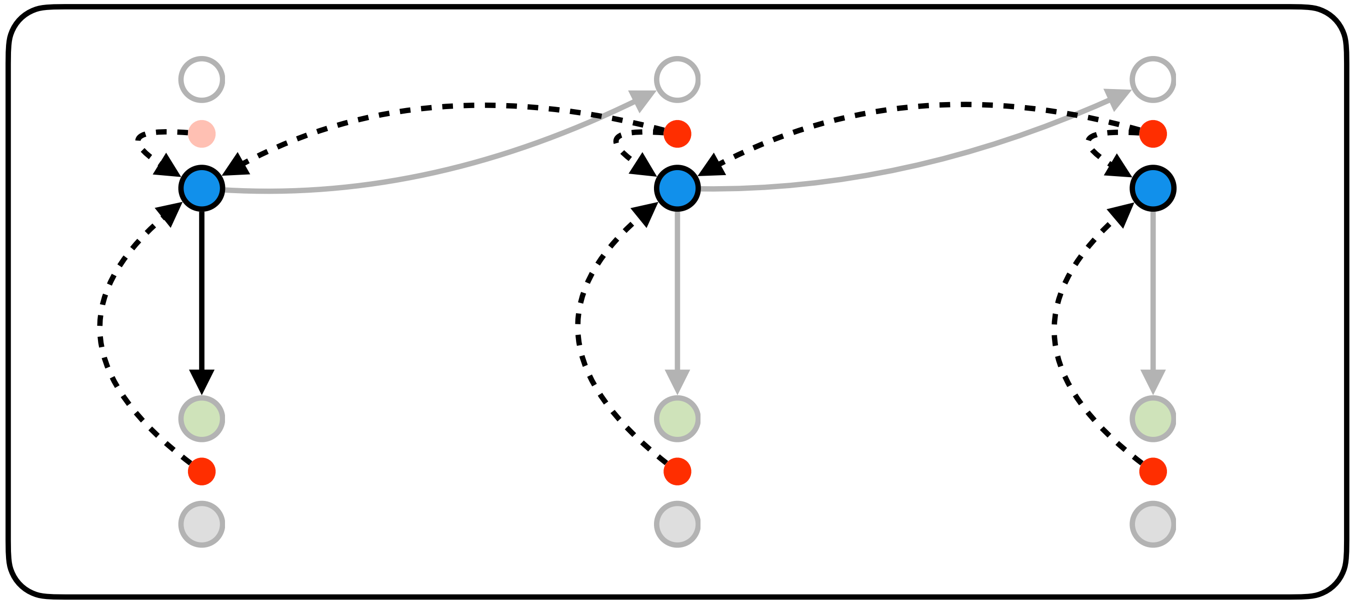


structure defines gradients,
which define inference

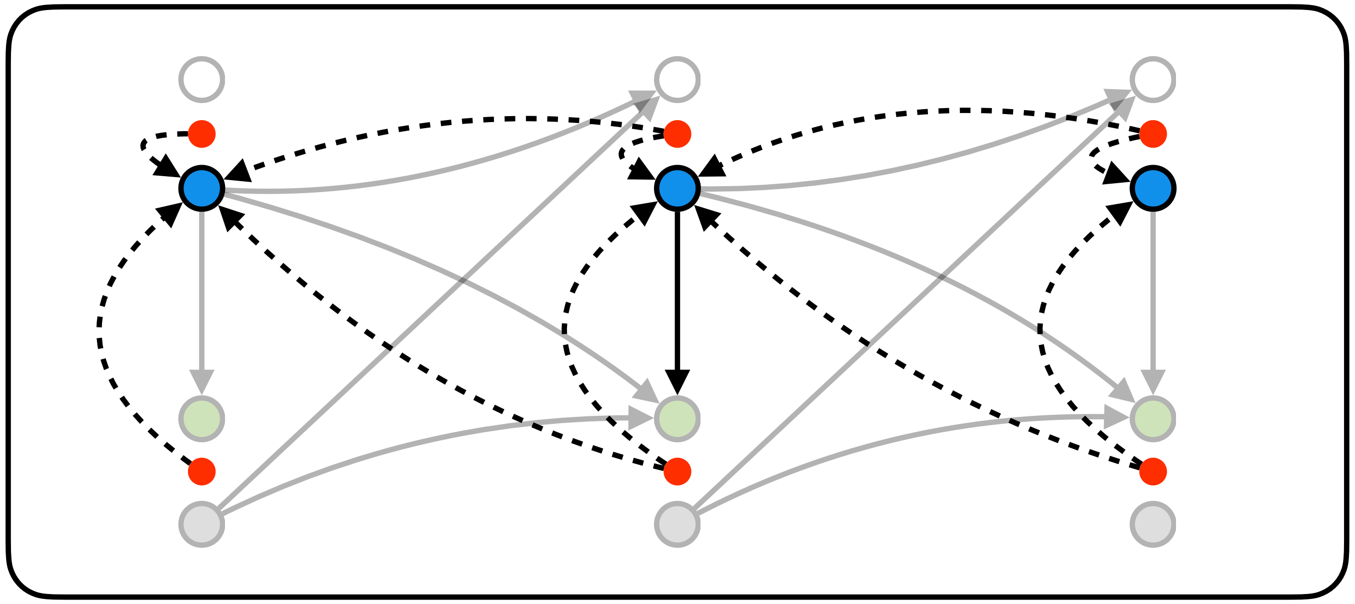


ITERATIVE AMORTIZED INFERENCE

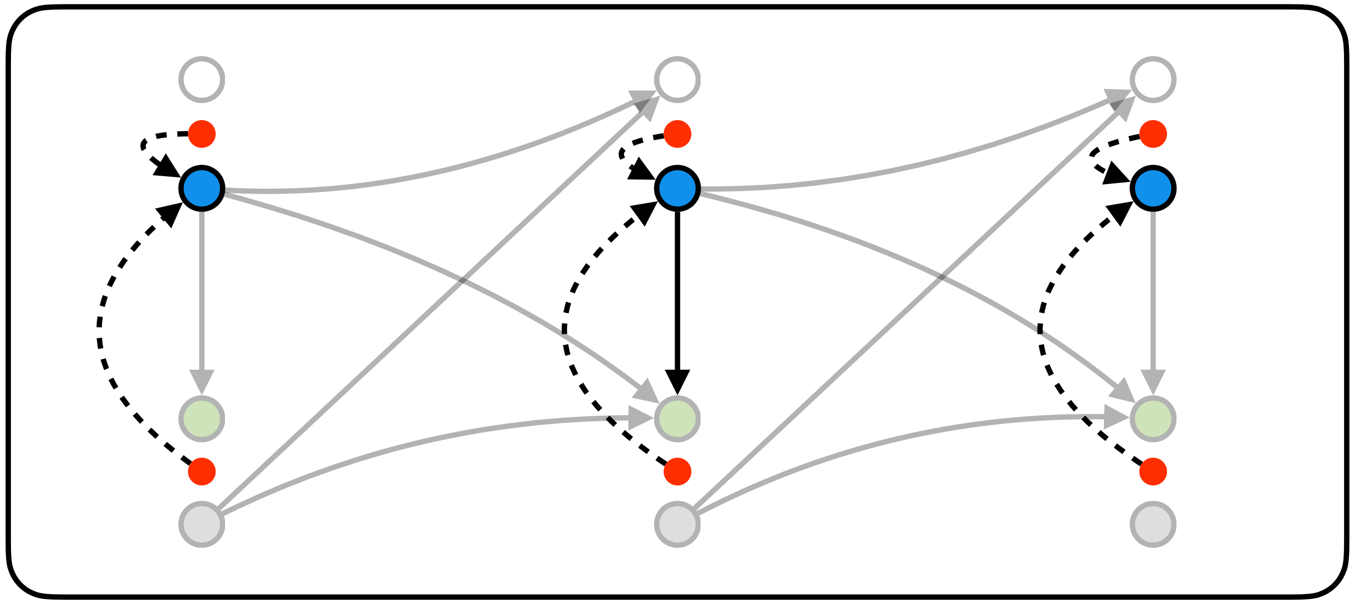
can be trivially extended
to structured models



structure defines gradients,
which define inference



FILTERING INFERENCE

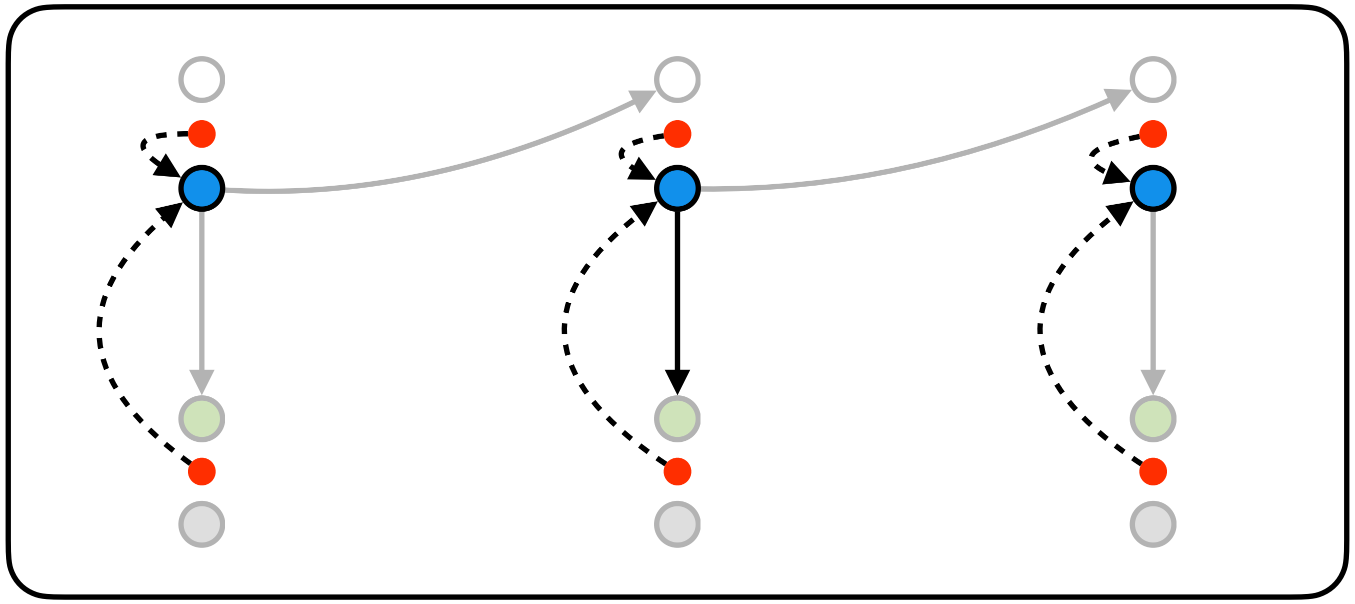


in dynamical models, *filtering* refers to performing inference using only **current** and **past** variables

filtering approximate posterior:

$$q(\mathbf{z}_{\leq T} | \mathbf{x}_{\leq T}) = \prod_{t=1}^{T} q(\mathbf{z}_t | \mathbf{x}_{\leq t}, \mathbf{z}_{<t})$$

KALMAN FILTERING

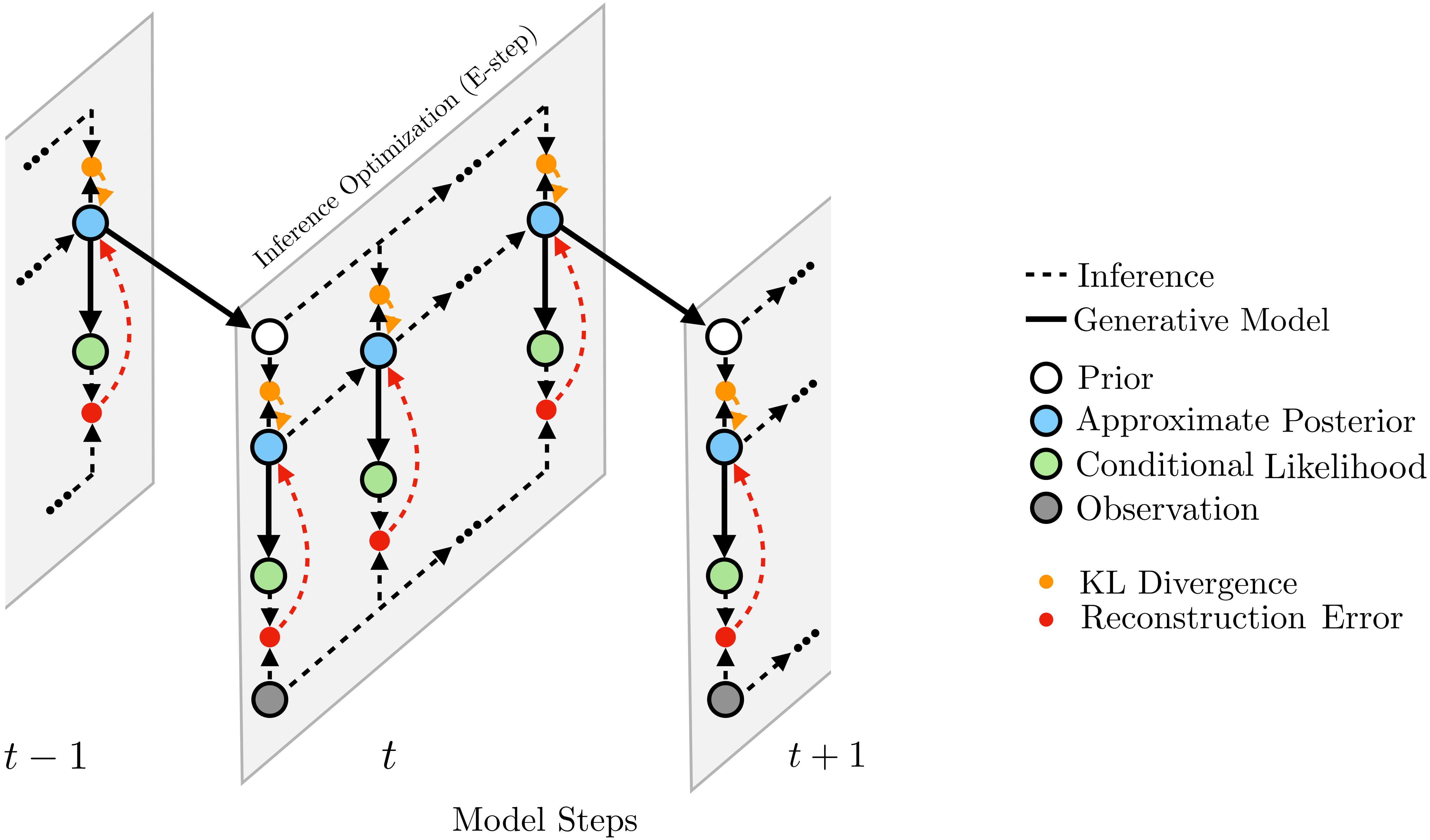


Kalman filtering performs exact inference in linear-Gaussian dynamical models

infer latent variable using prediction and residual error

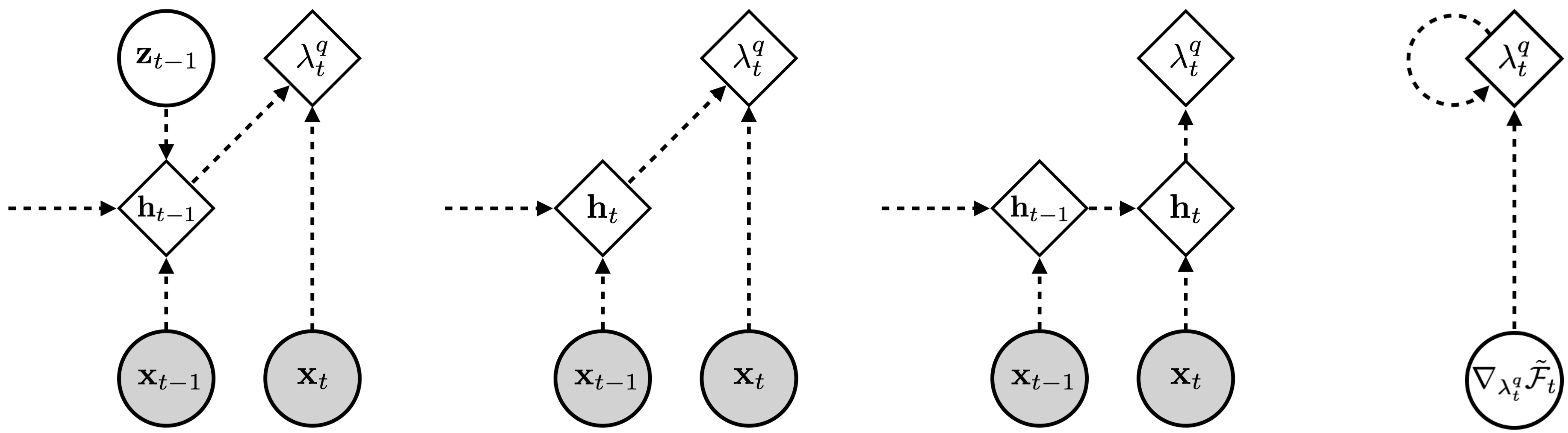
latent mean: $$\underbrace{\boldsymbol{\mu}_t}_{\text{updated estimate}} = \underbrace{\boldsymbol{\mu}_{t|t-1}}_{\text{predicted estimate}} + \underbrace{\mathbf{K}_t}_{\text{“Kalman gain”}} \underbrace{(\mathbf{x} - \hat{\mathbf{x}}_t)}_{\text{prediction error}}$$

FILTERING VARIATIONAL INFERENCE



Marino *et al.*, 2018b

FILTERING INFERENCE MODELS



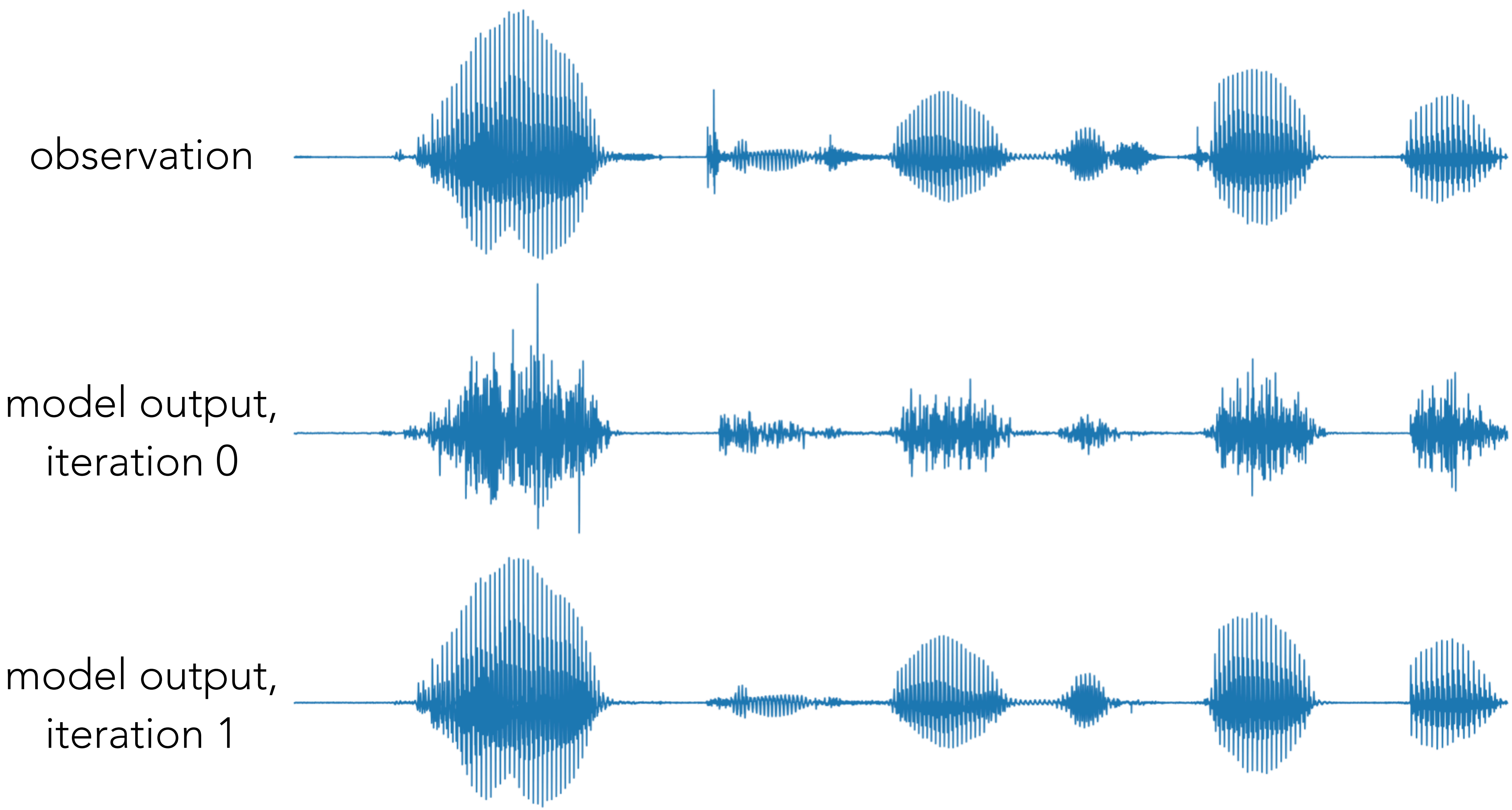
VRNN SRNN SVG AVF

custom-designed

Marino *et al.*, 2018b

VISUALIZING INFERENCE IMPROVEMENT

TIMIT audio waveforms

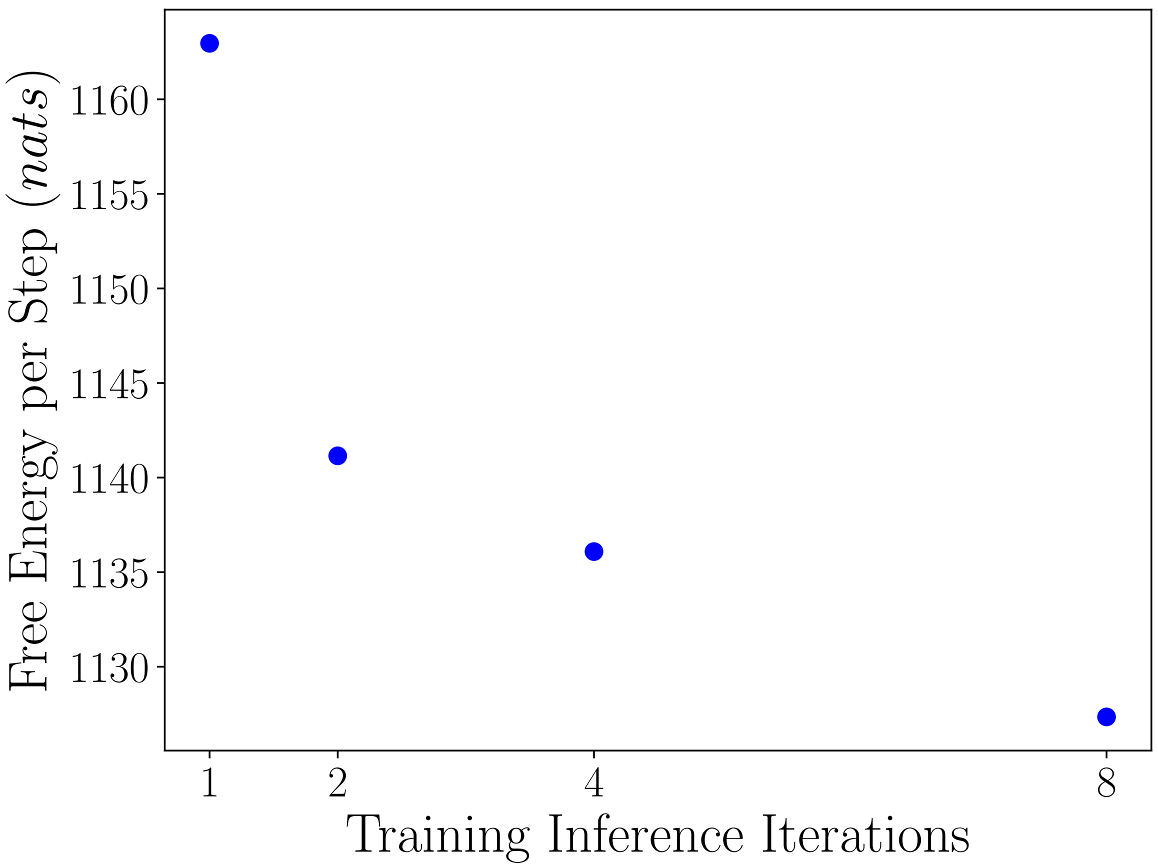


Marino *et al.*, 2018b

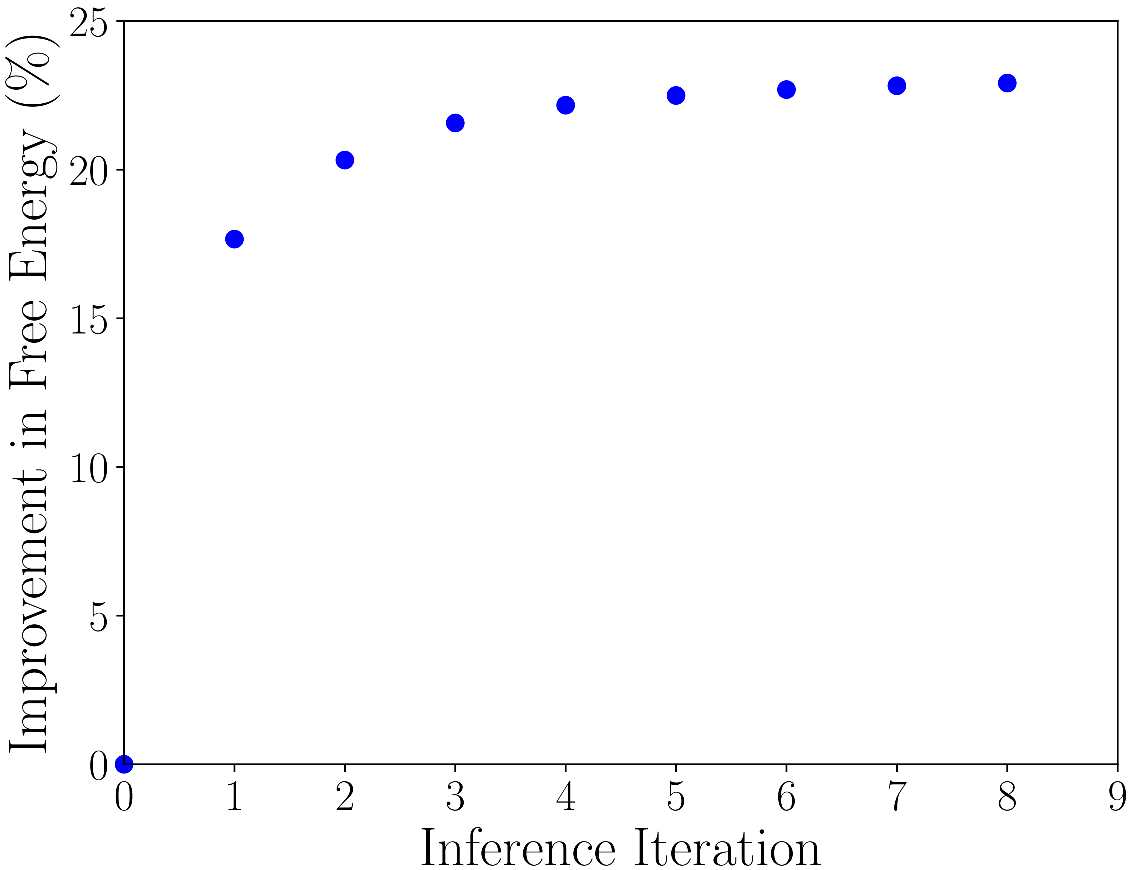
INFERENCE ITERATIONS

ON TIMIT VAL SET

training with additional inference iterations results in improved performance



each inference iteration yields decreasing relative improvement



Marino *et al.*, 2018b

QUANTITATIVE ANALYSIS

negative lower bound on test sets in *nats*

Speech	TIMIT
VRNN	
baseline	1,082
AVF (1 step)	1,105
AVF (2 step)	**1,071**
SRNN	
baseline	1,026
AVF (1 step)	**1,024**

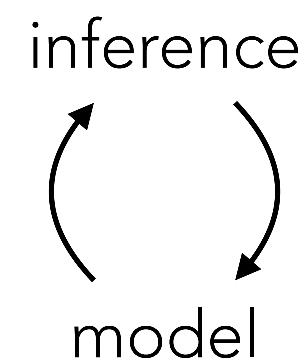
Video	KTH Actions
SVG	
baseline	15,097
AVF (1 step)	**11, 714**

Music	Piano-midi.de	MuseData	JSB Chorales	Nottingham
SRNN				
baseline [Fraccaro *et al.*]	8.20	6.28	4.74	2.94
baseline	8.19	6.27	6.92	3.19
AVF (1 step)	**8.12**	**5.99**	6.97	**3.13**
AVF (5 step)	–	–	**6.77**	–

Marino *et al.*, 2018b

CLOSING REMARKS

LAGGING INFERENCE NETWORKS

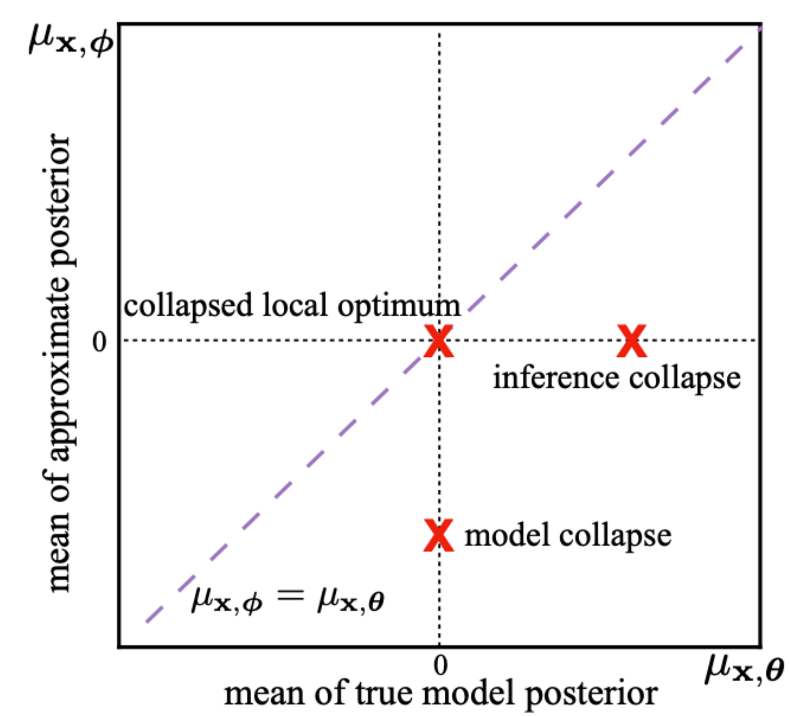


negative feedback between sub-optimal
inference and learning impairs training

$$\mathcal{L}(\mathbf{x}; q) = \mathbb{E}_q \left[\log p_\theta(\mathbf{x}|\mathbf{z}) - \log \frac{q(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z})} \right]$$

can enter local optimum when $q_\phi(\mathbf{z}|\mathbf{x}) = p_\theta(\mathbf{z})$

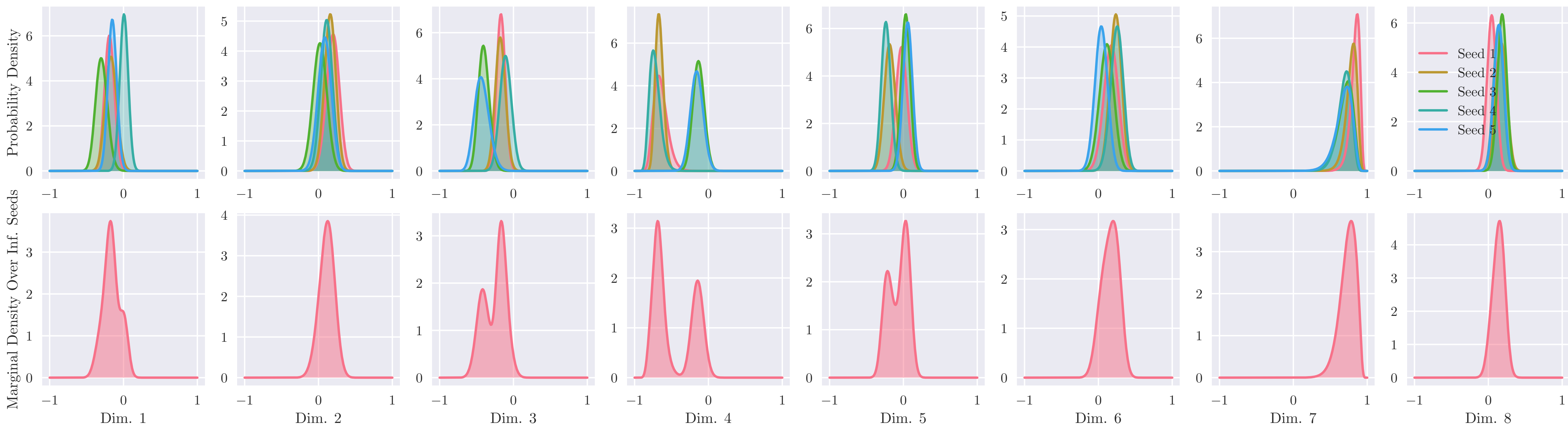
inference networks "*lag*" behind
optimal estimates, contributing to the problem



Lagging Inference Networks and Posterior Collapse in VAEs, He et al., 2019

ADDED STOCHASTICITY

by using stochastic inputs (gradients/errors), we end up with stochastic estimates



latent dimensions

yields a more expressive marginal distribution,
but can yield higher-variance estimates for individual inference seeds

degree of help/harm from this noise in unclear

Marino, unpublished

TRICKS —> THEORY?

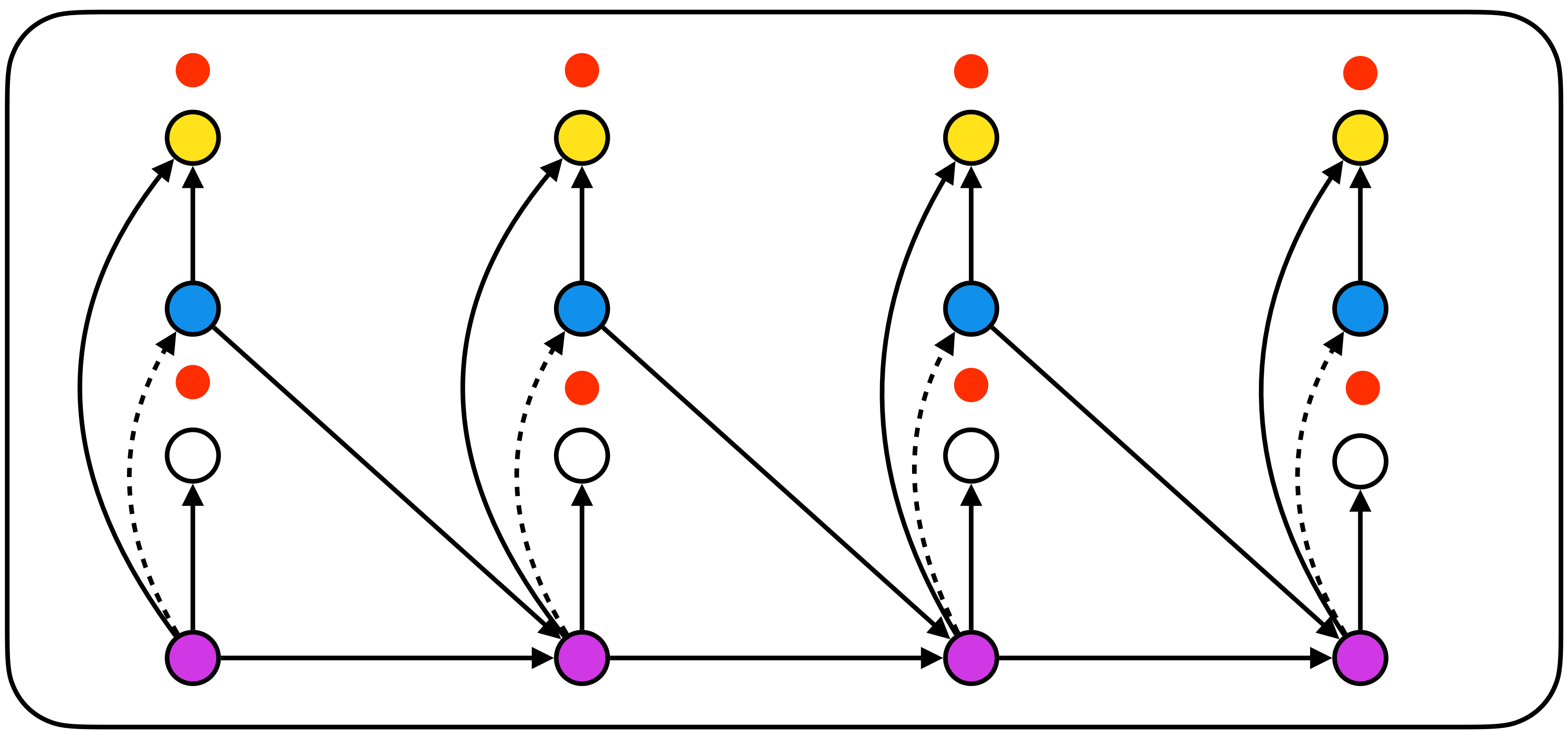
there are currently a lot of *tricks* to getting amortized/meta optimization to work well

- log-scale the inputs (Andrychowicz et al., 2016)
- layer-normalize the inputs, outputs (Marino et al., 2018ab)
- gated output (Marino et al., 2018a)
- recurrent or not? (Andrychowicz et al., 2016)
- input optimizer iteration (Lucas et al., 2018)
- optimizer truncation length (Metz et al., 2019)
- previous gradients (Metz et al., 2019)
- errors vs. gradients (Marino et al., 2018a)
- . . . (Li & Malik, 2017), (Wichrowska et al., 2017), . . .

still an empirical question whether things will work or fail, largely lacking theory

REINFORCEMENT LEARNING

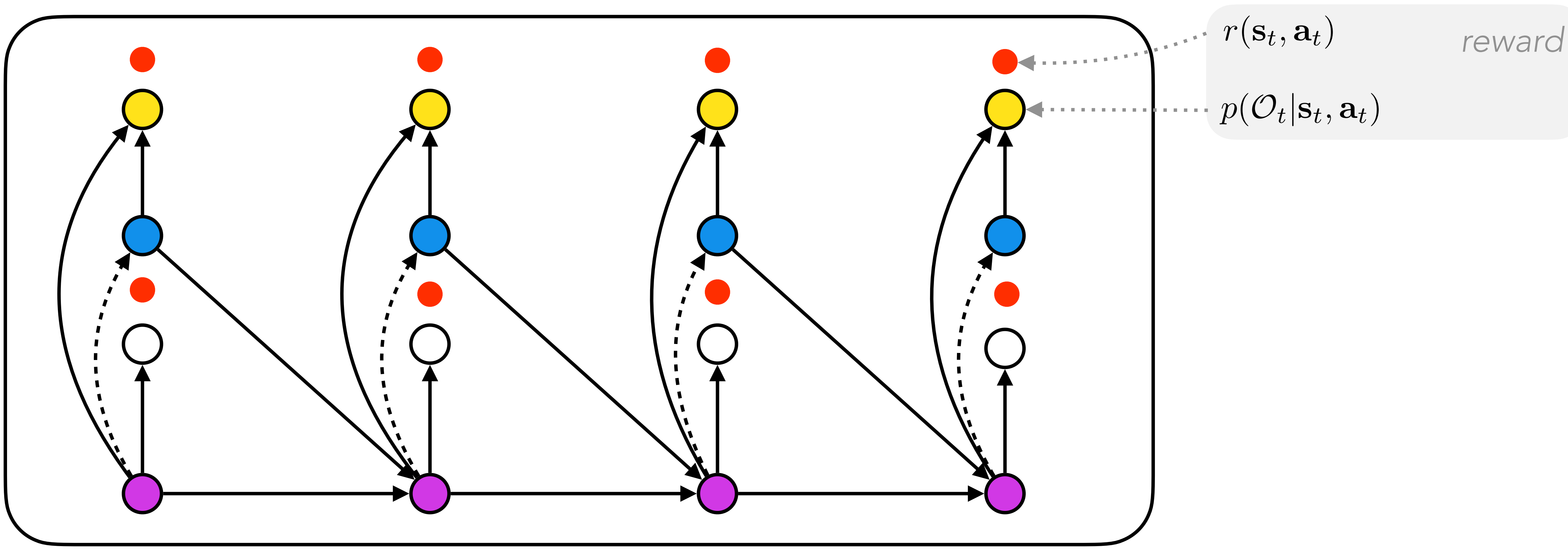
can cast reinforcement learning and control as variational inference



Levine, 2018

REINFORCEMENT LEARNING

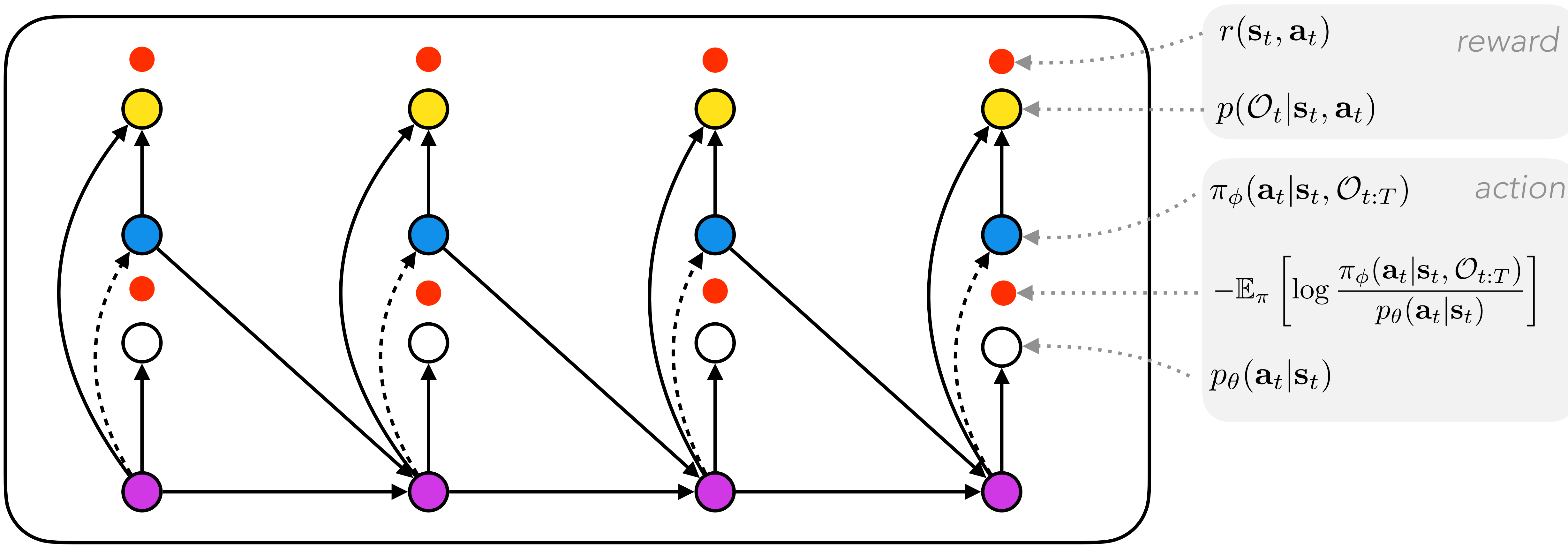
can cast reinforcement learning and control as variational inference



Levine, 2018

REINFORCEMENT LEARNING

can cast reinforcement learning and control as variational inference

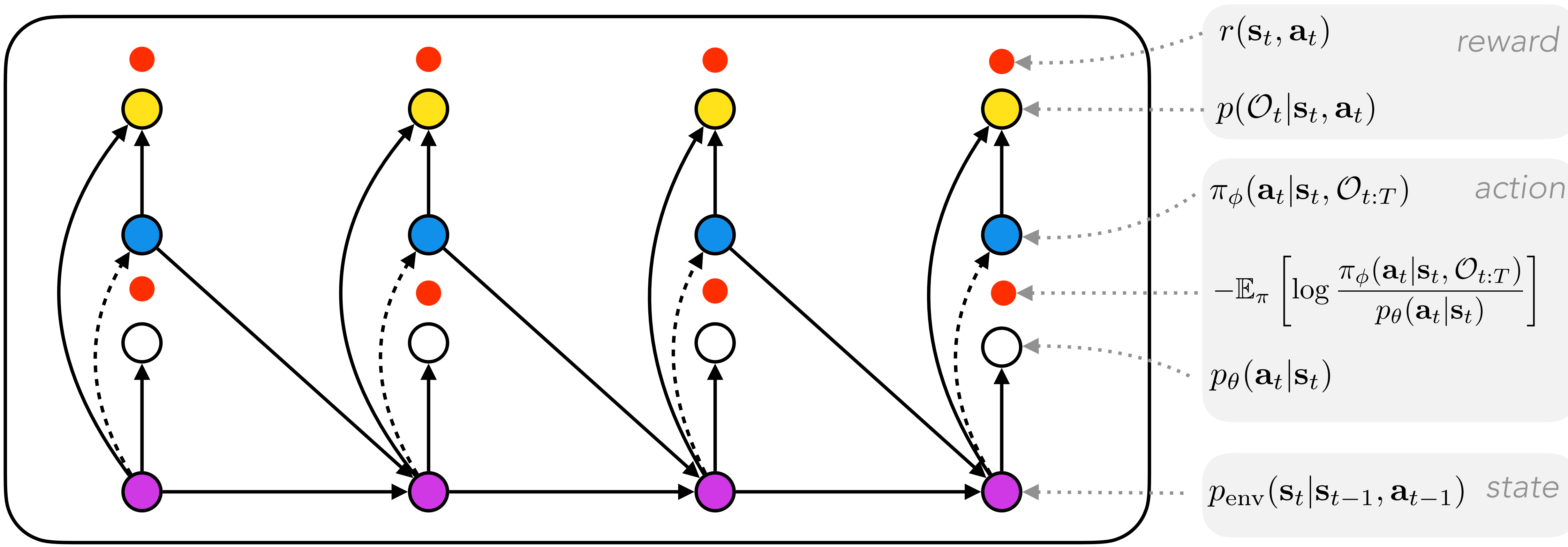


policy networks are a form of *amortized optimization!*

Levine, 2018

REINFORCEMENT LEARNING

can cast reinforcement learning and control as variational inference



policy networks are a form of *amortized optimization!*

Levine, 2018