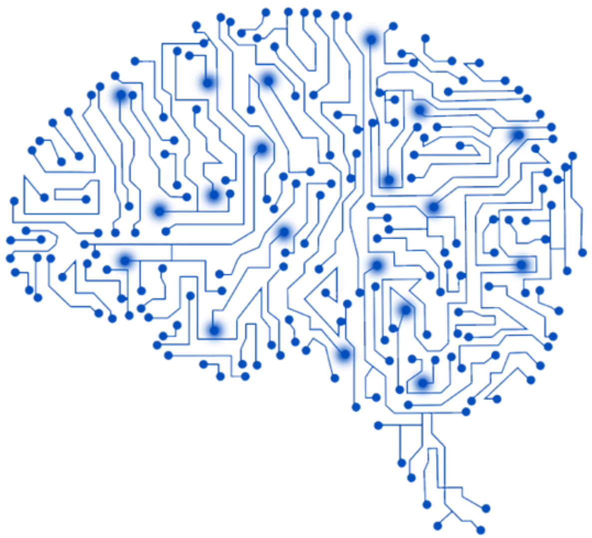
Improving Variational Inference for Complex Probabilistic Modeling



Anqi Wu

School of Computational Science and Engineering
Georgia Tech

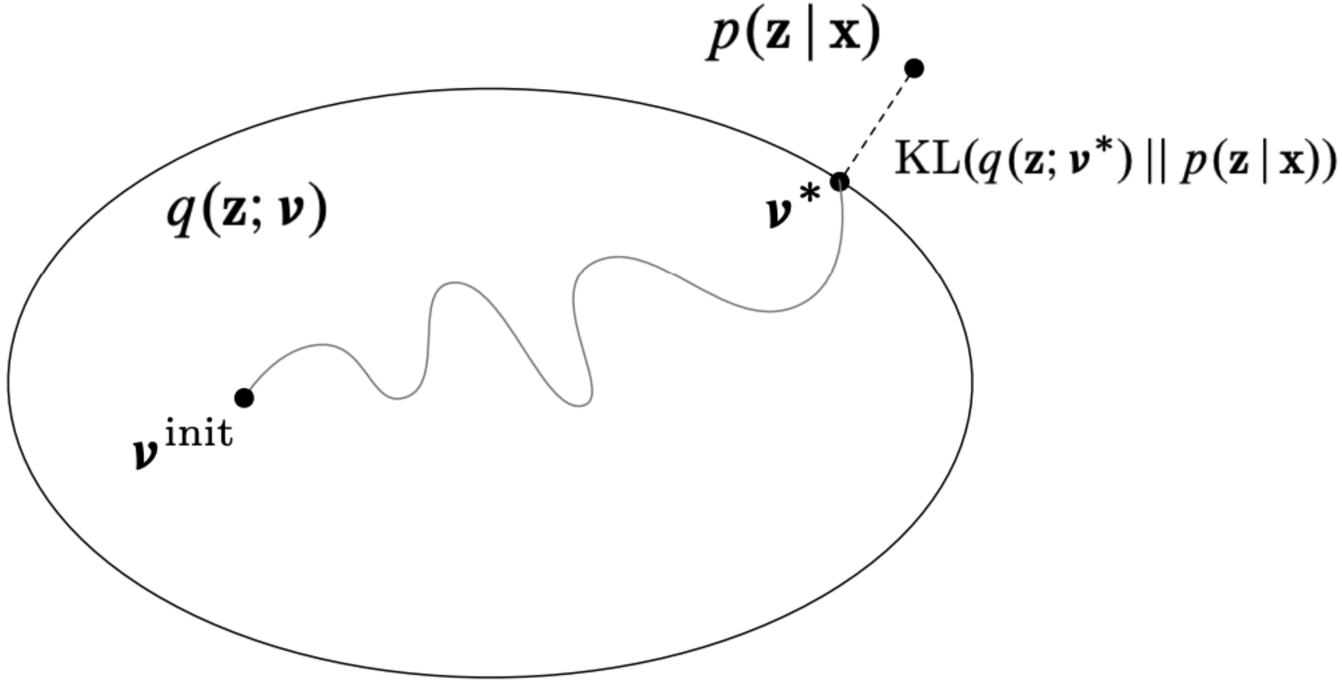
Probabilistic Machine Learning

- A probabilistic model is a joint distribution of hidden variables z and observed variables **x, p(z, x)**.

- Inference about the unknowns is through the **posterior**, the conditional distribution of the hidden variables given the observations **$p(z|x) = p(z, x)/p(x)$**.

- For most interesting models, the denominator is not tractable. We appeal to **approximate posterior inference**.

Variational Inference



- Variational inference turns **inference into optimization**.
- Posit a **variational family** of distributions over the latent variables, **q(z; v)**
- Fit the **variational parameters v** to be close (in KL) to the exact posterior. (There are alternative divergences, which connect to algorithms like EP, BP, and others.)

Variational Inference

- Assume q(z; v) is an approximate posterior distribution (mean-field Gaussian)

$$v^* = \text{argmin}_v \;\; D_{KL}[q(z; v) || p(z | x)]$$

Where

$$D_{KL}[q(z; v) || p(z | x)] = \mathbb{E}_q[\log \frac{q(z; v)}{p(z | x)}] = -\underbrace{\left(\mathbb{E}_{z \sim q}[\log p(x | z)] - D_{KL}(q(z; v) || p(z))\right)}_{\text{ELBO (evidence lower bound)}} + \underbrace{\log p(x)}_{\text{constant}}$$

- Thus, minimizing the KL is equivalent to maximizing the ELBO.

$$v^* = \text{argmax}_v \;\; \mathbb{E}_{z \sim q}[\log p(x | z)] - D_{KL}(q(z; v) || p(z))$$

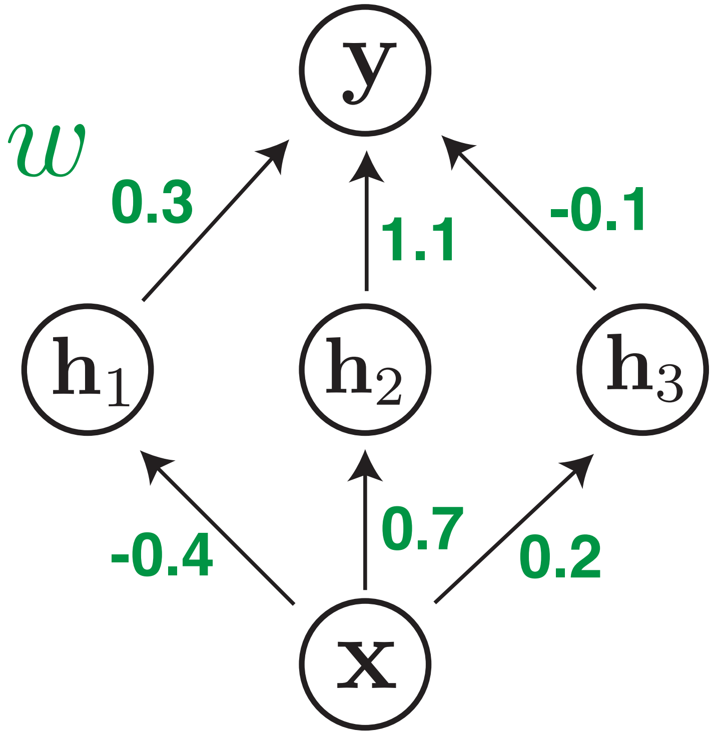
Still intractable!

Outline

- **Determinstic variational inference** for Bayesian neural networks
  - Eliminate gradient variance in evaluating the expectation term
  - Empirical Bayes to avoid the prior tuning (*general approach*)

- **Variational importance sampling** for partially observed multivariate Hawkes process
  - VIS provides a tighter bound than ELBO (*general approach*)
  - Novel forward-backward approximate distribution

Standard Neural Network



☺ Flexible class of models

Bayesian Neural Network

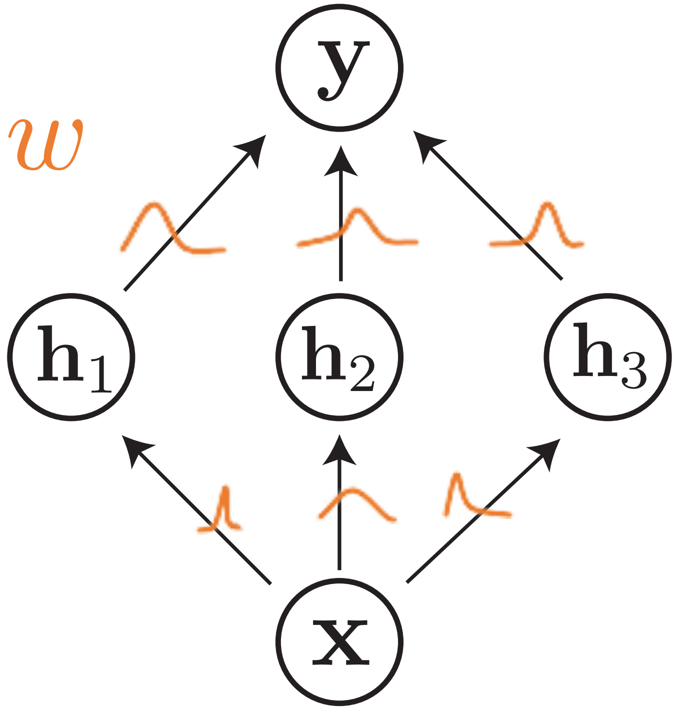


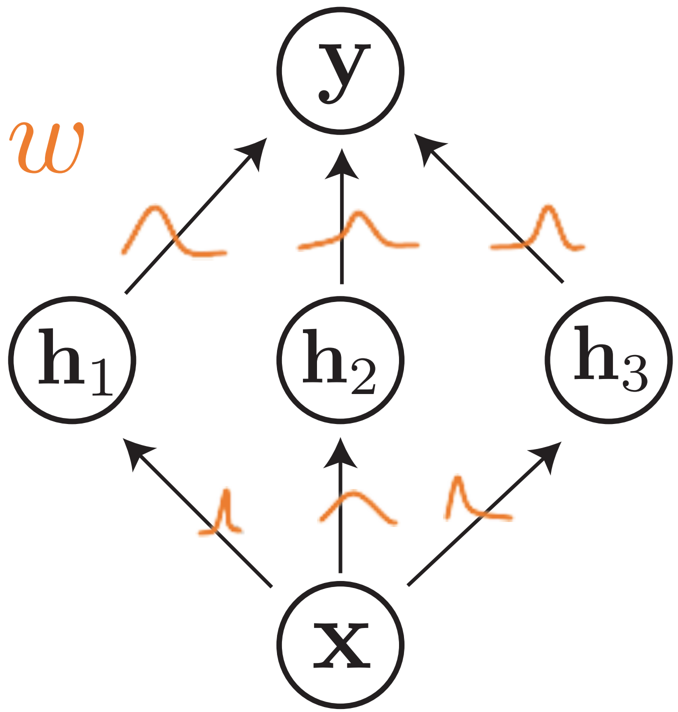
Image credit: Blundell et al., 2015

☺ Flexible class of models

☺ Principled handling of uncertainty

☺ Principled handling of regularization

Bayesian Neural Network



Goal The posterior distribution of w is $p(w|x, y)$.

Solution Variational Inference

variational approximate posterior $q_\theta(w) \sim p(w|x, y)$

ELBO (evidence lower bound)

$$\max_\theta \mathbb{E}_{q_\theta(w)} [\log p(y|x, w)] - D_{KL} [q_\theta(w) || p(w)]$$

Fit the data

challenge I

Don't stray far from the prior

challenge II

Challenge I: Gradient Variance

$$\mathbb{E}_{q_\theta(w)}\left[\log p(y|x,w)\right] - D_{KL}\left[q_\theta(w)||p(w)\right]$$

↑

Fit the data

↑

Monte Carlo sampling

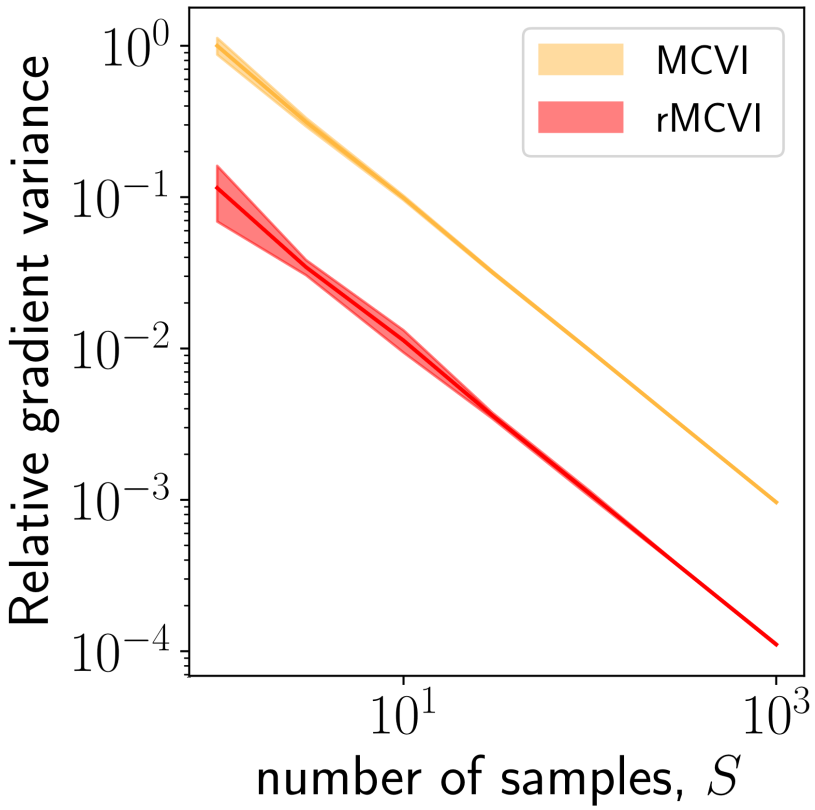
↑

gradient variance

Challenge I: Gradient Variance

MCVI: Monte Carlo Variational Inference

DVI: Deterministic Variational Inference



reduce gradient variance

local reparameterization trick
Kingma et al., 2015

variational dropout
Kingma et al., 2015, Molchanov et al., 2017

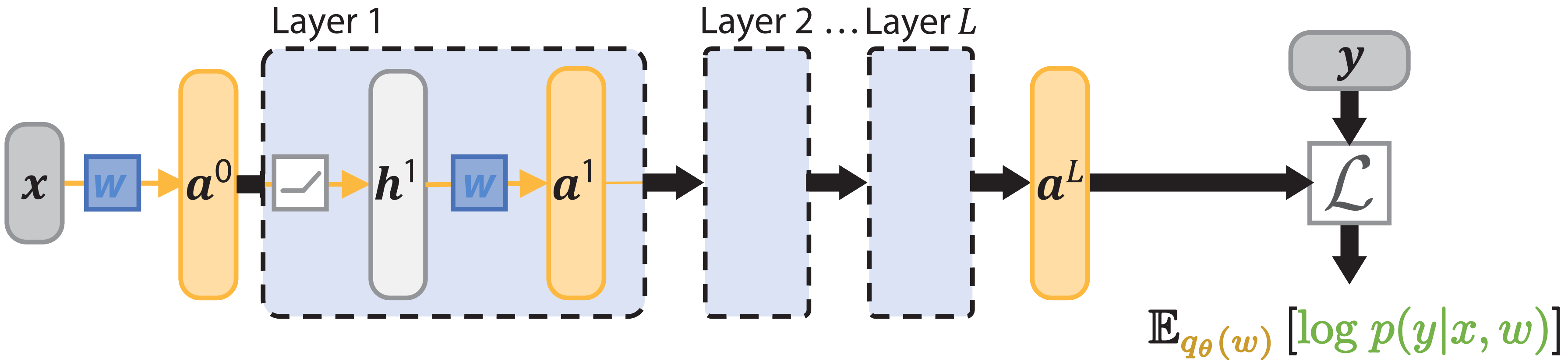
reparameterization gradient estimators
Miller et al., 2017

control variates
Zhu et al., 2018

Ours: deterministic approximation instead of MC, thus no gradient variance

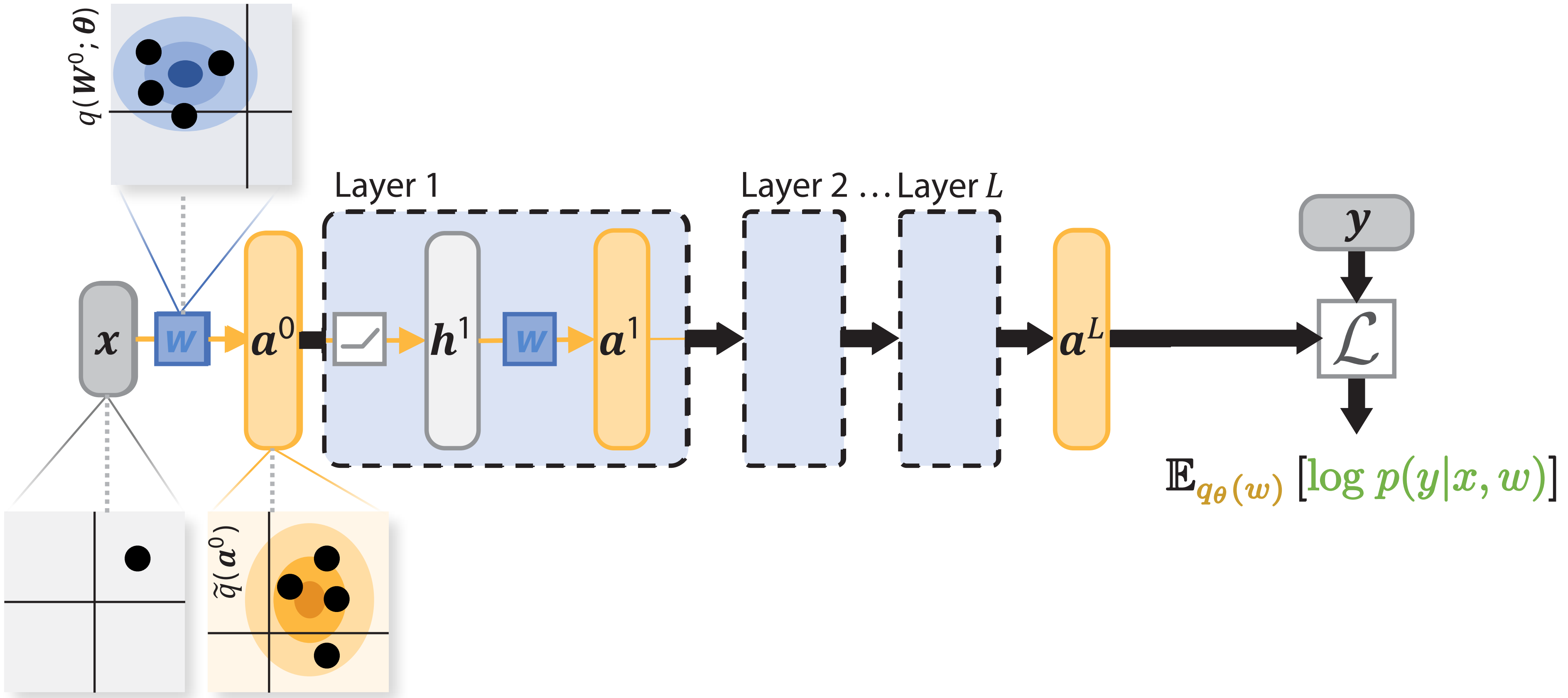
Monte Carlo Approximation for ELBO

$$\mathbb{E}_{q_\theta(w)}\left[\log p(y|x,w)\right] \approx \frac{1}{S}\sum_{s=1}^{S} \log p(y|w^{(s)},x), \quad w^{(s)} \sim q_\theta(w)$$



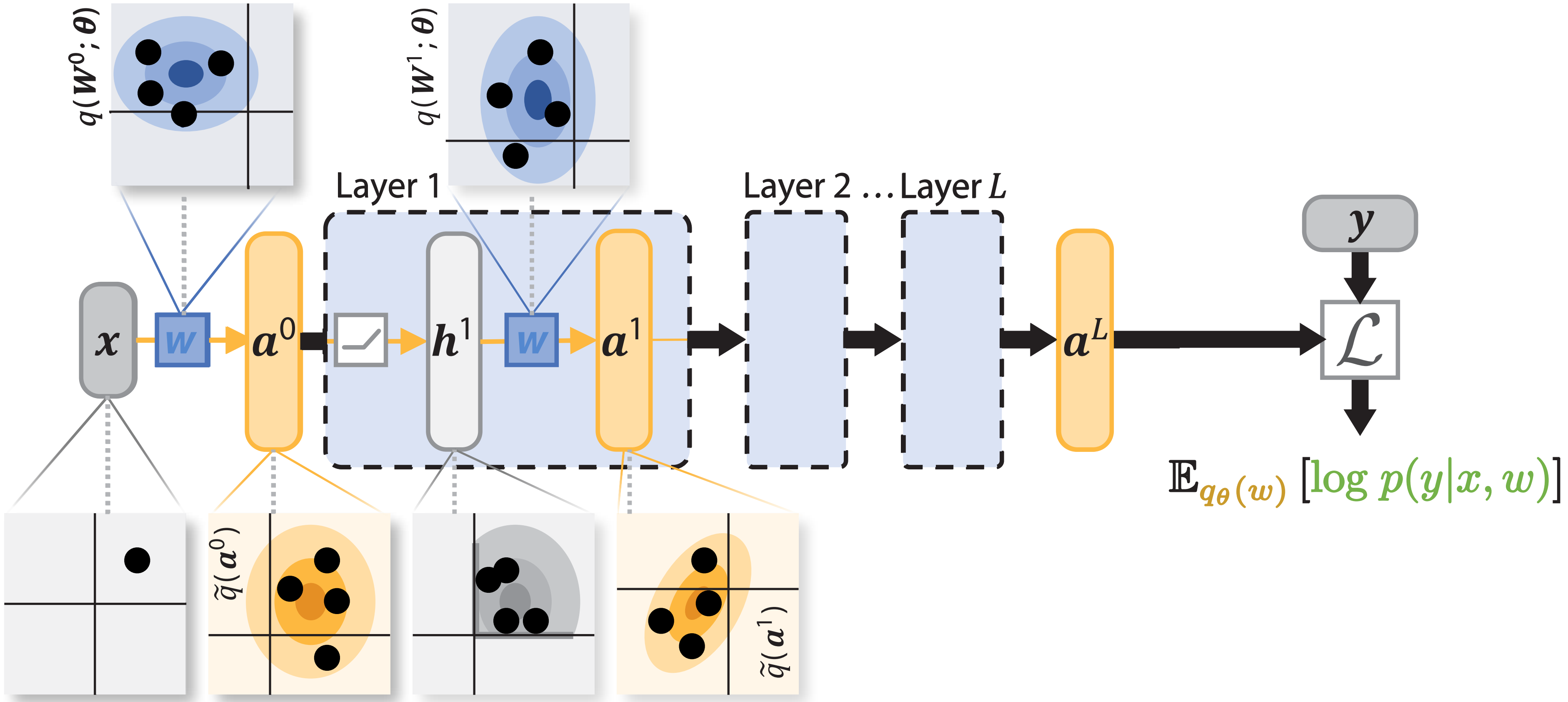
Monte Carlo Approximation for ELBO

$$\mathbb{E}_{q_\theta(w)}\left[\log p(y|x,w)\right] \approx \frac{1}{S}\sum_{s=1}^{S} \log p(y|w^{(s)},x), \quad w^{(s)} \sim q_\theta(w)$$



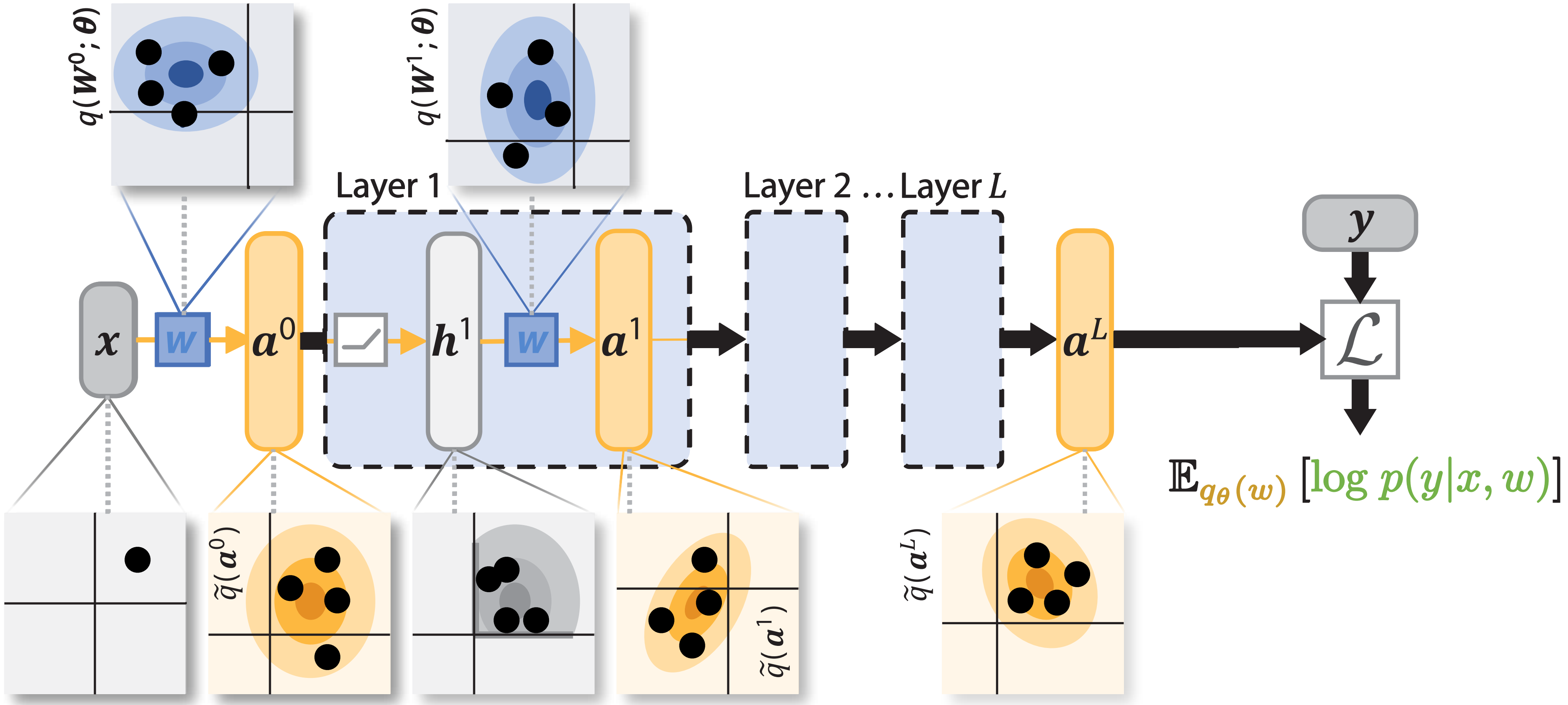
Monte Carlo Approximation for ELBO

$$\mathbb{E}_{q_\theta(w)}\left[\log p(y|x,w)\right] \approx \frac{1}{S}\sum_{s=1}^{S} \log p(y|w^{(s)},x), \quad w^{(s)} \sim q_\theta(w)$$

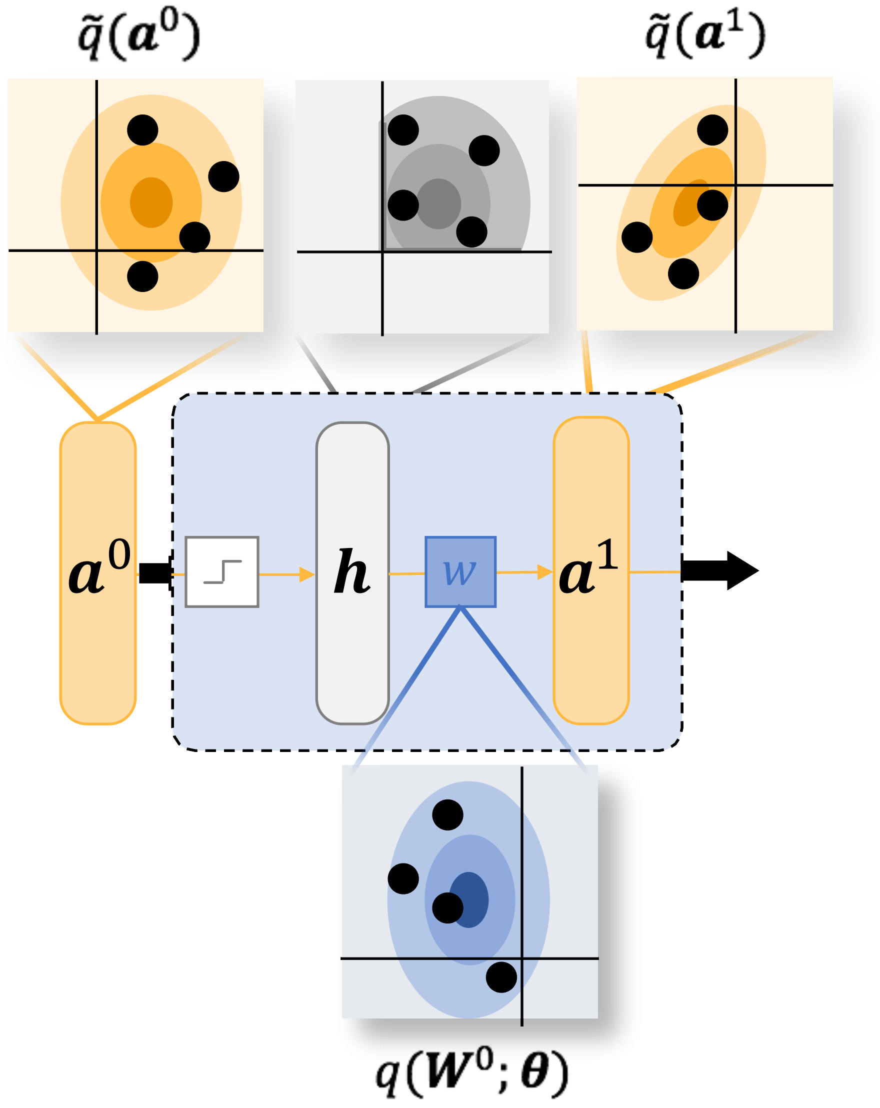


Monte Carlo Approximation for ELBO

$$\mathbb{E}_{q_\theta(w)}\left[\log p(y|x,w)\right] \approx \frac{1}{S}\sum_{s=1}^{S} \log p(y|w^{(s)},x), \quad w^{(s)} \sim q_\theta(w)$$



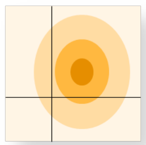
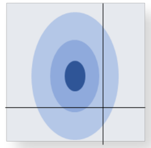
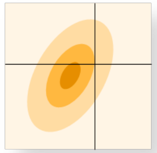
Challenge I: Deterministic Propagation of Uncertainties



Instead of propagating uncertainties via **samples**.

We can **deterministically** propagate **distributions**.

$\tilde{q}(\boldsymbol{a}^0)$ $q(\boldsymbol{W}^0;\boldsymbol{\theta})$ $\tilde{q}(\boldsymbol{a}^1)$

bnn.activation_layer( , ) = 

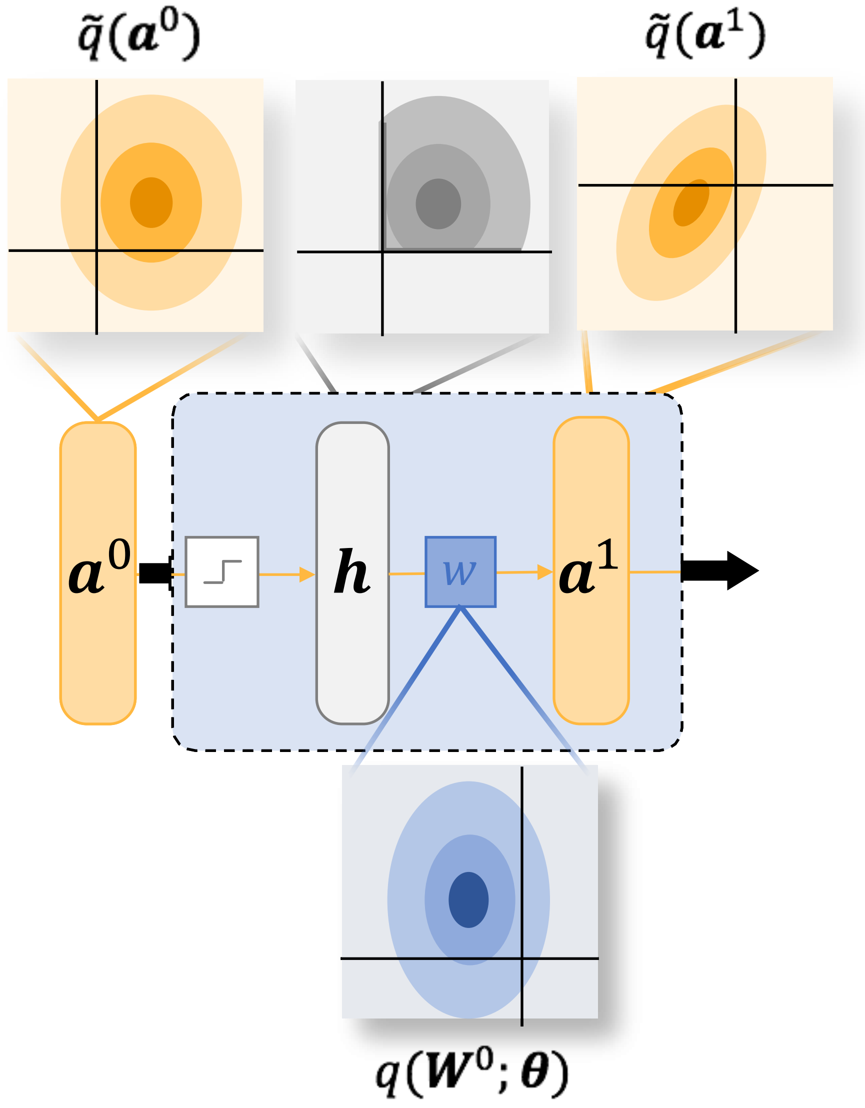
Gaussian

$$a_i^1 = \sum_{i=1}^{d} w_{ij} h_j \sim \mathcal{N}(\boldsymbol{\mu}^1, \boldsymbol{\Sigma}^1)$$

Central Limit Theorem:

1. $\boldsymbol{W}$ and $\boldsymbol{h}$ are i.i.d. samples
2. Large number of hidden nodes in $\boldsymbol{h}$

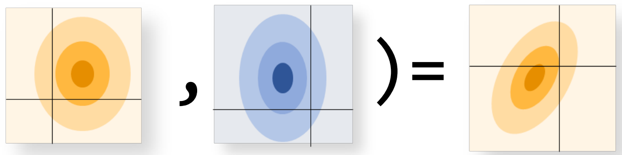
Challenge I: Deterministic Propagation of Uncertainties



Instead of propagating uncertainties via **samples**.

We can **deterministically** propagate **distributions**.

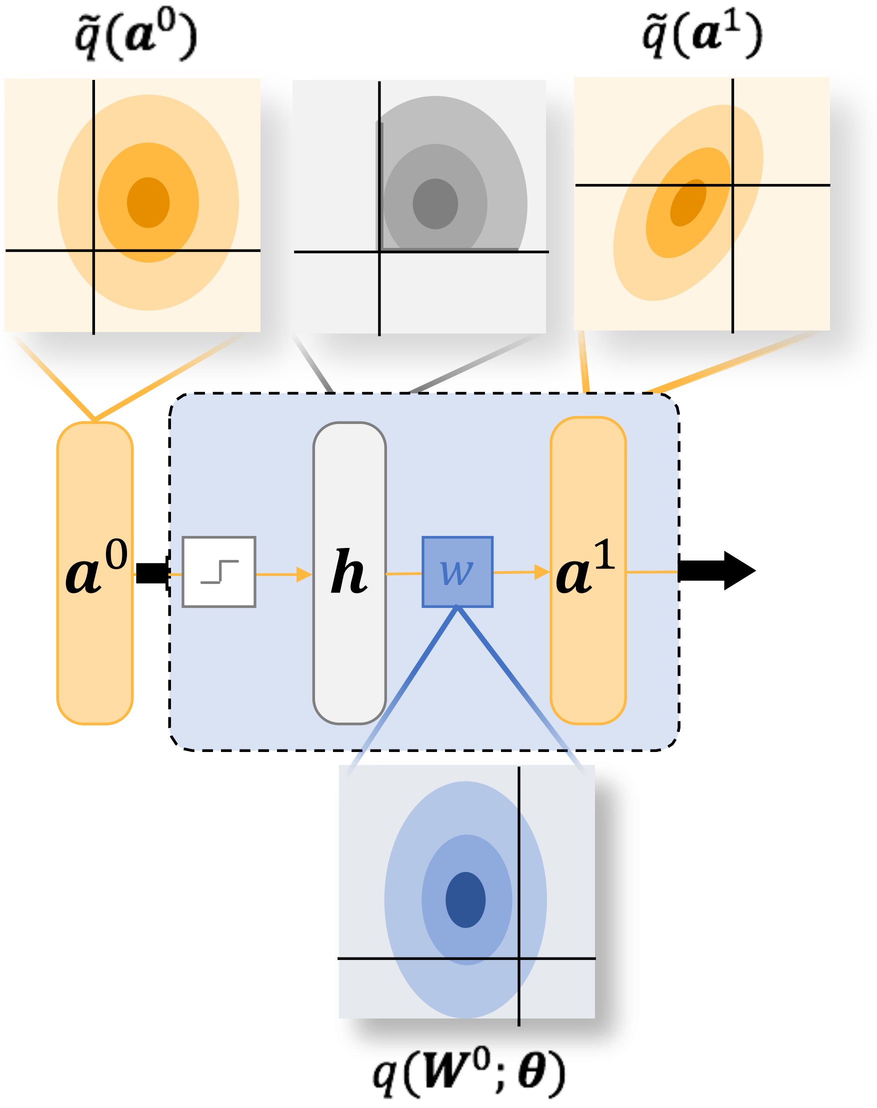
$\tilde{q}(\boldsymbol{a}^0)$ $q(\boldsymbol{W}^0;\boldsymbol{\theta})$ $\tilde{q}(\boldsymbol{a}^1)$

bnn.heaviside_layer() =

Example: 2-dimensional $\boldsymbol{a}^1$

$w_{ij} \sim \mathcal{N}^{2\times d}$ $\boldsymbol{h} \sim \text{truncated}\mathcal{N}^d$

Challenge I: Deterministic Propagation of Uncertainties

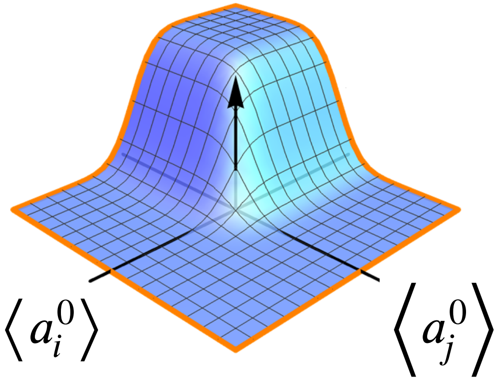


$$\boldsymbol{a}^1 \sim \mathcal{N}(\boldsymbol{\mu}^1, \boldsymbol{\Sigma}^1)$$

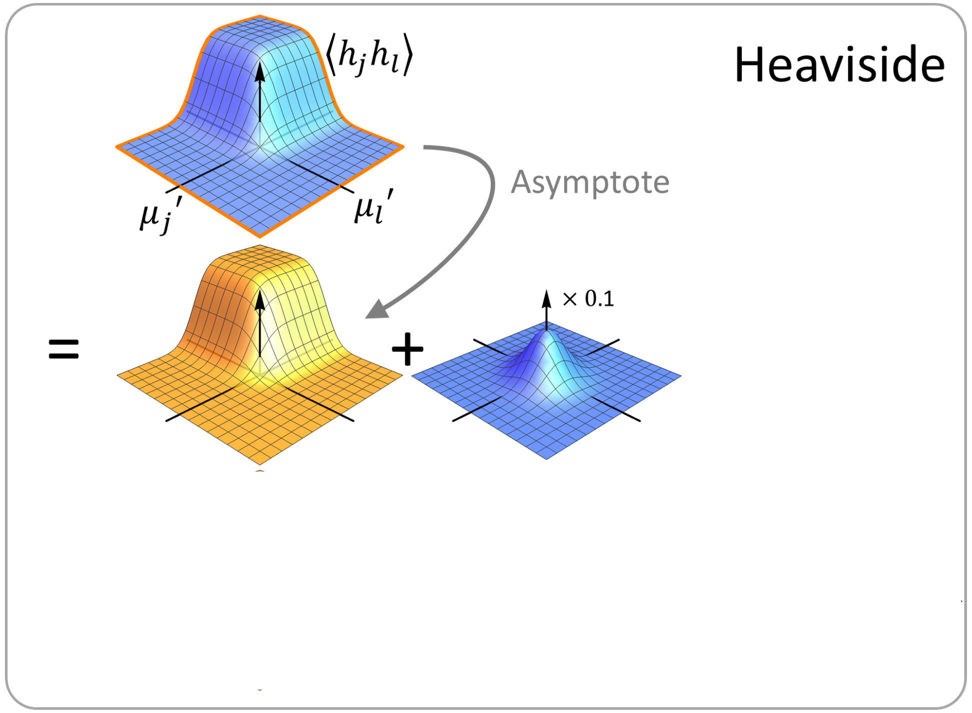
Just need moments: $\langle h_i \rangle, \langle h_i h_j \rangle$

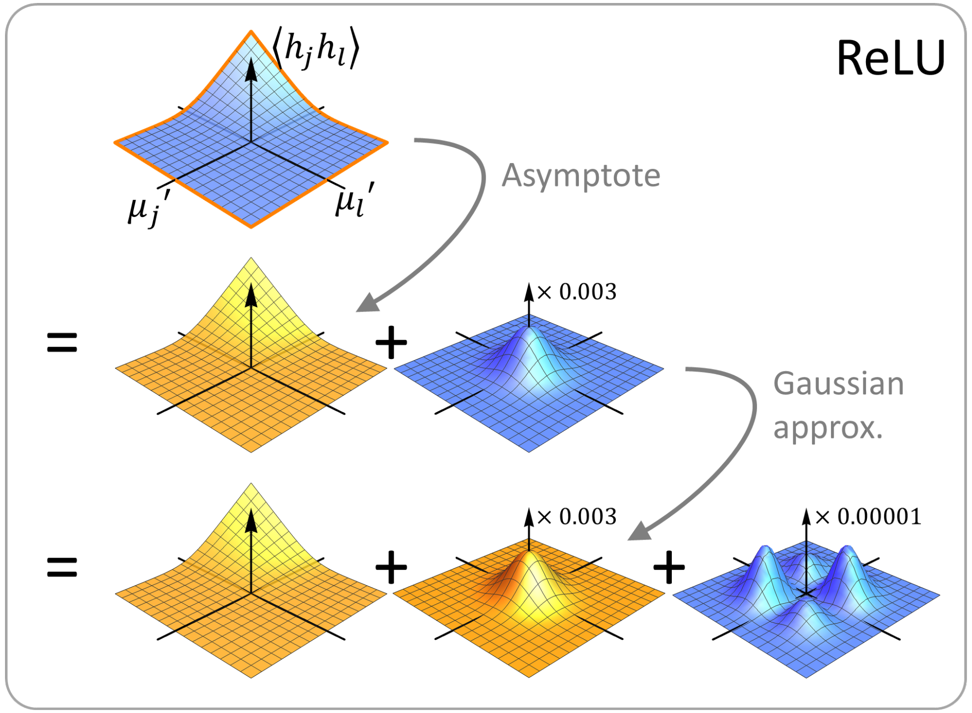
$$\langle h_i \rangle = \mathbb{E}_{\boldsymbol{a}^0 \sim \mathcal{N}(\boldsymbol{\mu}^0, \boldsymbol{\Sigma}^0)}\left[f\left(a_i^0\right)\right] = \int f(\alpha)\phi\left(\frac{\alpha - \langle a_i^0 \rangle}{\Sigma_{ii}^0}\right) d\alpha$$

$\langle h_i h_j \rangle =$



Challenge I: Deterministic Propagation of Uncertainties

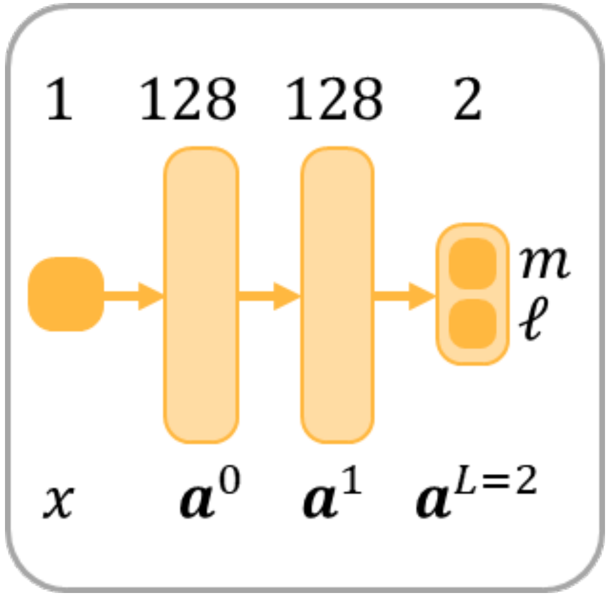


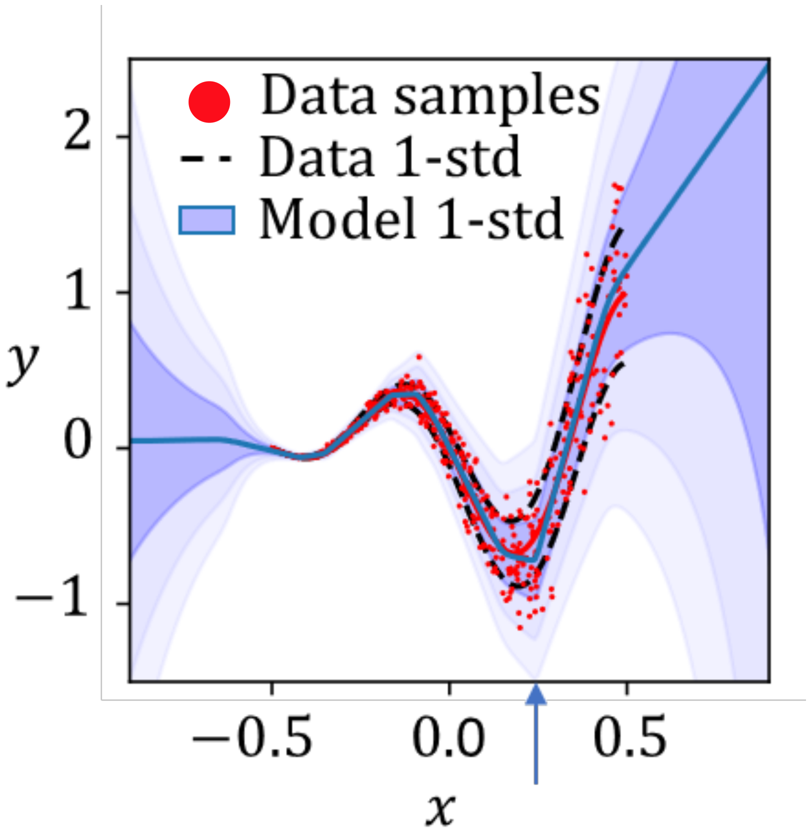


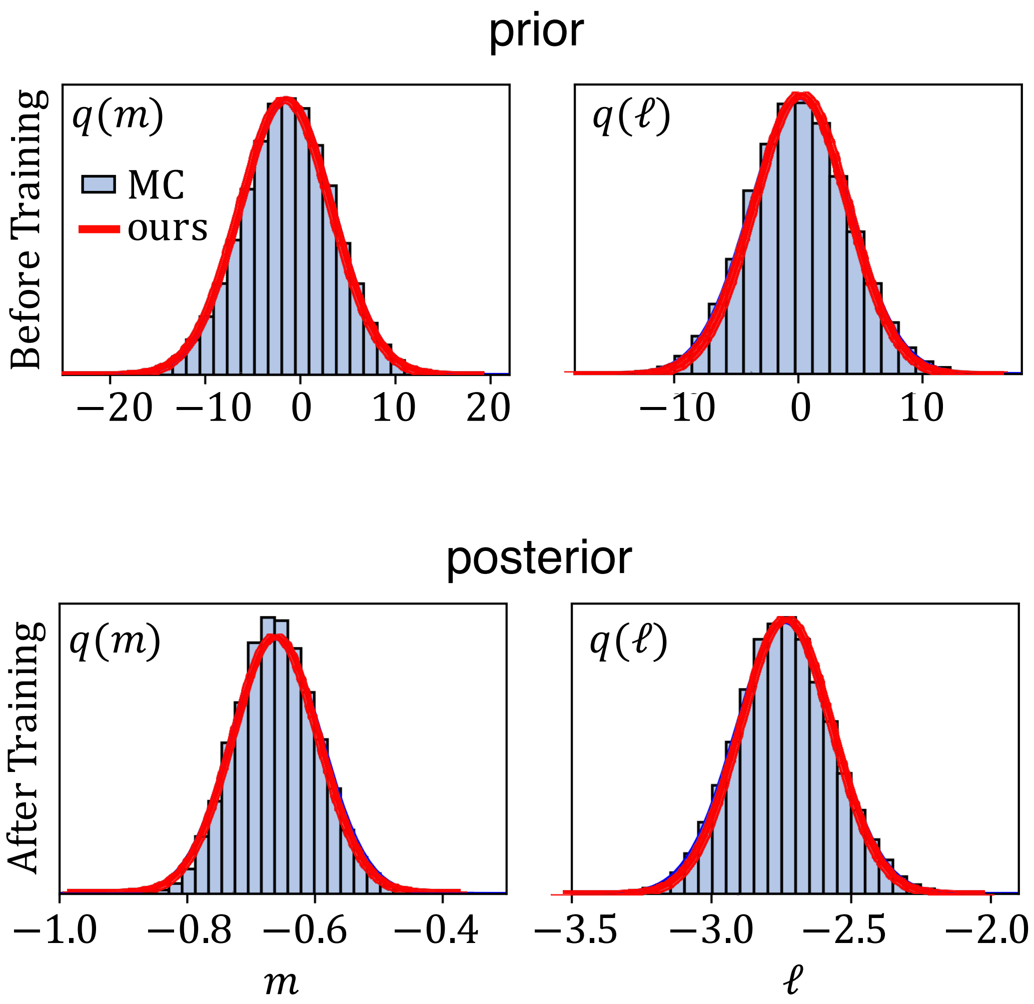

```
def heaviside_covariance(x):
    mu1 = tf.expand_dims(mu, 2)
    mu2 = tf.transpose(mu1, [0,2,1])

    s11s22 = tf.expand_dims(x_var_diag, axis=2) * tf.expand_dims(x_var_diag, axis=1)
    rho = x.var / (tf.sqrt(s11s22))# + EPSILON)
    rho = tf.clip_by_value(rho, -1/(1+EPSILON), 1/(1+EPSILON))

    return bu.heavy_g(rho, mu1, mu2)
```

```
def relu_covariance(x):
    mu1 = tf.expand_dims(mu, 2)
    mu2 = tf.transpose(mu1, [0,2,1])

    s11s22 = tf.expand_dims(x_var_diag, axis=2) * tf.expand_dims(x_var_diag, axis=1)
    rho = x.var / (tf.sqrt(s11s22))# + EPSILON)
    rho = tf.clip_by_value(rho, -1/(1+EPSILON), 1/(1+EPSILON))

    return x.var * bu.delta(rho, mu1, mu2)
```

Challenge I: Empirical Verification







Challenge II: Prior Tuning

$$\mathbb{E}_{q_\theta(w)}\left[\log p(y|x,w)\right] - D_{KL}\left[q_\theta(w)||p(w)\right]$$

Fit the data

gradient variance

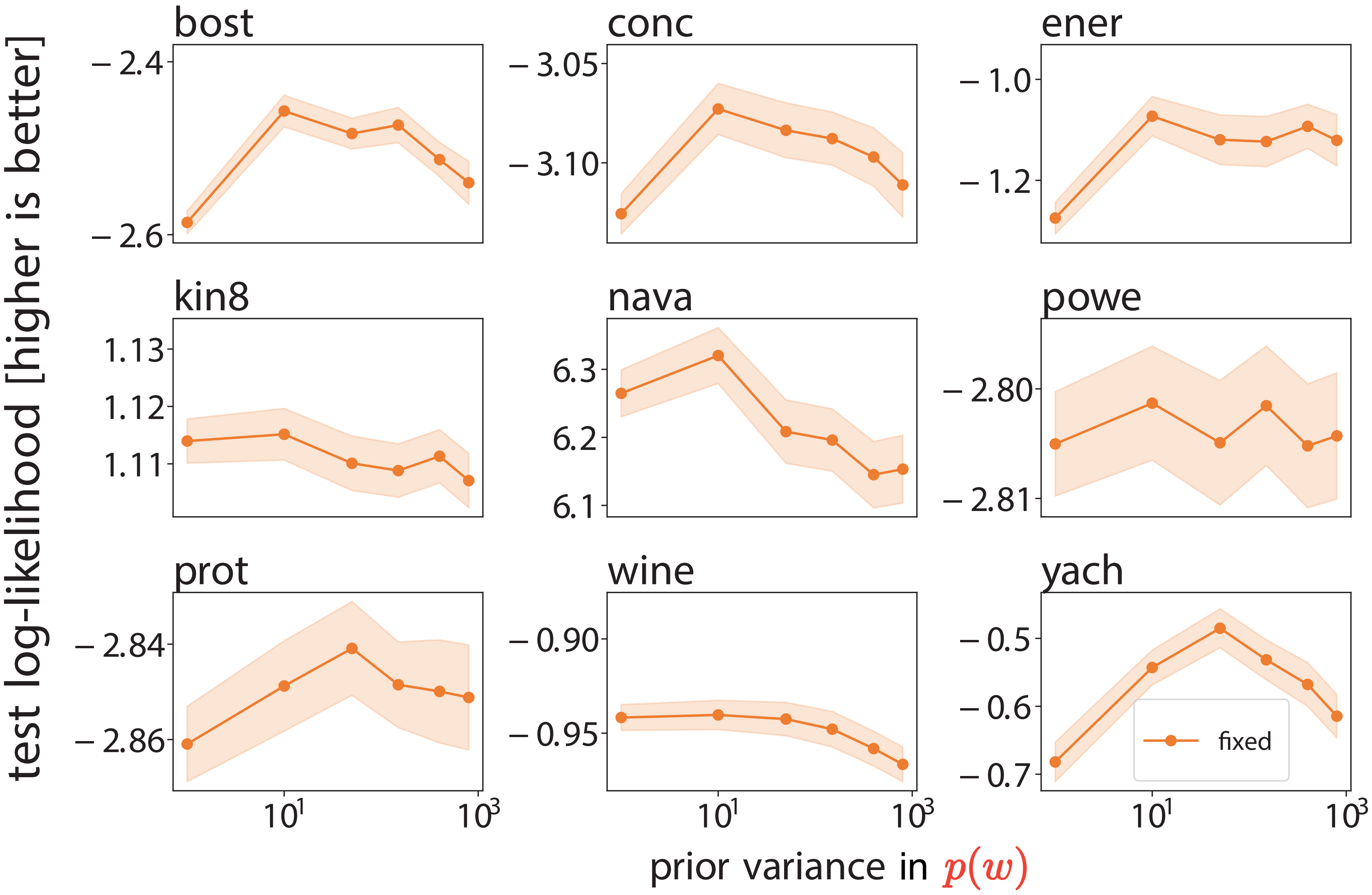
solved

Don't stray far from the prior

prior tuning

Challenge II: Prior Tuning

UCI regression datasets



Challenge II: Empirical Bayes for Prior Tuning

$$w \sim p(w|s) = \mathcal{N}(0, s)$$

prior variance

$$s \sim p(s) = \text{InvGamma}(\alpha, \beta)$$

scale

shape

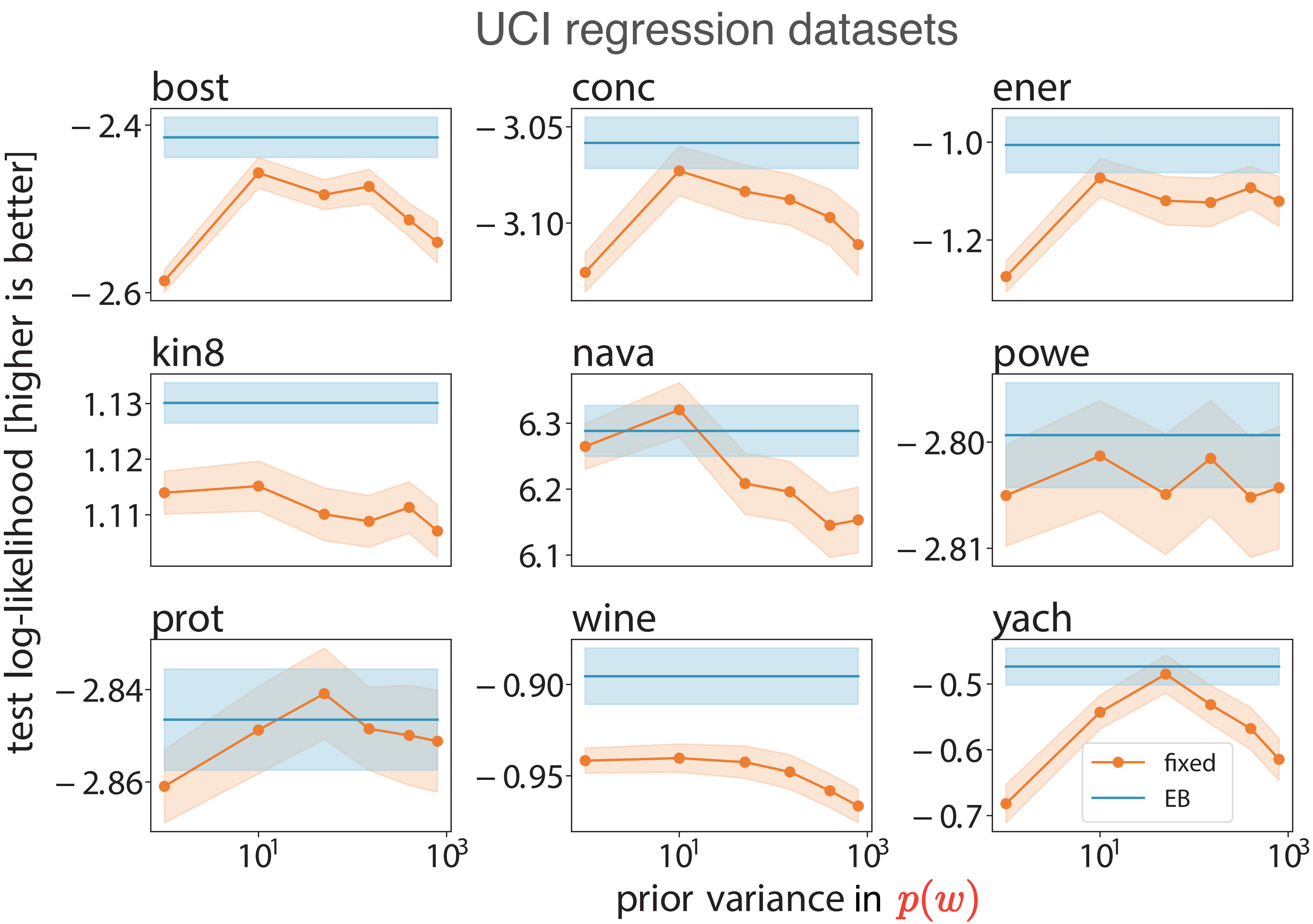
Optimize ELBO $$s^* = \underset{s}{\text{argmax}}\, ELBO(s, \theta) = \text{function}(\theta)$$

variational parameter

Empirical Bayes ELBO

$$\mathbb{E}_{q_\theta(w)}\left[\log p(y|x, w)\right] - D_{KL}\left[q_\theta(w)||p(w|s^*(\theta))\right]$$

Challenge II: Empirical Verification



Deterministic VI + Empirical Bayes

ELBO (evidence lower bound)

$$\mathbb{E}_{q_\theta(w)}\left[\log p(y|x,w)\right] - D_{KL}\left[q_\theta(w)||p(w)\right]$$

Fit the data

gradient variance

solved

Don't stray far from the prior

prior tuning

solved

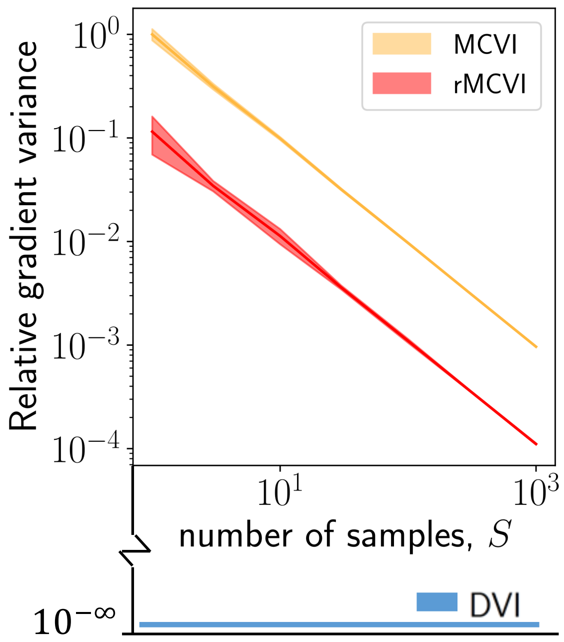
Deterministic VI + Empirical Bayes

UCI regression datasets: test log likelihood

Dataset	$\|\mathcal{D}\|$	d_x	Deterministic+EB	Deterministic+fixed	MonteCarlo+EB	MonteCarlo+fixed
bost	506	13	$\mathbf{-2.41 \pm 0.02}$	-2.46 ± 0.02	-2.46 ± 0.02	-2.48 ± 0.02
conc	1030	8	$\mathbf{-3.06 \pm 0.01}$	-3.07 ± 0.01	-3.07 ± 0.01	-3.07 ± 0.01
ener	768	8	$\mathbf{-1.01 \pm 0.06}$	-1.07 ± 0.04	-1.03 ± 0.04	-1.07 ± 0.04
kin8	8192	8	1.13 ± 0.00	1.12 ± 0.00	$\mathbf{1.14 \pm 0.00}$	1.13 ± 0.00
nava	11934	16	6.29 ± 0.04	$\mathbf{6.32 \pm 0.04}$	5.94 ± 0.05	6.00 ± 0.02
powe	9568	4	$\mathbf{-2.80 \pm 0.00}$	$\mathbf{-2.80 \pm 0.01}$	$\mathbf{-2.80 \pm 0.00}$	$\mathbf{-2.80 \pm 0.00}$
prot	45730	9	-2.85 ± 0.01	$\mathbf{-2.84 \pm 0.01}$	-2.87 ± 0.01	-2.89 ± 0.01
wine	1588	11	$\mathbf{-0.90 \pm 0.01}$	-0.94 ± 0.01	-0.92 ± 0.01	-0.94 ± 0.01
yach	308	6	$\mathbf{-0.47 \pm 0.03}$	-0.49 ± 0.03	-0.68 ± 0.03	-0.56 ± 0.03

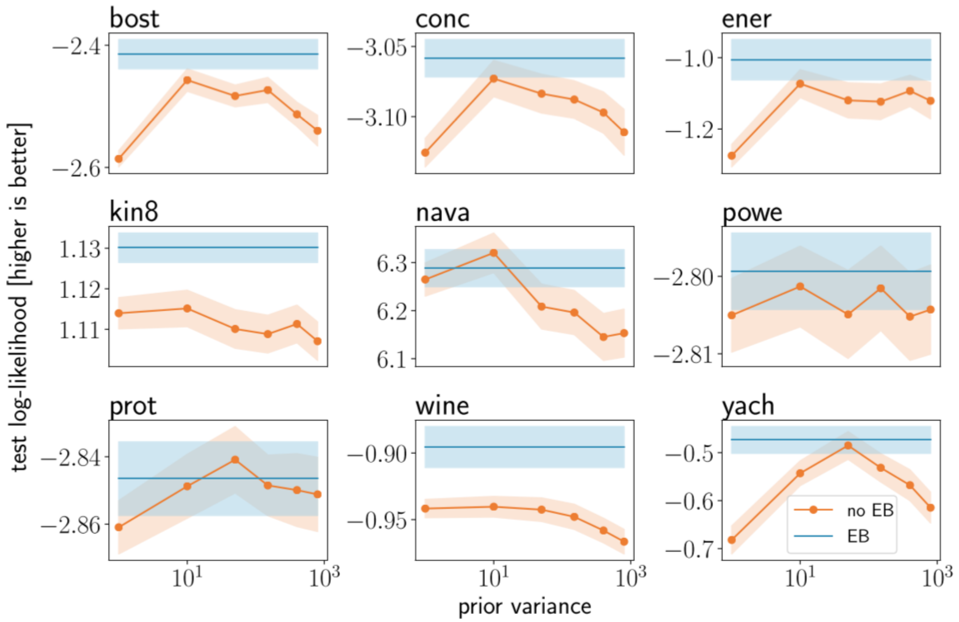
Deterministic:

Eliminate Gradient variance



Robust:

Less tuning required



Efficient:

Just a few special function calls

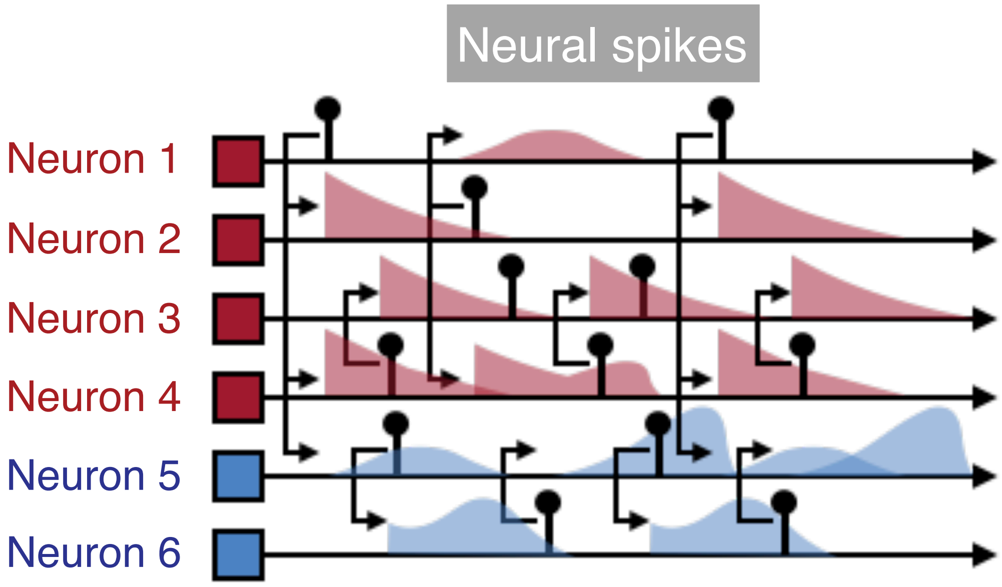
```
def heaviside_covariance(x):
    mu1 = tf.expand_dims(mu, 2)
    mu2 = tf.transpose(mu1, [0,2,1])

    s11s22 = tf.expand_dims(x_var_diag, axis=2) * tf.expand_dims(x_var_diag, axis=1)
    rho = x.var / (tf.sqrt(s11s22))# + EPSILON)
    rho = tf.clip_by_value(rho, -1/(1+EPSILON), 1/(1+EPSILON))

    return bu.heavy_g(rho, mu1, mu2)
```

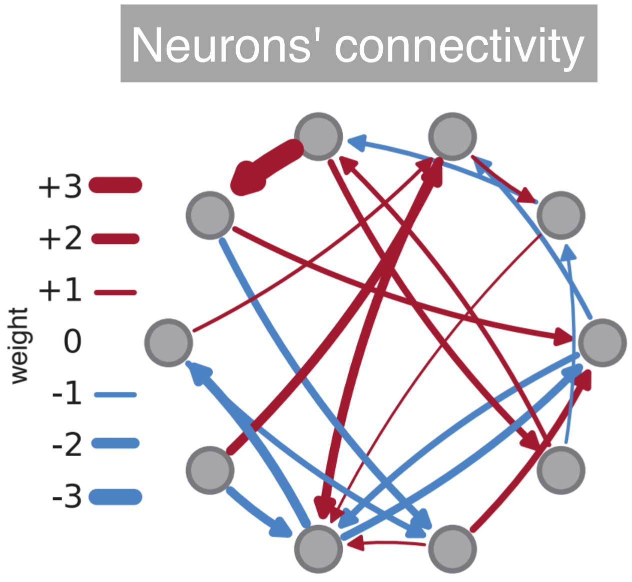
Outline

- **Determinstic variational inference** for Bayesian neural networks
  - Eliminate gradient variance in evaluating the expectation term
  - Empirical Bayes to avoid the prior tuning (*general approach*)

- **Variational importance sampling** for partially observed multivariate Hawkes process
  - VIS provides a tighter bound than ELBO (*general approach*)
  - Novel forward-backward approximate distribution

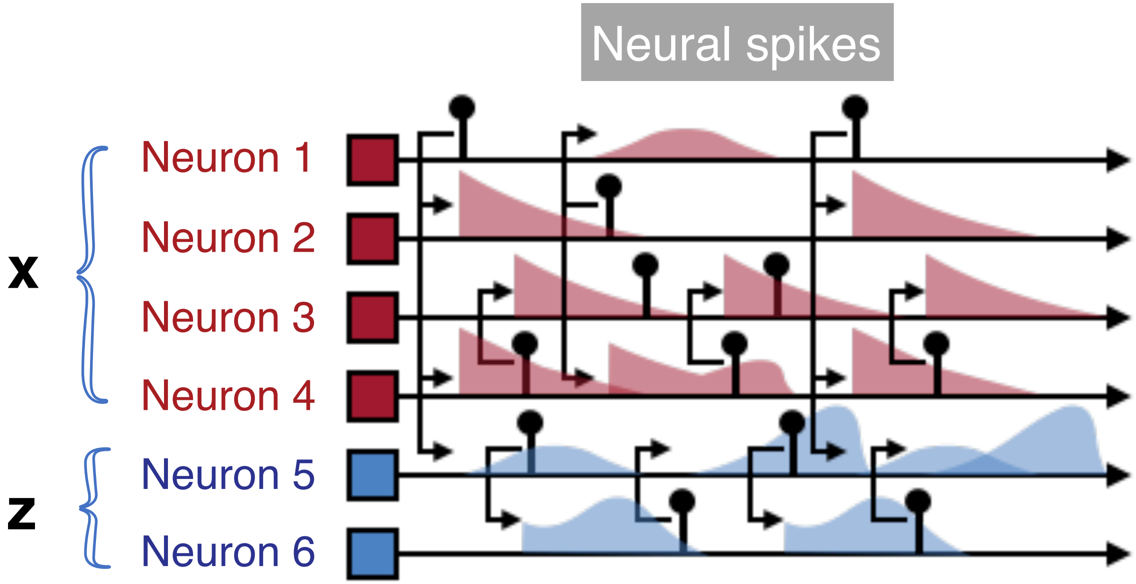
Partially observed multivariate Hawkes process (POMHP)

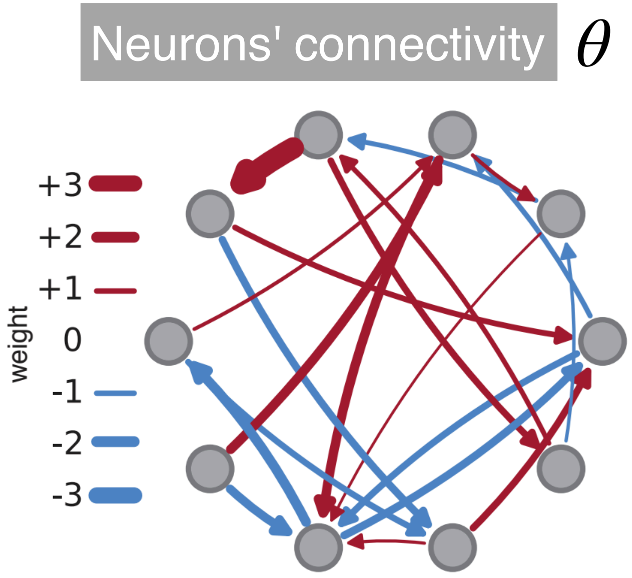




- Hawkes process is a self-exciting point process to describe neural spiking time.
- In a multivariate Hawkes process, each event can influence the occurrence of future events, **not just in the same dimension but also in other dimensions**.
- Partially observed means some events might be **hidden or unobserved**.
- Applications: finance, social networks, **neuroscience**, and so forth.

Variational Inference





- Events from observed neurons, denoted as **x**.
- Events from unobserved neurons, denoted as **z**.
- Maximum likelihood estimation to maximize the marginal **p(x;** θ **)** with respect to the model parameters θ (such as **connectivity weights**).

intractable!

Variational Inference

- Maximize

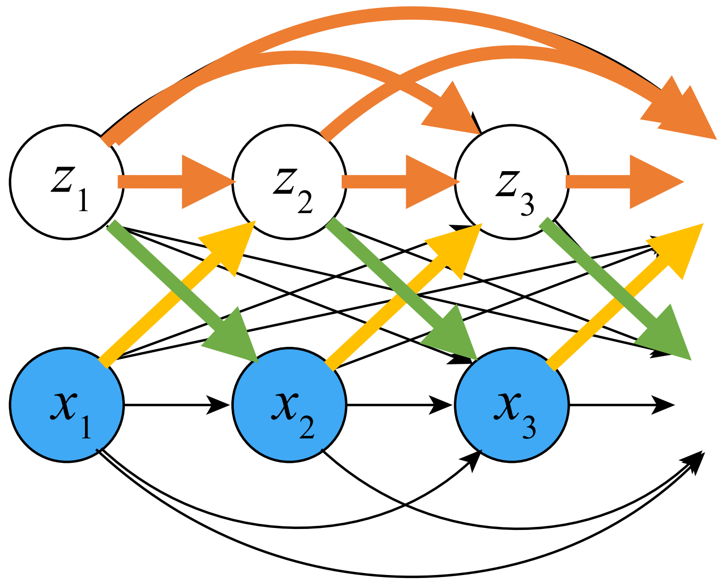
$$\log p(x;\theta) = \log \int p(x,z;\theta)dz = \log \int q(z;\phi)\frac{p(x,z;\theta)}{q(z;\phi)}dz$$

(Jensen's inequality) $\geq \mathbb{E}_{q(z;\phi)}[\log p(x,z;\theta) - \log q(z;\phi)]$

(ELBO) $= \mathbb{E}_{z\sim q}[\log p(x\,|\,z,\theta)] - D_{KL}(q(z;\phi)\,||\,p(z))$

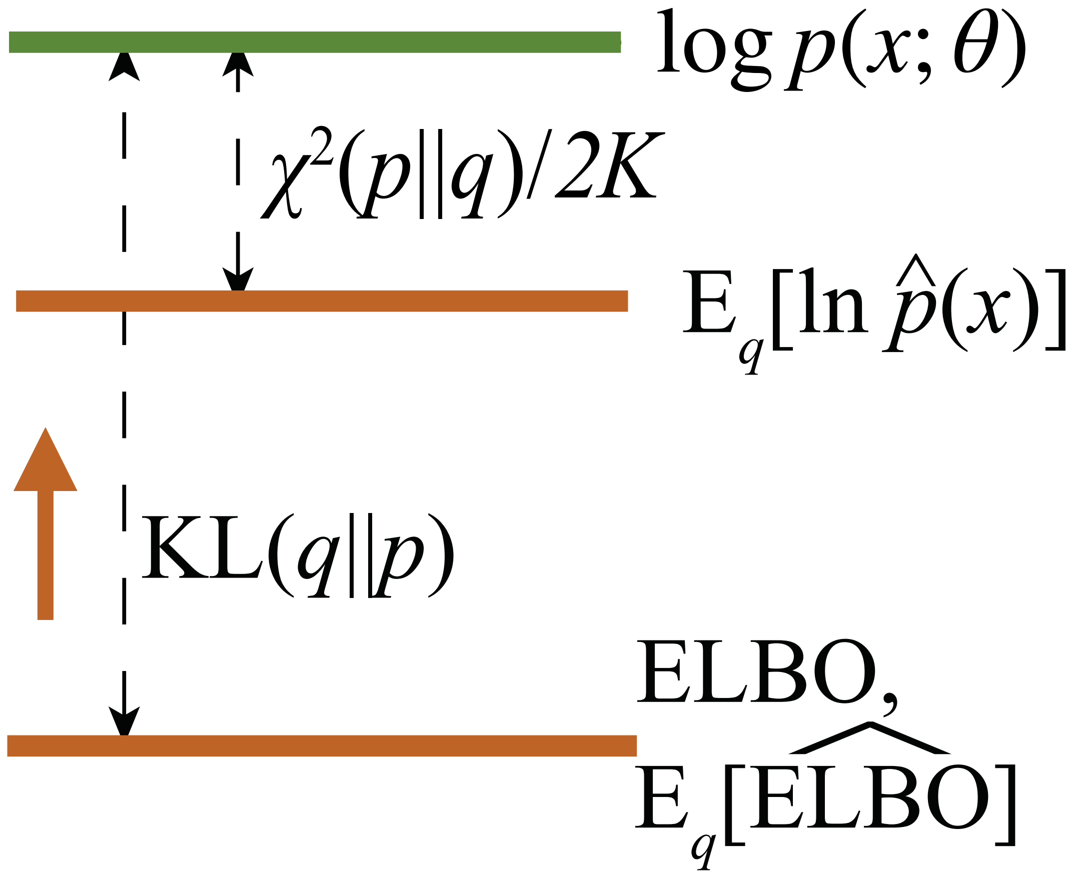
- **Two challenges**:
- ELBO doesn't always promise good parameter estimation or give the tightest lower bound, especially when a problem is very complicated like POMHP.

Variational Inference



- **Two challenges**:
- ELBO doesn't always promise good parameter estimation or give the tightest lower bound, especially when a problem is very complicated like POMHP.
- The generally chosen $q(z;\phi)$ is an MHP considering only influence from visible neurons and sampled history hidden neurons to future hidden neurons.
  - inference is slow.
  - omits the influence from hidden neurons to visible neurons.

Challenge I: Tighter Lower Bound



Variational Inference: ELBO

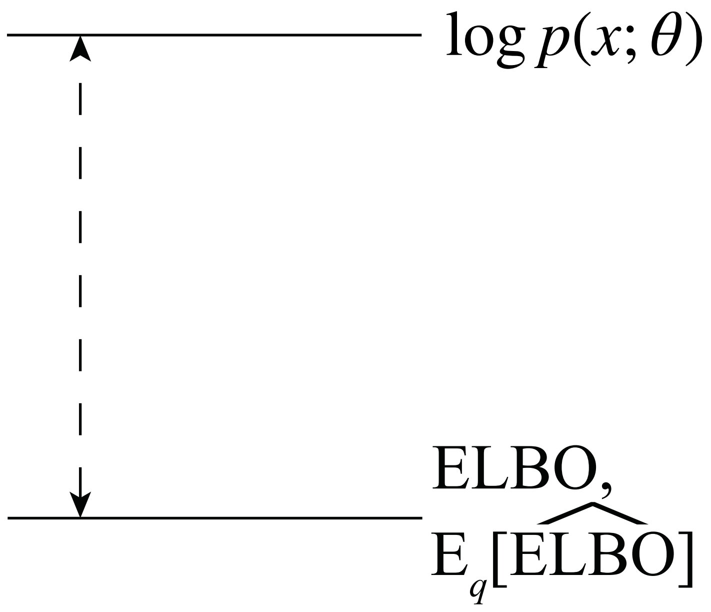
- ELBO $= \mathbb{E}_{z \sim q}[\log p(x \mid z, \theta)] - D_{KL}(q(z;\phi) \,||\, p(z))$

- $\widehat{\text{ELBO}} = \dfrac{1}{K}\sum_{k=1}^{K} [\log p(x, z^k; \theta) - \log q(z^k; \phi)]$

  where $\{z^k\}_{k=1}^{K}$ are K Monte Carlo samples from $q(z;\phi)$.

- $\widehat{\text{ELBO}}$ is an unbiased estimator of ELBO and **a down-biased estimator** of $\log p(x;\theta)$.

  i.e., $E_q[\widehat{\text{ELBO}}] = \text{ELBO} \leq \log p(x;\theta)$.



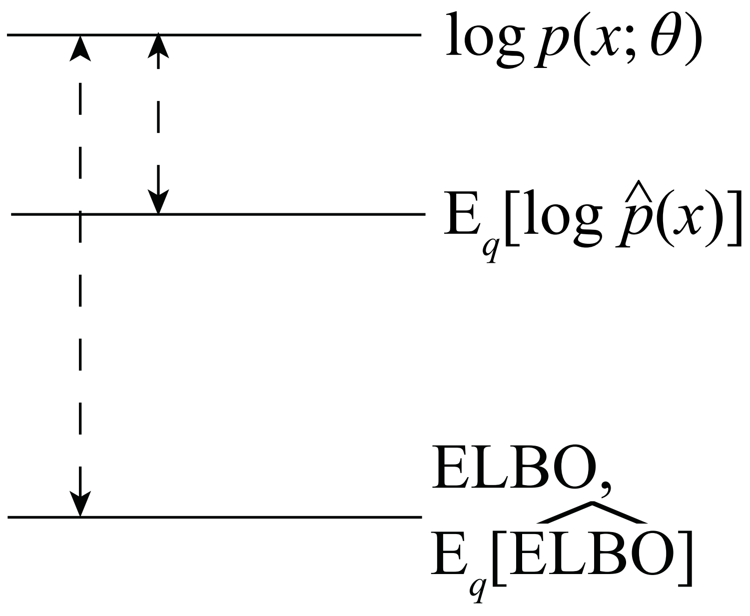
Importance Sampling

- Estimate the marginal with a proposal distribution $q(z;\phi)$

$$p(x;\theta) = \int p(x,z;\theta)dz = \int q(z;\phi)\frac{p(x,z;\theta)}{q(z;\phi)}dz$$

$$\approx \frac{1}{K}\sum_{k=1}^{K}\frac{p(x,z^k;\theta)}{q(z^k;\phi)} =: \hat{p}(x;\theta)$$

where $\{z^k\}_{k=1}^{K}$ are K Monte Carlo samples from $q(z;\phi)$.



- Since $E_q[\hat{p}(x;\theta)] = \frac{1}{K}KE_q[\frac{p(x,z;\theta)}{q(z;\phi)}] = \int p(x,z;\theta)dz = p(x;\theta)$

$\hat{p}(x;\theta)$ is an unbiased estimator of $p(x;\theta)$.

- Moreover, given Jensen's inequality $E_q[\log \hat{p}(x;\theta)] \leq \log E_q[\hat{p}(x;\theta)] = \log p(x;\theta)$

$\log \hat{p}(x;\theta)$ is **a down-biased estimator** of $\log p(x;\theta)$.

VI vs IS

- The bias of $\widehat{\text{ELBO}}$

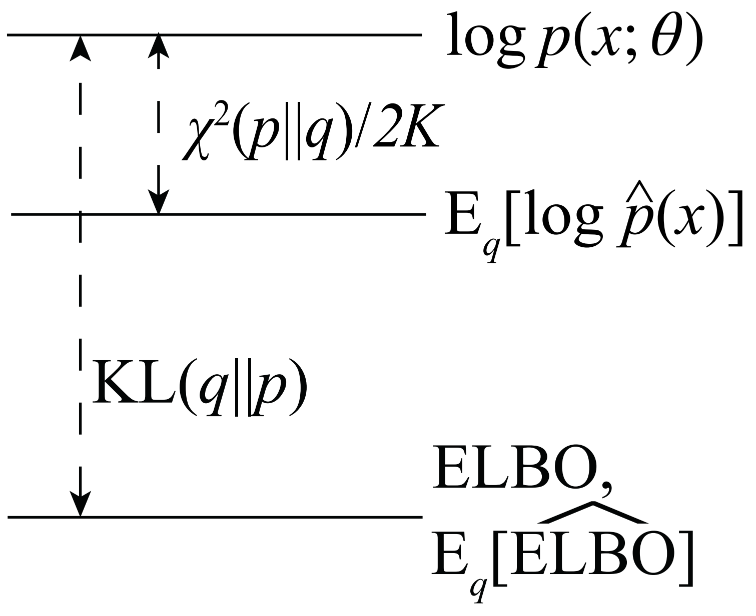
$$E_q[\widehat{\text{ELBO}} - \log p(x;\theta)] = \text{ELBO} - \log p(x;\theta)$$

$$= -D_{KL}(q(z;\phi)\,||\,p(z\,|\,x,\theta))$$

- The bias of $\log \hat{p}(x;\theta)$ [Struski et al 2022]

$$E_q[\log \hat{p}(x;\theta) - \log p(x;\theta)] \approx -\frac{1}{2K}\chi^2(p(z\,|\,x,\theta)\,||\,q(z;\phi))$$

which converges to 0 when $K \to \infty$.



- When K=1, $\log \hat{p}(x;\theta) = \widehat{\text{ELBO}}$. Thus, $\log \hat{p}(x;\theta)$ is an <u>asymptotically tighter</u> lower bound compared with $\widehat{\text{ELBO}}$.

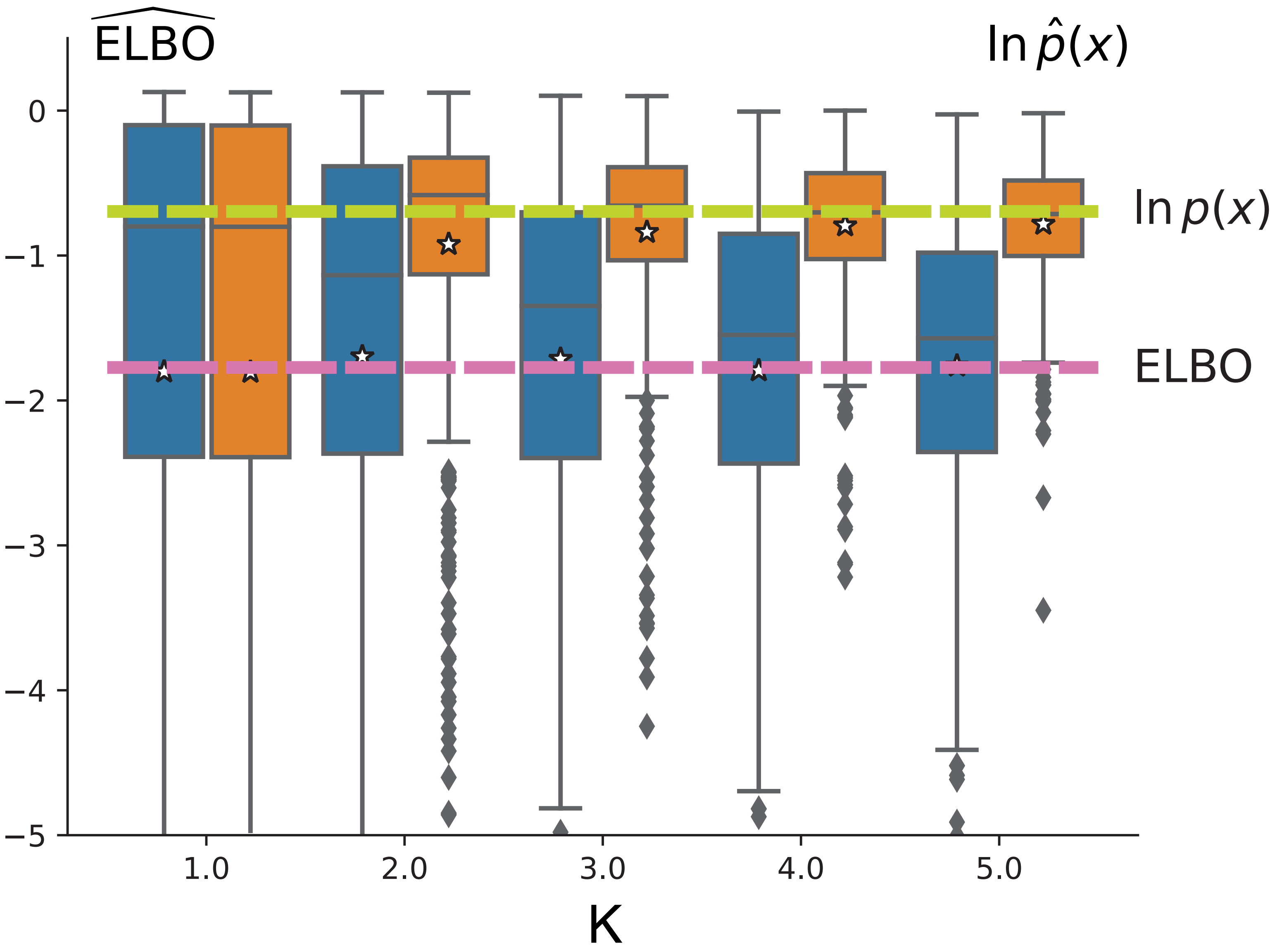
Variational Importance Sampling

Algorithm 1 variational importance sampling

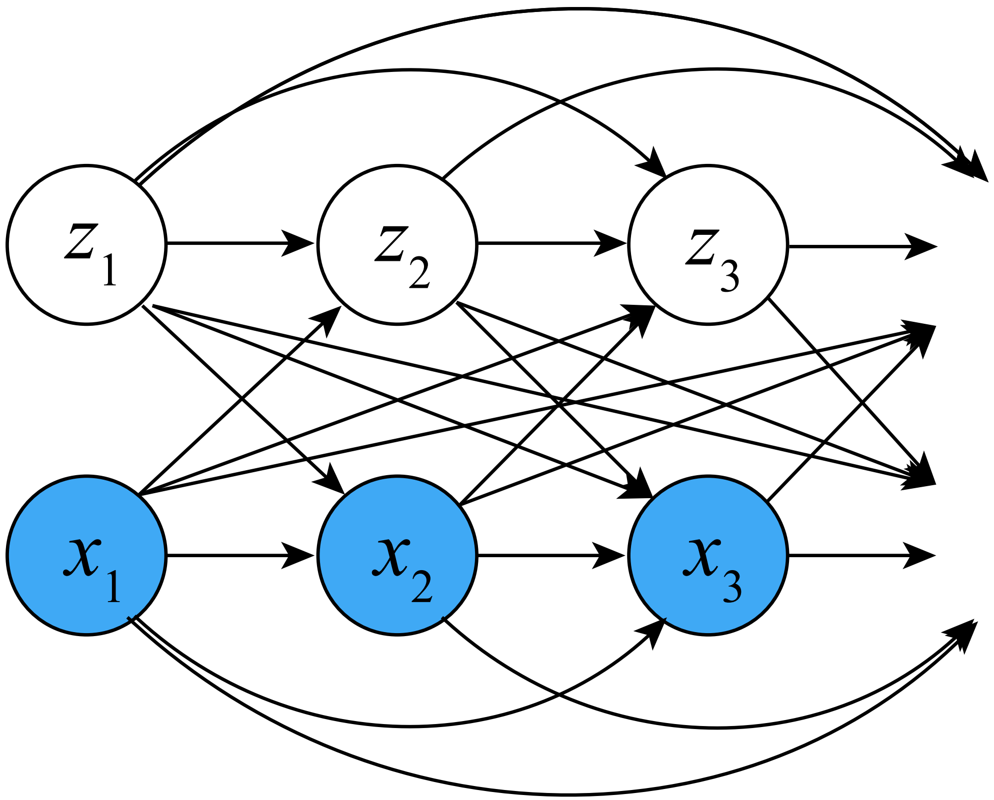
```
function VIS(x, p(x, z; θ), q(z|x; φ))
    for i = 0:N-1 do
        Sample {z^(k)}_{k=1}^K from q(z|x; φ).
        Update θ by maximizing ln p̂(x; θ).
        Update φ by minimizing χ²(p(x, z; θ)||q(z|x; φ)).
    end for
    return θ, φ.
end function
```

- Inference with **importance sampling**.
- The **proposal distribution** is from minimizing $\chi^2(p(\boldsymbol{x}, \boldsymbol{z}; \theta)||q(\boldsymbol{z}|\boldsymbol{x}; \phi))$.

Numerical Simulation



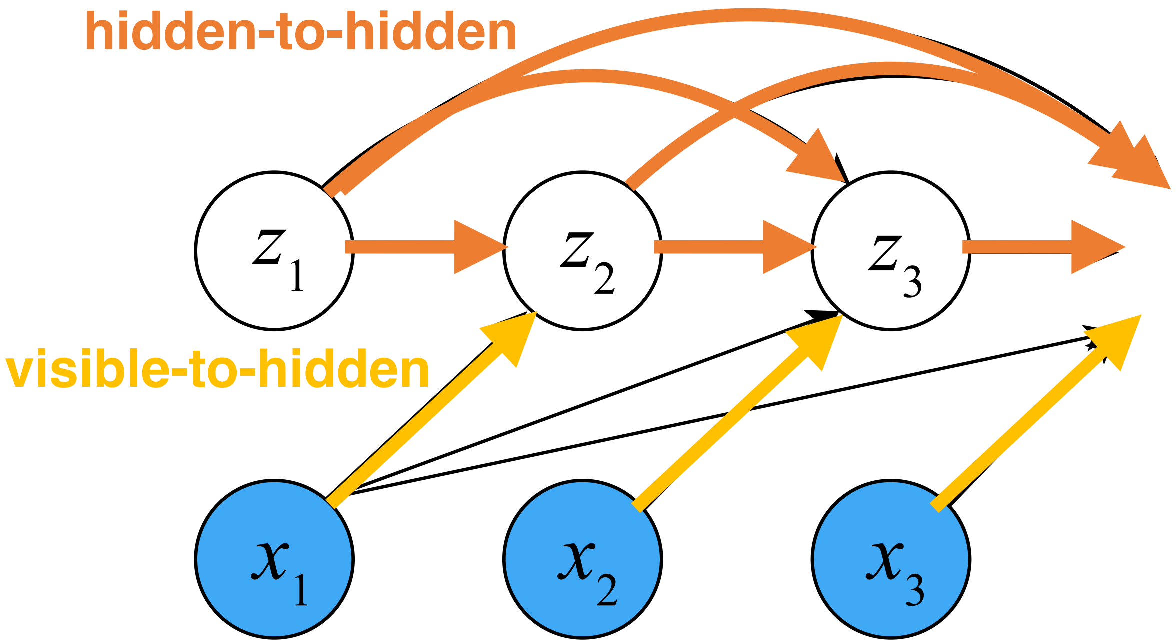
Challenge II: VIS for POMHP



The choice of the variational distribution family, $q(z \mid x; \phi)$, is important!

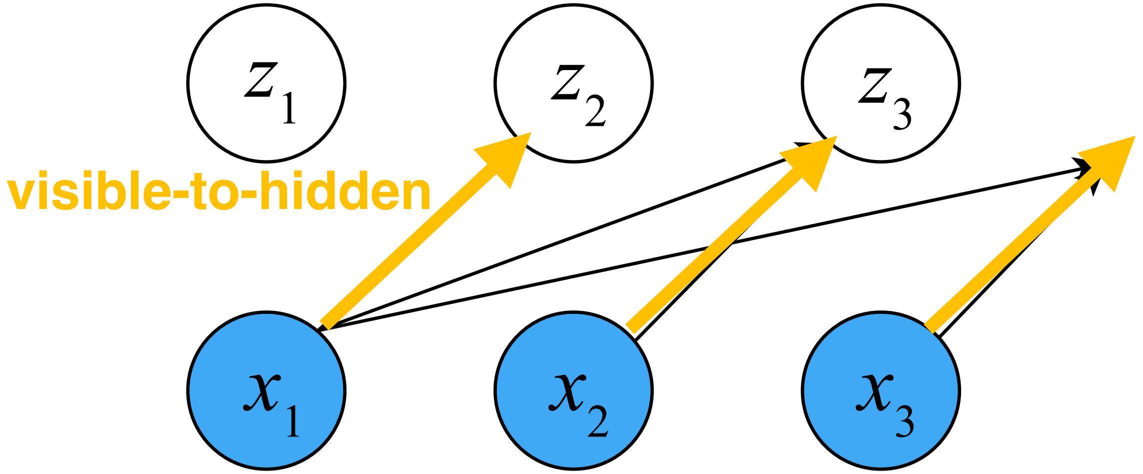
Previous choices

Forward-self sampling
to formulate $q(z \mid x; \phi)$



☹ inefficient sampling
😐 reasonable accuracy

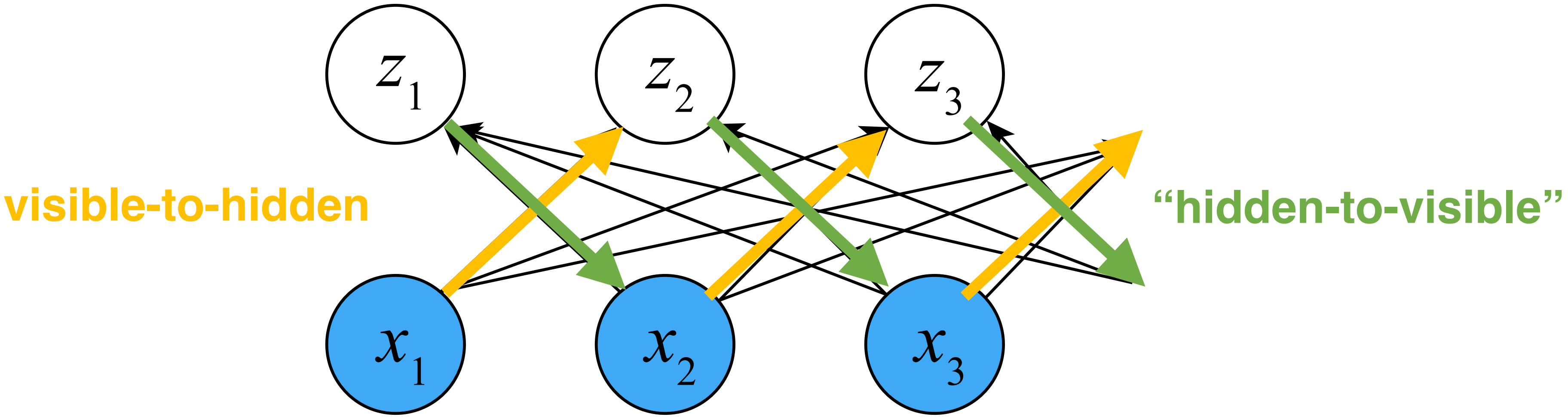
Forward sampling
to formulate $q(z \mid x; \phi)$



☺ efficient sampling
☹ low accuracy

Our choice

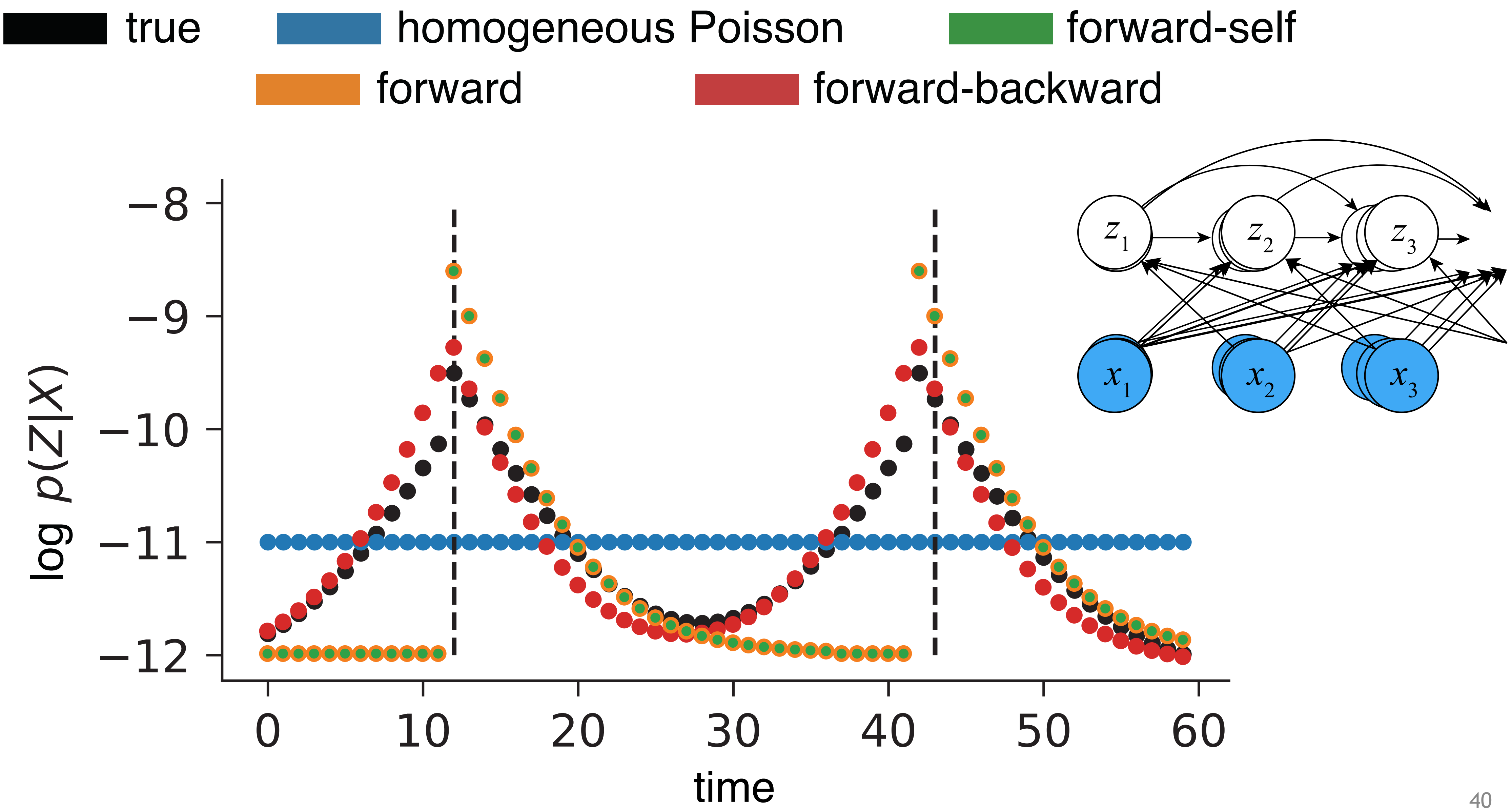
Forward-backward sampling
to formulate $q(z \mid x; \phi)$



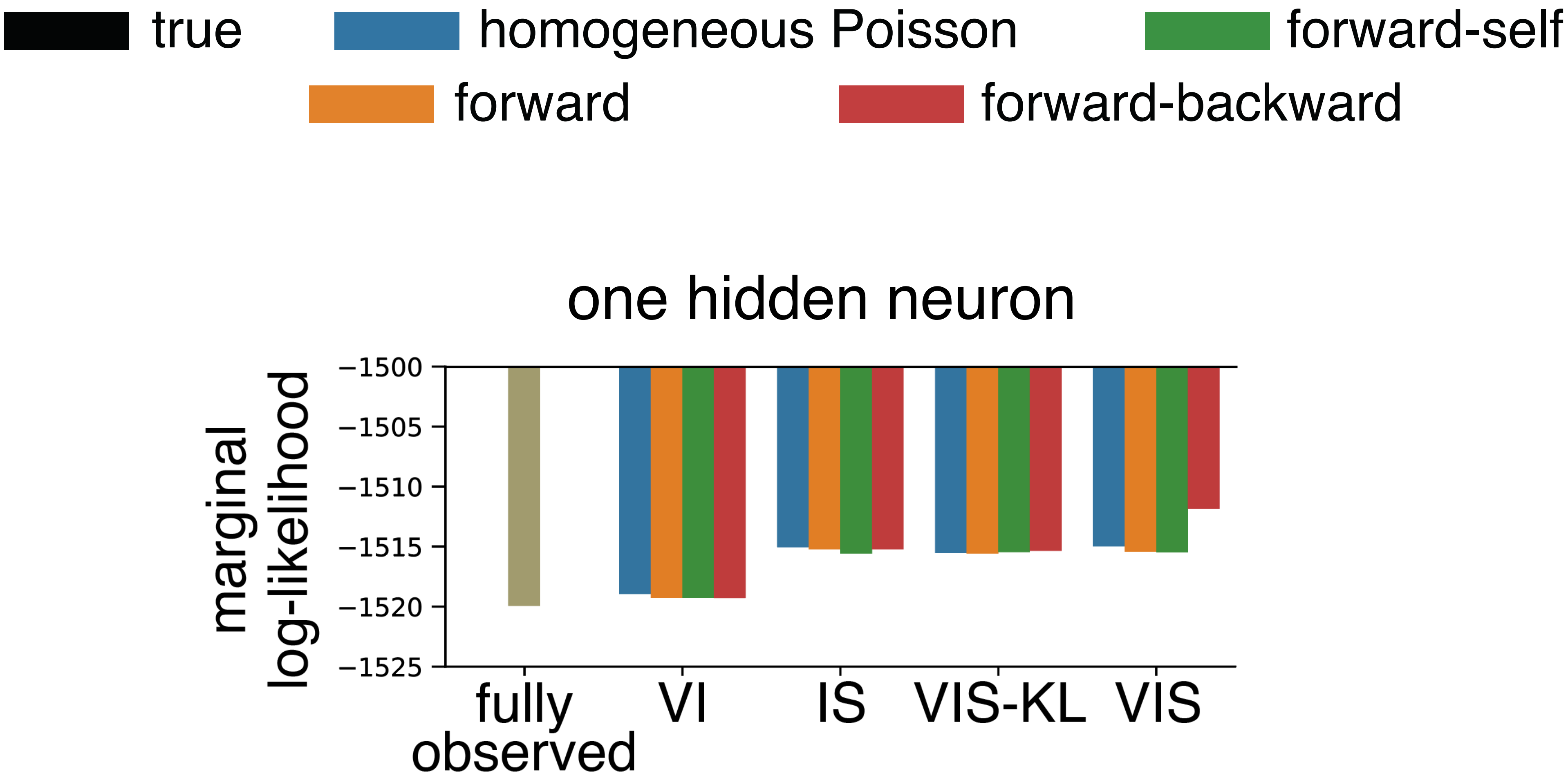
☺ efficient sampling

☺ improved accuracy

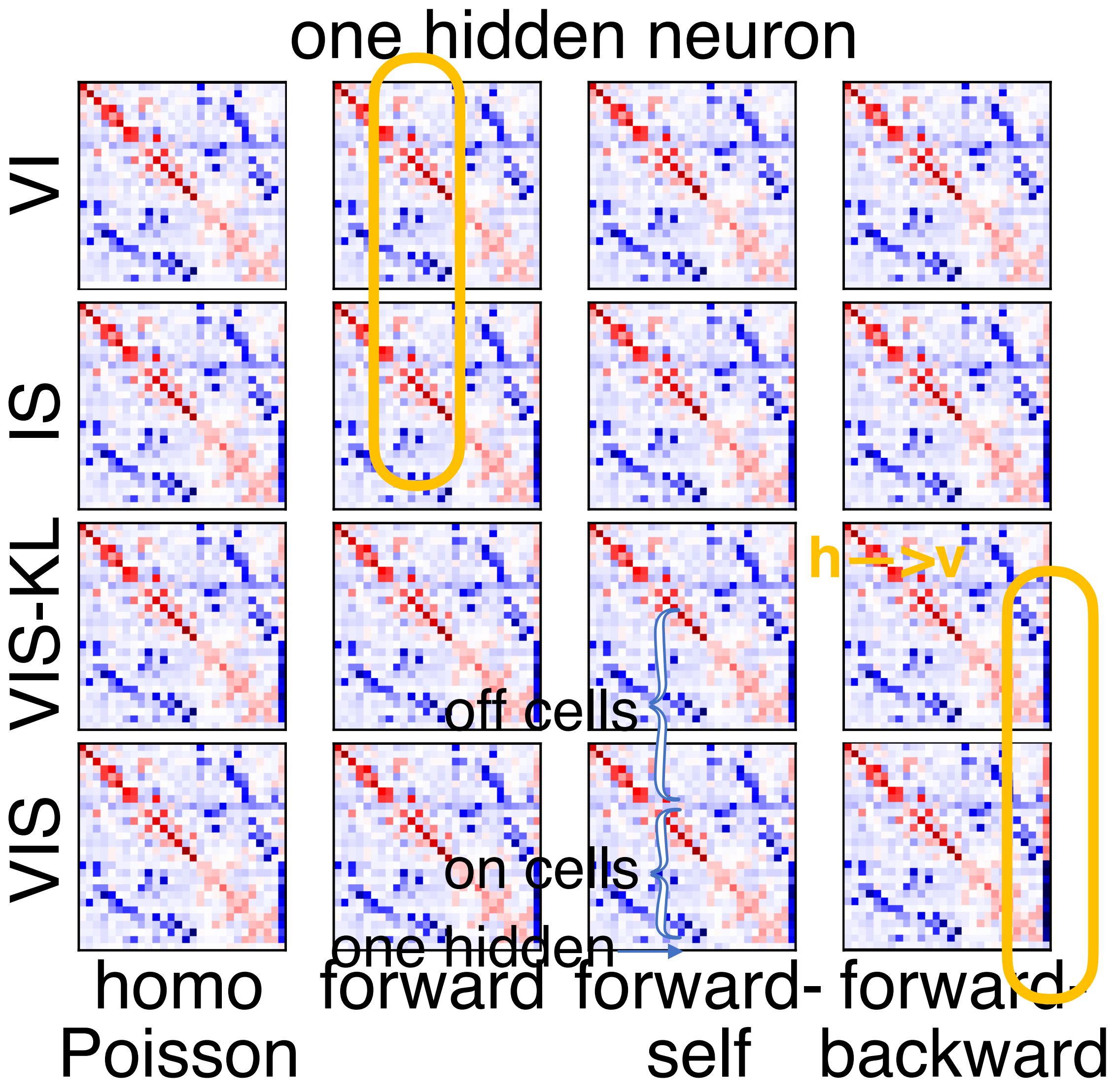
A Toy Example



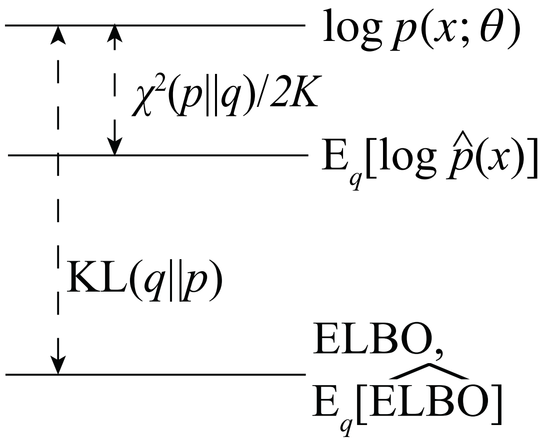
Retinal ganglion cell (RGC) dataset

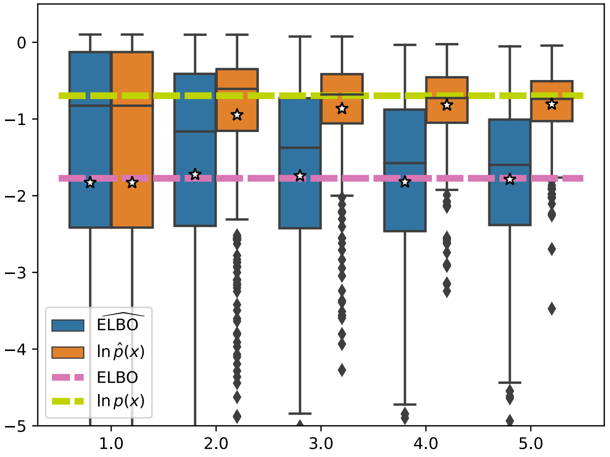


Connectivity Weights

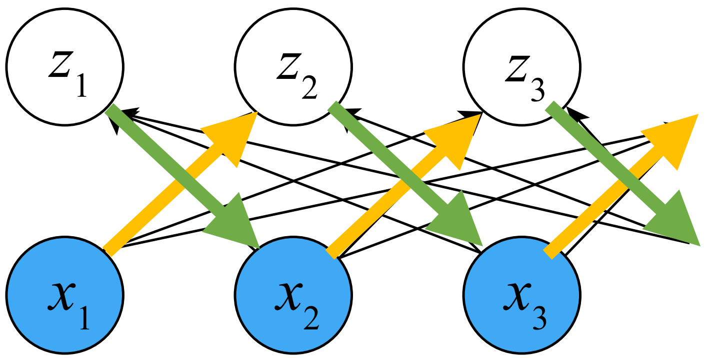


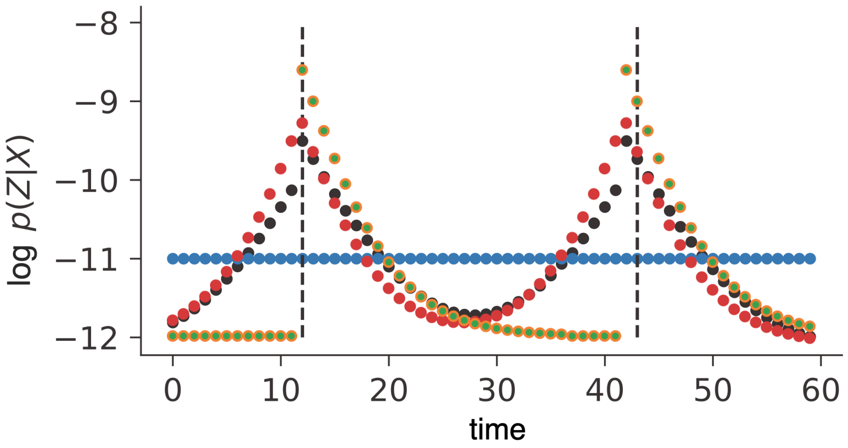
Asymptotically tighter lower bound





Better variational distribution family
Efficient parallel sampling





Acknowledgement

Collaborators:
Sebastian Nowozin
Edward Meeds
Richard E. Turner
José Miguel Hernández-Lobato
Alexander L. Gaunt

Lab members:
Chengrui Li
Yule Wang
Feiyang Wu
Weihan Li
Jingyang Ke
Mithilesh Vaidya
Zijing Wu

BRAINML

NIH National Institutes of Health

ALFRED P. SLOAN FOUNDATION

GT