Forward χ^2 Divergence Based Variational Importance Sampling

Georgia Tech

Chengrui Li, Yule Wang, Weihan Li, Anqi Wu @ GaTech CSE

ICLR 2024

Latent Variable Model

marginal likelihood ⟵ $p(\boldsymbol{x};\theta) = \int p(\boldsymbol{x},\boldsymbol{z};\theta)\,\mathrm{d}\boldsymbol{z}$ ⟶ latent variable

observed variable, intractable integral, complete likelihood, parameter set

Maximum likelihood estimation (MLE): $\hat{\theta} = \arg\max_\theta p(\boldsymbol{x};\theta)$

VI and IS, and their biases

	variational inference (VI)	importance sampling (IS)
$q(\boldsymbol{z}\|\boldsymbol{x};\phi)$	variational distribution	porposal distribution
target function	$\mathrm{ELBO}(\boldsymbol{x};\theta,\phi)$	$\ln p(\boldsymbol{x};\theta)$
numerical estimator	$\widehat{\mathrm{ELBO}}(\boldsymbol{x};\theta,\phi)$	$\ln \hat{p}(\boldsymbol{x};\theta,\phi)$

$$\mathrm{ELBO}(\boldsymbol{x};\theta,\phi) = \mathbb{E}_q[\ln p(\boldsymbol{x},\boldsymbol{z};\theta) - \ln q(\boldsymbol{z}|\boldsymbol{x};\phi)]$$

$$\widehat{\mathrm{ELBO}}(\boldsymbol{x};\theta,\phi) = \frac{1}{K}\sum_{k=1}^{K}\left[\ln p\left(\boldsymbol{x},\boldsymbol{z}^{(k)};\theta\right) - \ln q\left(\boldsymbol{z}^{(k)}|\boldsymbol{x};\phi\right)\right]$$

samples $\{\boldsymbol{z}^{(k)}\}_{k=1}^{K} \sim q(\boldsymbol{z}|\boldsymbol{x};\phi)$

$$\ln p(\boldsymbol{x};\theta) = \ln \mathbb{E}_q[p(\boldsymbol{x},\boldsymbol{z};\theta)/q(\boldsymbol{z}|\boldsymbol{x};\phi)]$$

$$\ln \hat{p}(\boldsymbol{x};\theta,\phi) = \mathrm{logsumexp}\left[\ln p\left(\boldsymbol{x},\boldsymbol{z}^{(k)};\theta\right) - \ln q\left(\boldsymbol{z}^{(k)}|\boldsymbol{x};\phi\right)\right] - \ln K$$

$\ln p(\boldsymbol{x};\theta)$

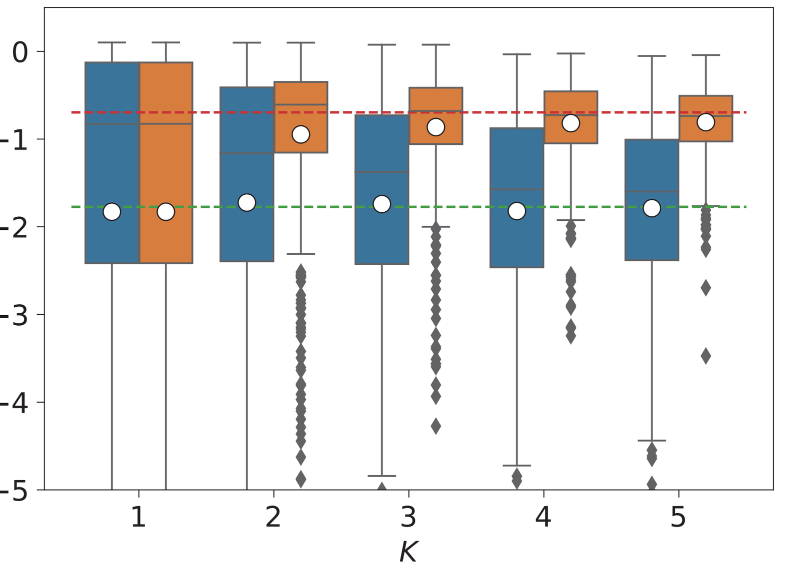
$\chi^2(p(\boldsymbol{z}|\boldsymbol{x};\theta)\|q(\boldsymbol{z}|\boldsymbol{x};\phi))/(2K)$

$\mathbb{E}_q[\ln \hat{p}(\boldsymbol{x};\theta,\phi)]$

$\mathrm{KL}(q(\boldsymbol{z}|\boldsymbol{x};\phi)\|p(\boldsymbol{z}|\boldsymbol{x};\theta))$

$\mathrm{ELBO}(\boldsymbol{x};\theta,\phi)$

$\mathbb{E}_q\left[\widehat{\mathrm{ELBO}}(\boldsymbol{x};\theta,\phi)\right]$



- Compared with $\widehat{\mathrm{ELBO}}(\boldsymbol{x};\theta,\phi)$, $\ln \hat{p}(\boldsymbol{x};\theta,\phi)$ is an asymptotically tighter lower bound of $\ln p(\boldsymbol{x};\theta)$.
- When $K=1$, $\widehat{\mathrm{ELBO}}(\boldsymbol{x};\theta,\phi) = \ln \hat{p}(\boldsymbol{x};\theta,\phi)$.
- If $K \geqslant 2$, IS estimates $\ln p(\boldsymbol{x};\theta)$ better than VI, so we can use IS to learn θ.

VIS as the best way of doing IS

In fact, the effectiveness of the IS estimator is

$$\mathrm{Var}_q\left[\hat{p}(\boldsymbol{x};\theta,\phi)\right] = \frac{p(\boldsymbol{x};\theta)^2}{K}\chi^2(p(\boldsymbol{z}|\boldsymbol{x};\theta)\|q(\boldsymbol{z}|\boldsymbol{x};\phi))$$

So, we should minimize this forward χ^2 divergence w.r.t. ϕ to find the optimal choice of the proposal distribution $q(\boldsymbol{z}|\boldsymbol{x};\phi)$ for the current $p(\boldsymbol{z}|\boldsymbol{x};\theta)$.

$$\chi^2(p(\boldsymbol{z}|\boldsymbol{x};\theta)\|q(\boldsymbol{z}|\boldsymbol{x};\phi)) = \frac{1}{p(\boldsymbol{x};\theta)^2}\int\frac{p(\boldsymbol{x},\boldsymbol{z};\theta)^2}{q(\boldsymbol{z}|\boldsymbol{x};\phi)}\mathrm{d}\boldsymbol{z} - 1 =: \frac{1}{p(\boldsymbol{x};\theta)^2}V(\boldsymbol{x};\theta,\phi) - 1$$

It is converted to minimizing $V(\boldsymbol{x};\theta,\phi)$, which should be estimated and minimized in log space for numerical stability.

$$\ln V(\boldsymbol{x};\theta,\phi) \approx \mathrm{logsumexp}\left[2\ln p\left(\boldsymbol{x},\boldsymbol{z}^{(k)};\theta\right) - 2\ln q\left(\boldsymbol{z}^{(k)}|\boldsymbol{x};\phi\right)\right] - \ln K = \ln \hat{V}(\boldsymbol{x};\theta,\phi)$$

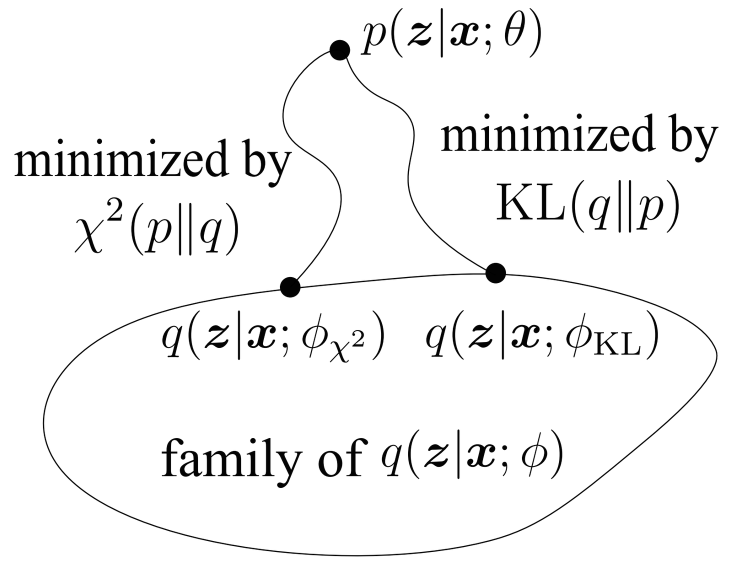
	VI	variational importance sampling (VIS)
Sample	$\{\boldsymbol{z}^{(k)}\}_{k=1}^{K} \sim q(\boldsymbol{z}\|\boldsymbol{x};\phi)$	
E-step	Minimize $\mathrm{KL}(q(\boldsymbol{z}\|\boldsymbol{x};\phi)\|p(\boldsymbol{z}\|\boldsymbol{x};\theta))$ w.r.t. ϕ by maximizing $\widehat{\mathrm{ELBO}}(\boldsymbol{x};\theta,\phi)$ w.r.t. ϕ	Minimize $\mathrm{KL}(q(\boldsymbol{z}\|\boldsymbol{x};\phi)\|p(\boldsymbol{z}\|\boldsymbol{x};\theta))$ w.r.t. ϕ by maximizing $\widehat{\mathrm{ELBO}}(\boldsymbol{x};\theta,\phi)$ w.r.t. ϕ
M-step	Maximize $\ln p(\boldsymbol{x};\theta)$ w.r.t. θ by maximizing $\widehat{\mathrm{ELBO}}(\boldsymbol{x};\theta,\phi)$ w.r.t. θ	Maximize $\ln p(\boldsymbol{x};\theta)$ w.r.t. θ by maximizing $\ln \hat{p}(\boldsymbol{x};\theta,\phi)$ w.r.t. θ

VIS only changes two lines of the code.

The score function gradient estimators in the E-step:

$$\frac{\partial\,\widehat{\mathrm{ELBO}}(\boldsymbol{x};\theta,\phi)}{\partial\phi} \approx \frac{1}{K}\sum_{k=1}^{K}\left\{\left[\ln p\left(\boldsymbol{x},\boldsymbol{z}^{(k)};\theta\right) - \ln q\left(\boldsymbol{z}^{(k)}|\boldsymbol{x};\phi\right)\right] \times \frac{\partial \ln q\left(\boldsymbol{z}^{(k)}|\boldsymbol{x};\phi\right)}{\partial\phi}\right\}$$

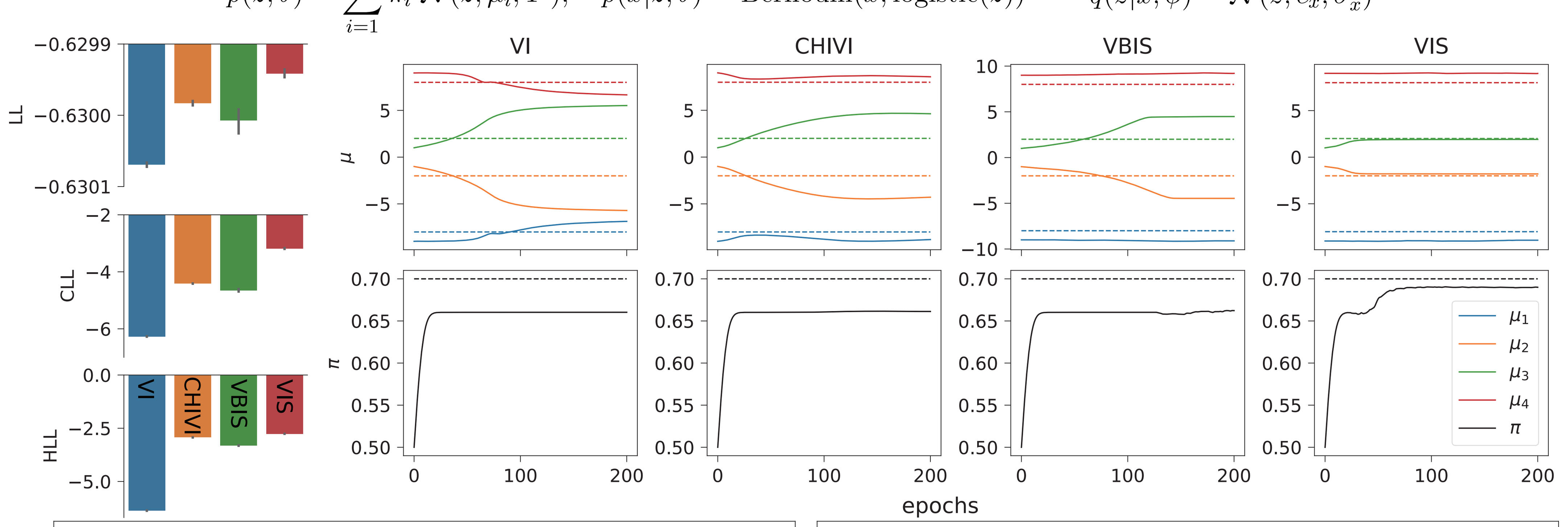
$$\frac{\partial \ln V(\boldsymbol{x};\theta,\phi)}{\partial\phi} \approx \frac{\partial}{\partial\phi}\frac{1}{2}\ln\hat{V}(\boldsymbol{x};\theta,\phi)$$

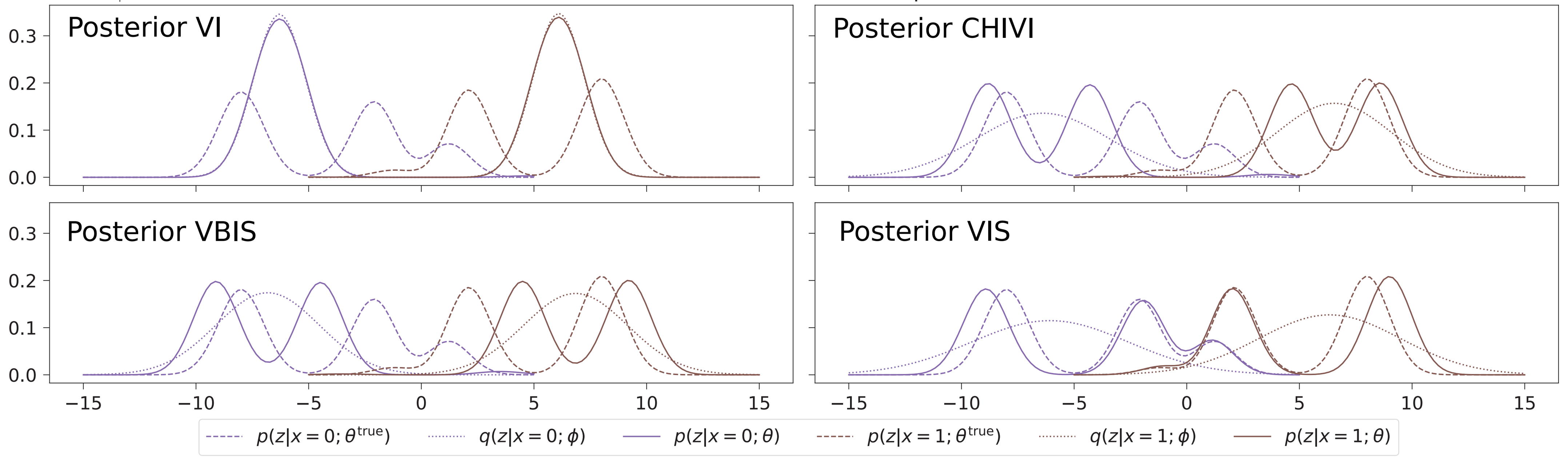


Experiments

Mixture model: GMM-Bernoulli

$p(z;\theta) = \sum_{i=1}^{4}\pi_i\,\mathcal{N}(z;\mu_i,1^2)$, $p(x|z;\theta) = \mathrm{Bernoulli}(x;\mathrm{logistic}(z))$ $q(z|x;\phi) = \mathcal{N}(z;c_x,\sigma_x^2)$



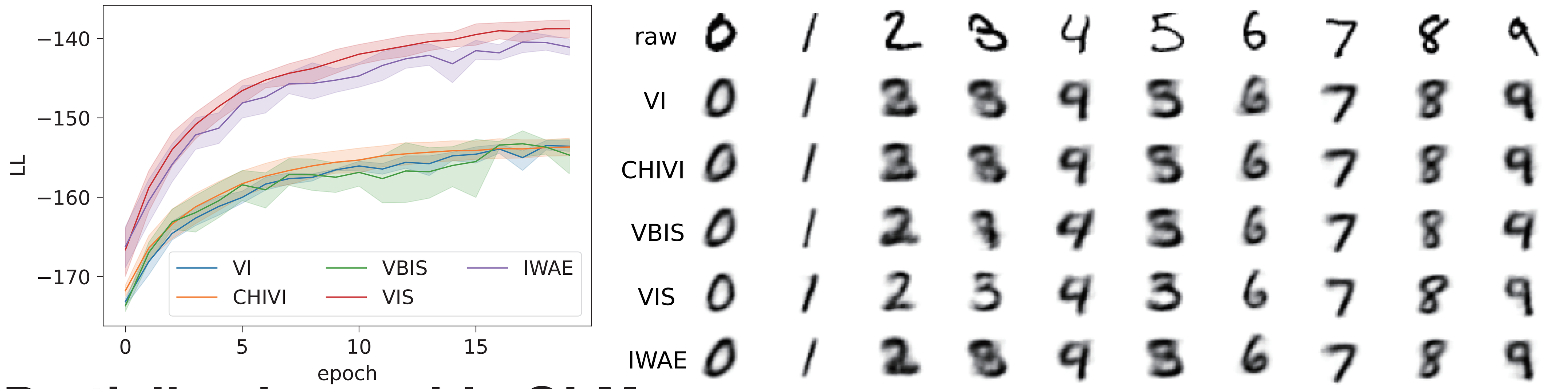


- VI: zero-forcing/mode-seeking behavior of minimizing the reverse KL. Good ELBO, reverse KL is nearly 0, but in fact both $p(z|x;\theta)$ and $q(z|x;\phi)$ are far from $p(z|x;\theta^{\mathrm{true}})$.
- VIS: mass-covering/mean-seeking behavior of minimizing the forward χ^2. This enlarges the effective support range of $q(z|x;\phi)$ for sampling.

VAE on MNIST

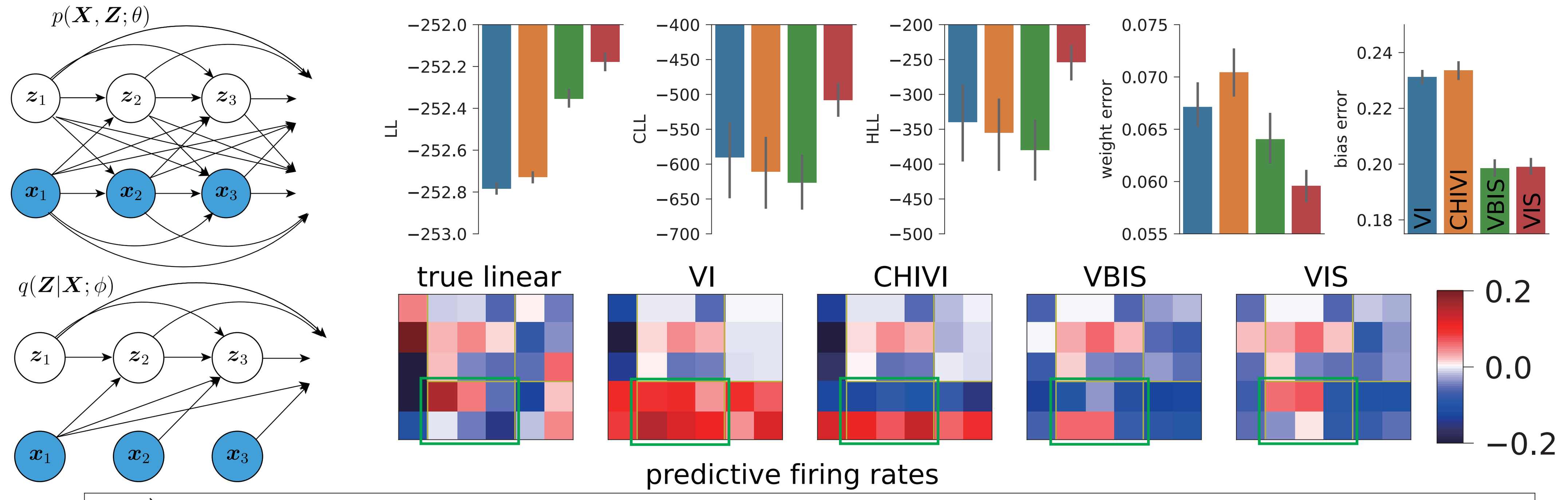
$p(\boldsymbol{z};\theta) = \mathcal{N}(\boldsymbol{z};\boldsymbol{0},\boldsymbol{I})$, $p(\boldsymbol{x}|\boldsymbol{z};\theta) = \mathrm{Bernoulli}(\boldsymbol{x};\mathrm{logistic}(\mathrm{decoder}(\boldsymbol{z})))$

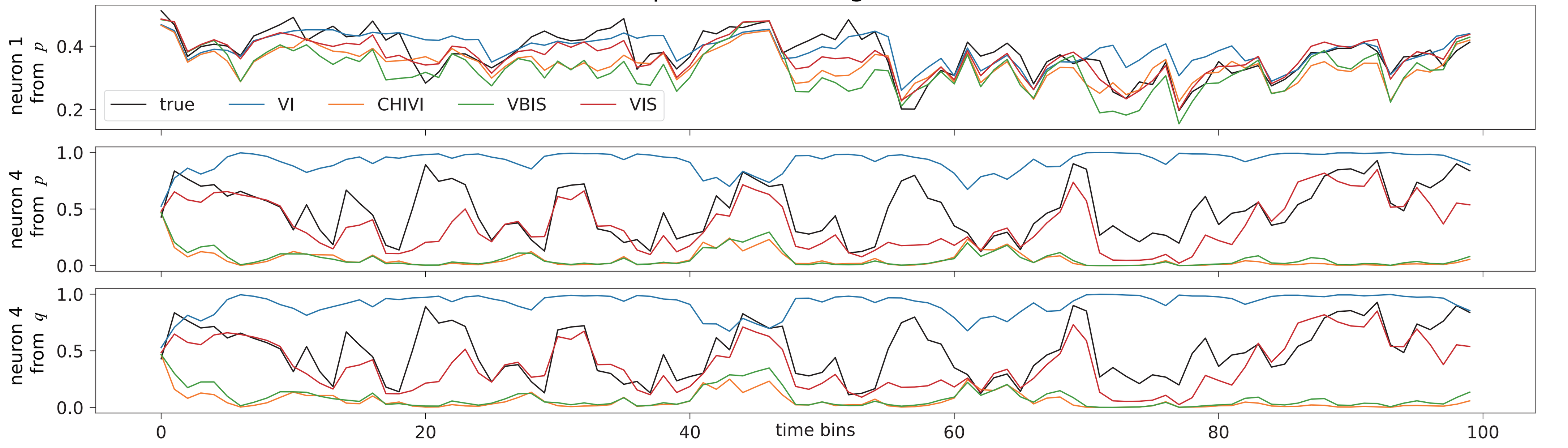
$q(\boldsymbol{z}|\boldsymbol{x};\phi) = \mathcal{N}\left(\boldsymbol{x};\boldsymbol{\mu}(\boldsymbol{x}),\mathrm{diag}\,\boldsymbol{\sigma}^2(\boldsymbol{x})\right)$, $\boldsymbol{\mu}(\boldsymbol{x}),\boldsymbol{\sigma}^2(\boldsymbol{x}) = \mathrm{decoder}(\boldsymbol{x})$



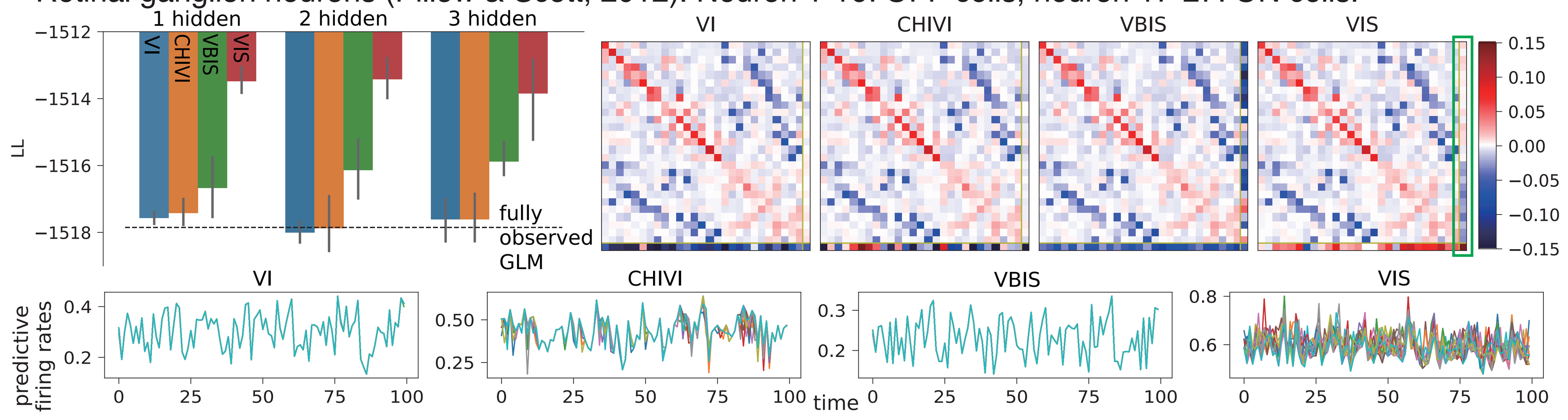
Partially observable GLM

- A very hard problem since $p(\boldsymbol{x};\boldsymbol{z};\theta)$ cannot be explicitly factored as $p(\boldsymbol{x}|\boldsymbol{z};\theta)p(\boldsymbol{z};\theta)$.
- $\boldsymbol{X}$ are spike trains from V visible neurons and $\boldsymbol{Z}$ are spike trains from $H = N - V$ hidden neurons.
- $\theta = \{\boldsymbol{b}\in\mathbb{R}^N, \boldsymbol{W}\in\mathbb{R}^{N\times N}\}$. $w_{n\leftarrow n'}$ is the weight from neuron n to neuron n'.





- Retinal ganglion neurons (Pillow & Scott, 2012). Neuron 1-16: OFF cells, neuron 17-27: ON cells.



References: [1] Burda et al., arXiv, 2015. [2] Dieng et al., NeurIPS, 2017. [3] Finke & Thiery, arXiv, 2019. [4] Jerfel et al., PMLR, 2018. [5] Domke & Sheldon, NeurIPS, 2021. [6] Su & Chen, Comp. Stat., 2021. [7] Akyildiz & Miguez, Stat. Comp., 2021.