Forward χ^2 Divergence Based Variational Importance Sampling

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

ICLR 2024 [Spotlight]

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

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

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

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

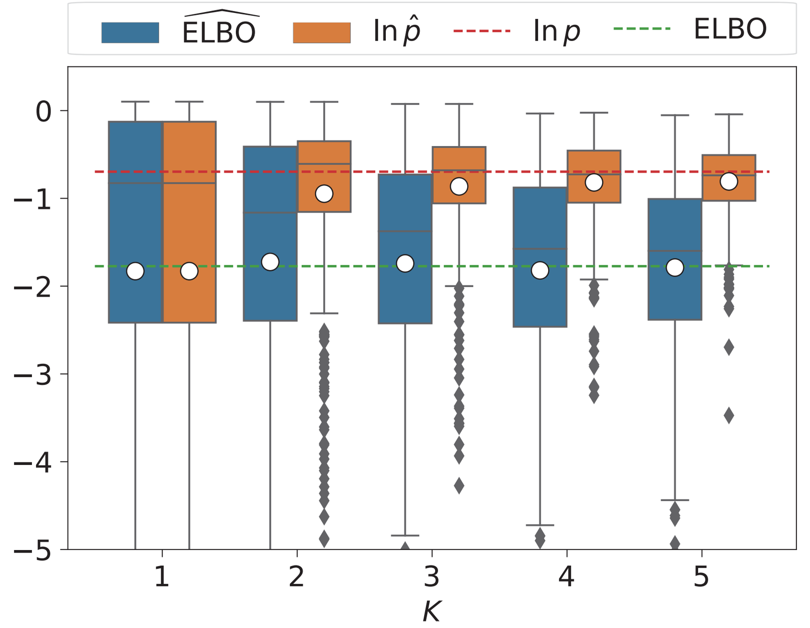
$\chi^2(p(\boldsymbol{z}|\boldsymbol{x};\theta)\|q(\boldsymbol{z}|\boldsymbol{x};\phi))/(2K)$

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

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

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

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



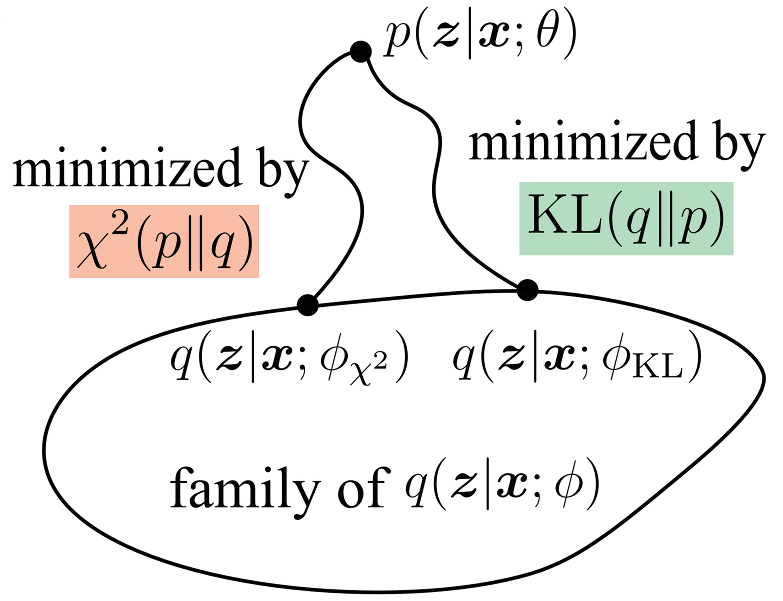
VIS as the best way of doing IS

In fact, the effectiveness of the IS estimator is

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

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

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



VIS algorithm

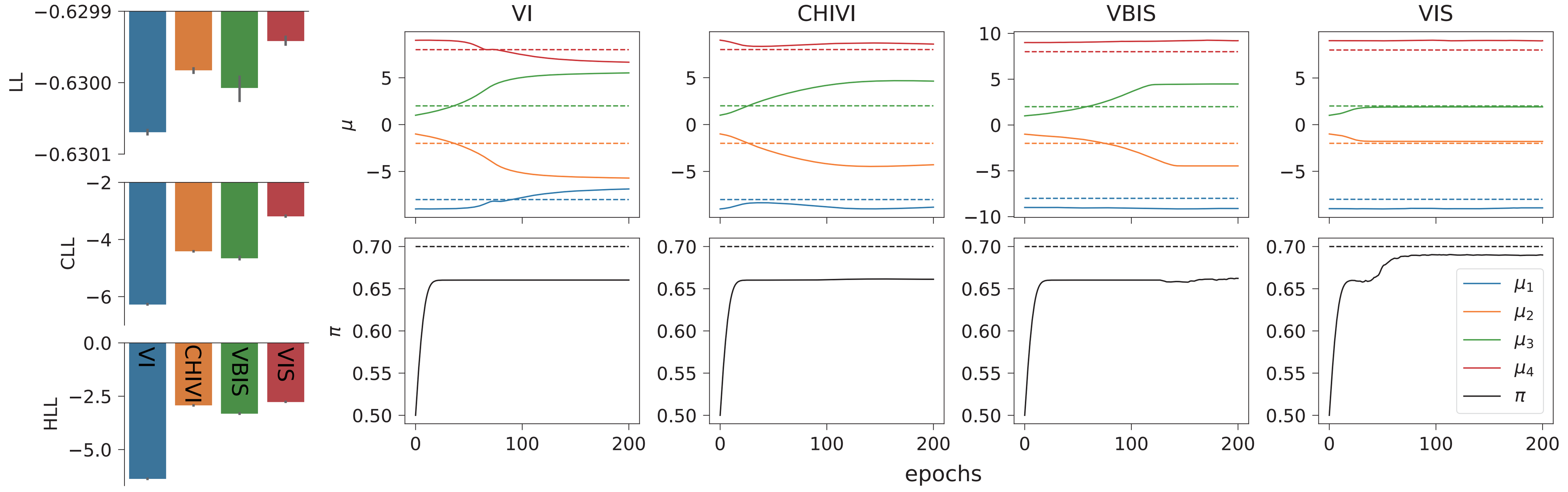
	VI	variational importance sampling (VIS)
Sample	$\{\boldsymbol{z}^{(k)}\}_{k=1}^{K} \sim q(\boldsymbol{z}\vert\boldsymbol{x};\phi)$	
E-step	$\min_\phi \mathrm{KL}(q(\boldsymbol{z}\vert\boldsymbol{x};\phi)\Vert p(\boldsymbol{z}\vert\boldsymbol{x};\theta)) \to \min_\phi \mathrm{ELBO}(\boldsymbol{x};\theta,\phi)$; $\frac{\partial\,\mathrm{ELBO}(\boldsymbol{x};\theta,\phi)}{\partial\phi} \approx \frac{\partial}{\partial\phi}\frac{-1}{2K}\sum_{k=1}^{K}\left[\ln p\left(\boldsymbol{x},\boldsymbol{z}^{(k)};\theta\right) - \ln q\left(\boldsymbol{z}^{(k)}\vert\boldsymbol{x};\phi\right)\right]^2$	$\min_\phi \chi^2(p(\boldsymbol{z}\vert\boldsymbol{x};\theta)\Vert q(\boldsymbol{z}\vert\boldsymbol{x};\phi)) \to \min_\phi \ln V(\boldsymbol{x};\theta,\phi)$; $\frac{\partial \ln V(\boldsymbol{x};\theta,\phi)}{\partial\phi} \approx \frac{\partial}{\partial\phi}\frac{1}{2}\ln\hat{V}(\boldsymbol{x};\theta,\phi)$
M-step	$\max_\theta \ln p(\boldsymbol{x};\theta) \to \max_\theta \mathrm{ELBO}(\boldsymbol{x};\theta,\phi)$; $\frac{\partial\,\mathrm{ELBO}(\boldsymbol{x};\theta,\phi)}{\partial\theta} \approx \frac{\partial}{\partial\theta}\widehat{\mathrm{ELBO}}(\boldsymbol{x};\theta,\phi)$	$\max_\theta \ln p(\boldsymbol{x};\theta)$; $\frac{\partial \ln p(\boldsymbol{x};\theta)}{\partial\theta} \approx \frac{\partial}{\partial\theta}\ln\hat{p}(\boldsymbol{x};\theta,\phi)$

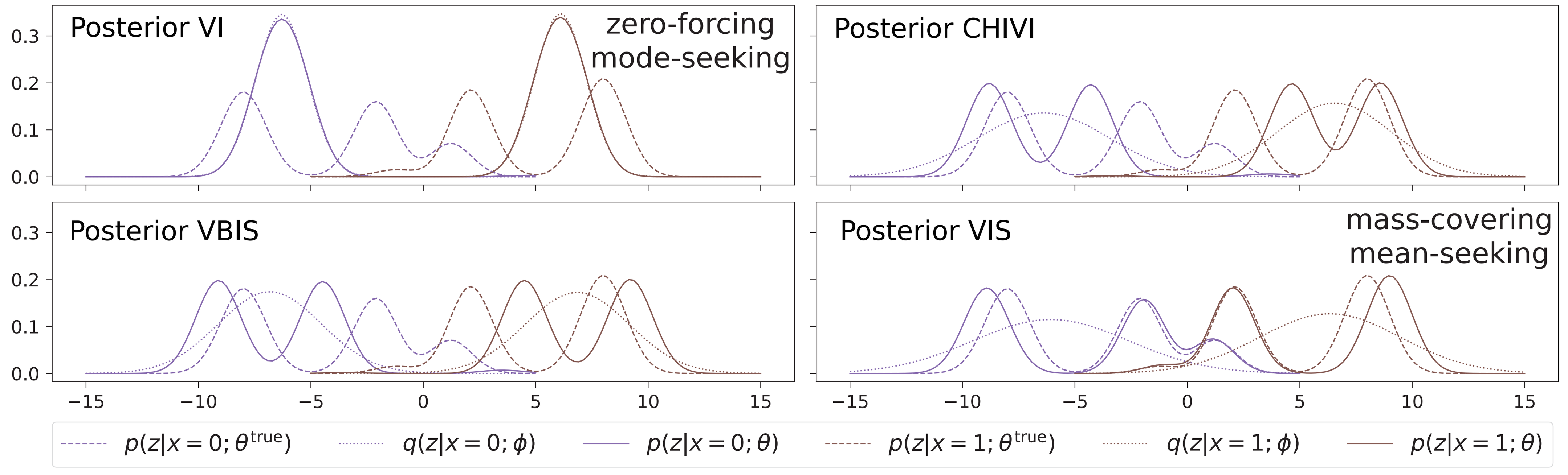
Experiments

Mixture model: GMM-Bernoulli

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

$$q(z|x;\phi) = \mathcal{N}(z;c_x,\sigma_x^2)$$

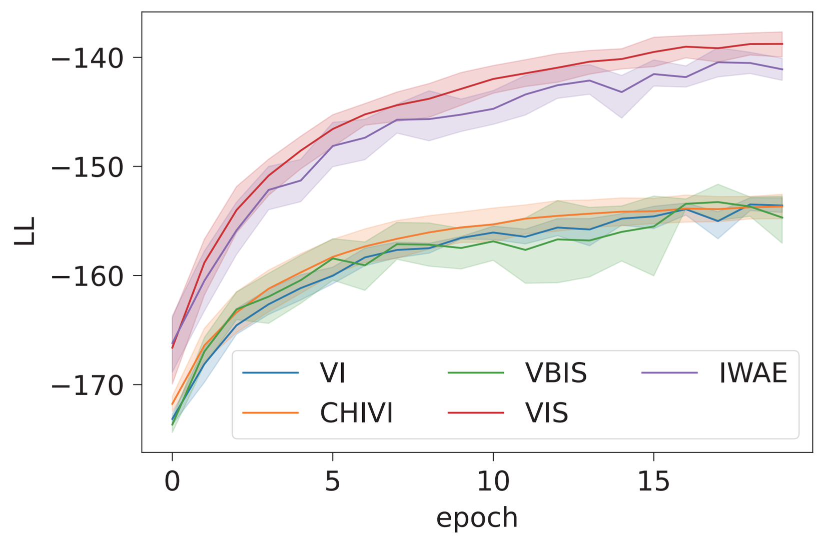


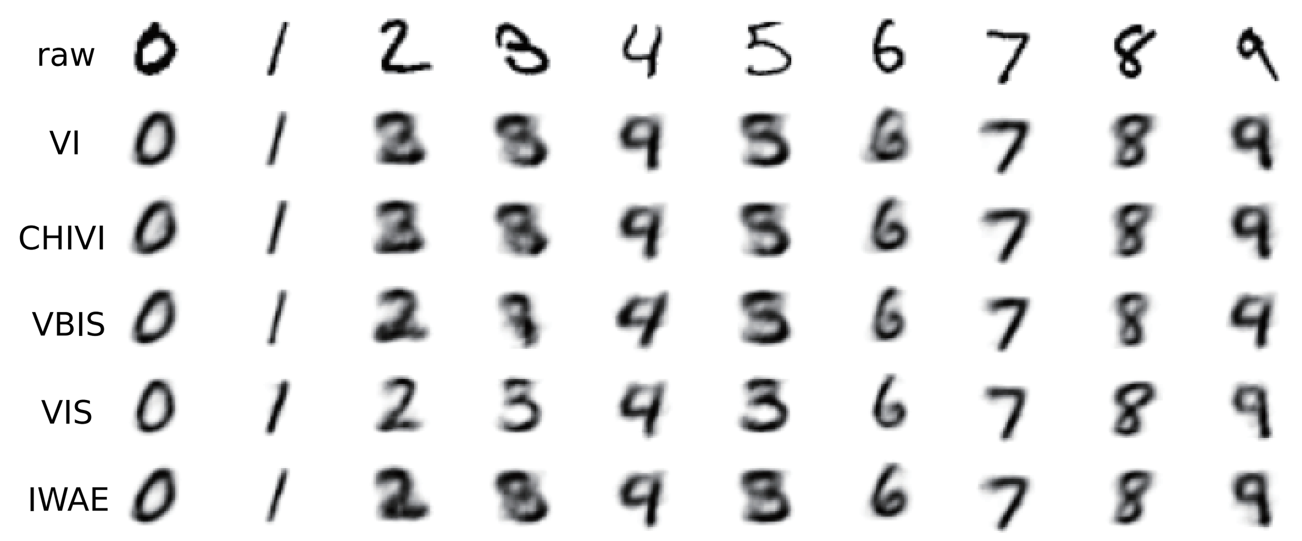


VAE on MNIST

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

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

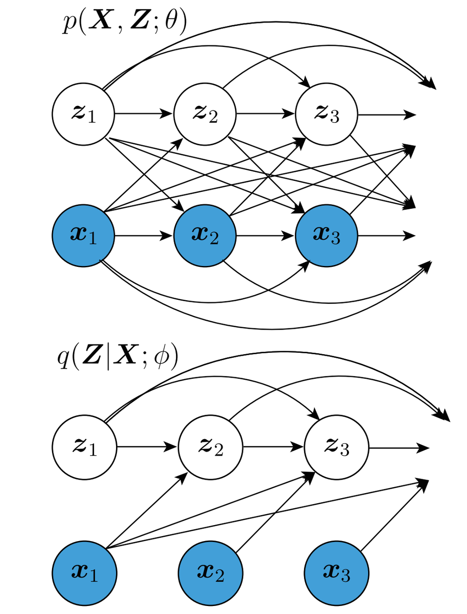




Partially observable GLM

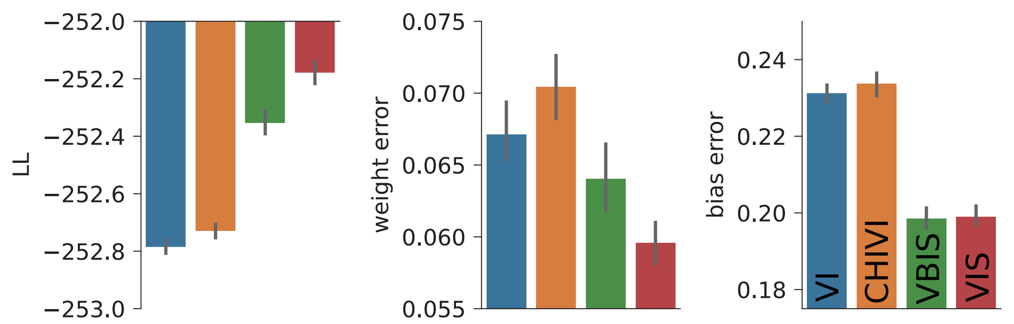
A very hard problem since $p(\boldsymbol{x},\boldsymbol{z};\theta)$ cannot be explicitly factored as $p(\boldsymbol{x}|\boldsymbol{z};\theta)p(\boldsymbol{z};\theta)$.

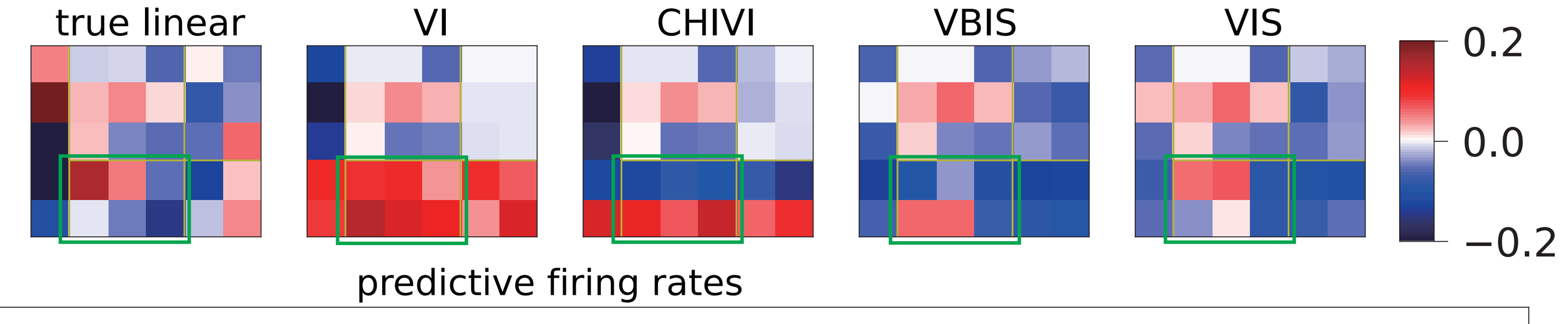
3-Visible-2-hidden synthetic

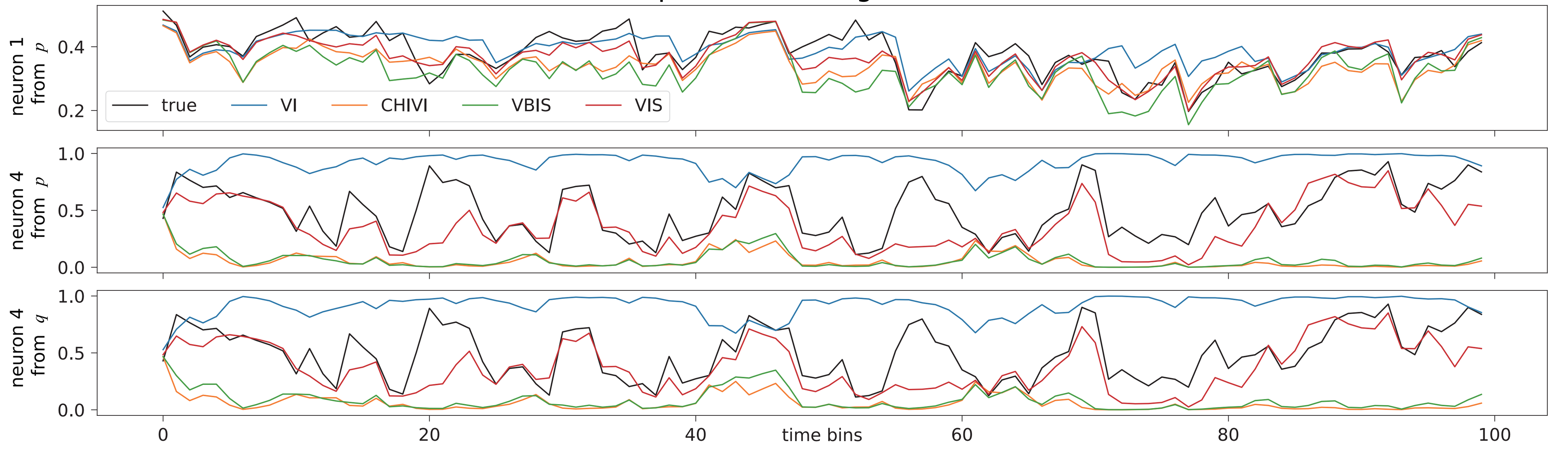


$$\begin{bmatrix}\boldsymbol{x}_{t+1}\\ \boldsymbol{z}_{t+1}\end{bmatrix} = \mathrm{Poisson}\left(\boldsymbol{W}\begin{bmatrix}\boldsymbol{x}_t\\ \boldsymbol{z}_t\end{bmatrix} + \boldsymbol{b}\right)$$

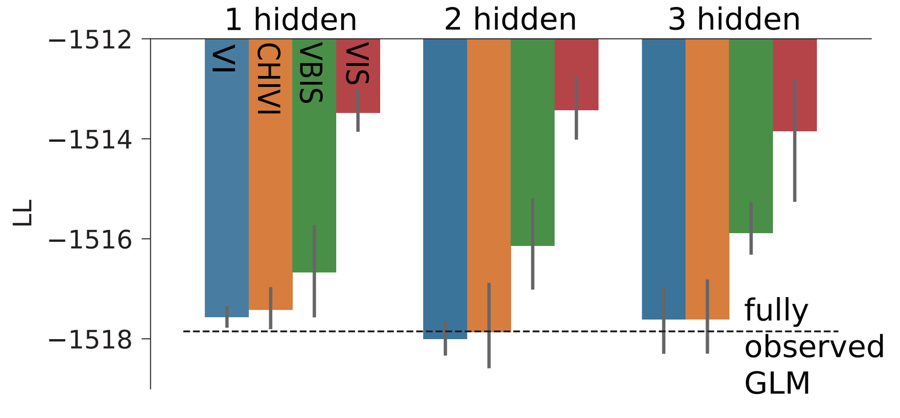
$$\theta = \{\boldsymbol{b}\in\mathbb{R}^N, \boldsymbol{W}\in\mathbb{R}^{N\times N}\}$$

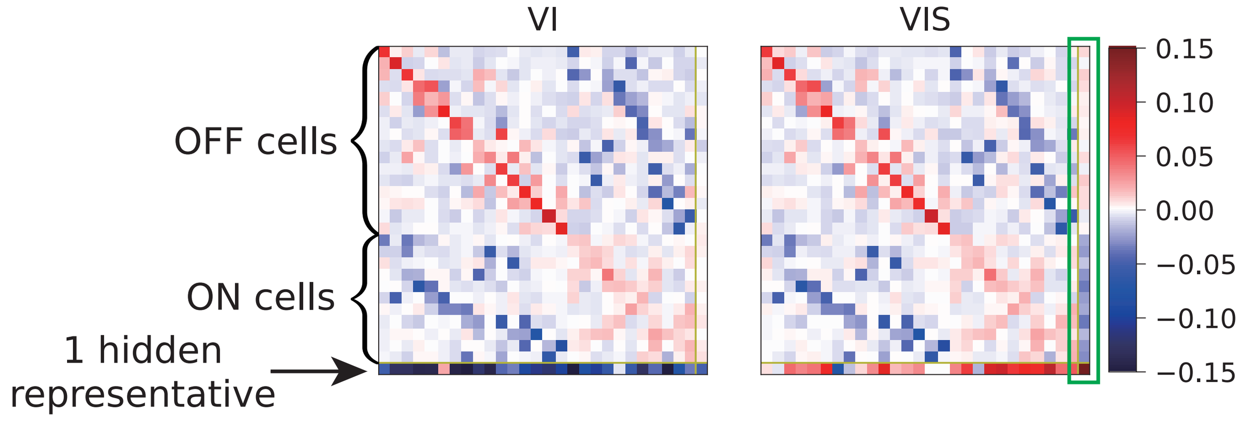






Retinal ganglion neurons (Pillow & Scott, 2012)





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.