Forward χ^2 divergence based variational importance sampling

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

Contents

1 Recap: variational inference

Latent variable model (LVM)

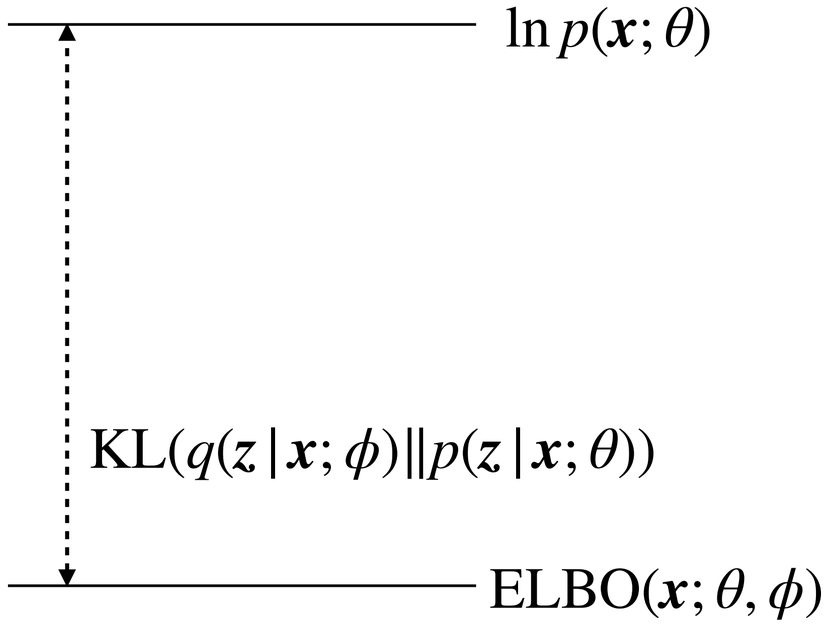
- Notation:
  - $\boldsymbol{x}$: observed variables
  - $\boldsymbol{z}$: latent variables
  - θ: parameter set
- Complete data likelihood: $p(\boldsymbol{x}, \boldsymbol{z}; \theta)$
- Marginal likelihood: $p(\boldsymbol{x}; \theta) = \int p(\boldsymbol{x}, \boldsymbol{z}; \theta) \, \mathrm{d}\boldsymbol{z}$, which is intractable when the problem is complicated.
- Maximum likelihood estimation: $\hat{\theta} = \mathrm{argmax}_{\theta} \, p(\boldsymbol{x}; \theta)$

Variational inference

- Use a variational distribution $q(\boldsymbol{z}\,|\,\boldsymbol{x};\boldsymbol{\phi})$ to approximate $p(\boldsymbol{z}\,|\,\boldsymbol{x};\boldsymbol{\theta})$.
- Consider the reverse KL divergence:

$$\begin{aligned}\mathrm{KL}(q(\boldsymbol{z}\,|\,\boldsymbol{x};\boldsymbol{\phi})\|p(\boldsymbol{z}\,|\,\boldsymbol{x};\boldsymbol{\theta})) &= \int q(\boldsymbol{z}\,|\,\boldsymbol{x};\boldsymbol{\phi})\ \ln\frac{q(\boldsymbol{z}\,|\,\boldsymbol{x};\boldsymbol{\phi})}{p(\boldsymbol{z}\,|\,\boldsymbol{x};\boldsymbol{\theta})}\,\mathrm{d}\boldsymbol{z}\\ &= \ln p(\boldsymbol{x};\boldsymbol{\theta}) - \mathbb{E}_q[\ln p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}) - \ln q(\boldsymbol{z}\,|\,\boldsymbol{x};\boldsymbol{\phi})]\\ &= \ln p(\boldsymbol{x};\boldsymbol{\theta}) - \mathrm{ELBO}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\phi})\end{aligned}$$

- So, ELBO is a lower bound of $\ln p(\boldsymbol{x};\boldsymbol{\theta})$.
- Maximize ELBO w.r.t. $\boldsymbol{\theta}$ increases $\ln p(x;\boldsymbol{\theta})$, and maximize ELBO w.r.t. $\boldsymbol{\phi}$ reduces the reverse KL divergence $\mathrm{KL}(q(\boldsymbol{z}\,|\,\boldsymbol{x};\boldsymbol{\phi})\|p(\boldsymbol{z}\,|\,\boldsymbol{x};\boldsymbol{\theta}))$.



Numerical estimator of ELBO

- Numerical estimator of ELBO:

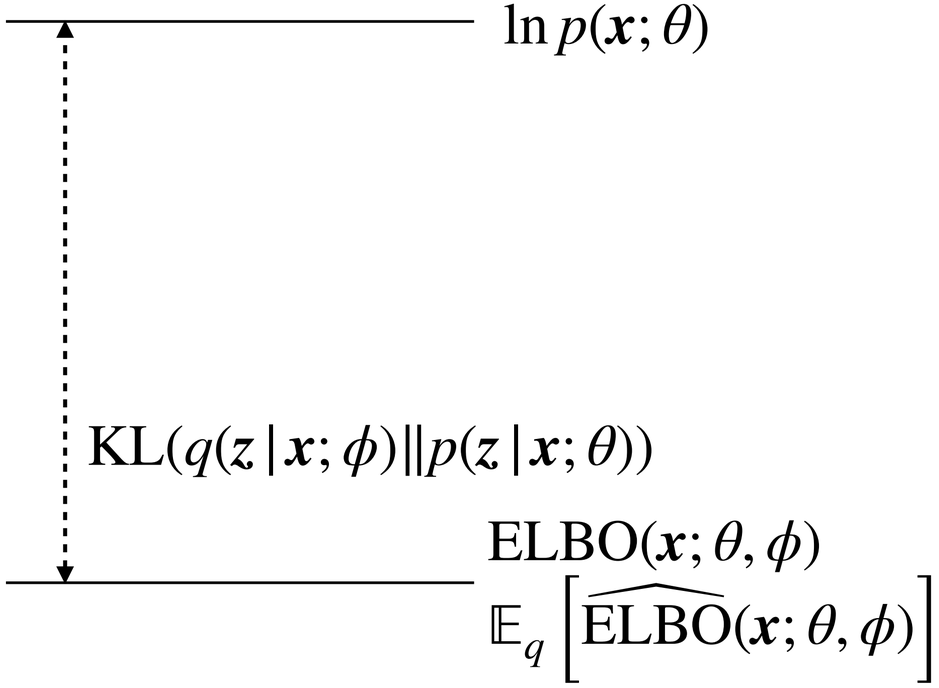
$$\widehat{\mathrm{ELBO}}(\boldsymbol{x};\theta,\phi) = \frac{1}{K}\sum_{k=1}^{K}\left[\ln p\left(\boldsymbol{x},\boldsymbol{z}^{(k)};\theta\right) - \ln q\left(\boldsymbol{z}^{(k)}\middle|\boldsymbol{x};\phi\right)\right]$$

  where $\left\{\boldsymbol{z}^{(k)}\right\}_{k=1}^{K} \sim q(\boldsymbol{z}\,|\,\boldsymbol{x};\phi).$

- The bias of the ELBO estimator:

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

- Is there any other estimator that approximates $\ln p(\boldsymbol{x};\theta)$ better?



2 Importance sampling

Directly approximate the marginal

- Importance sampling uses a proposal distribution $q(z \mid x; \phi)$ to estimate the integration:

$$p(\boldsymbol{x}; \theta) = \int p(\boldsymbol{x}, \boldsymbol{z}; \theta) \, \mathrm{d}\boldsymbol{z} = \mathbb{E}_q[p(\boldsymbol{x}, \boldsymbol{z}; \theta)/q(\boldsymbol{z} \mid \boldsymbol{x}; \boldsymbol{\phi})] \approx \frac{1}{K} \sum_{k=1}^{K} \frac{p\left(\boldsymbol{x}, \boldsymbol{z}^{(k)}; \theta\right)}{q\left(\boldsymbol{z}^{(k)} \middle| \boldsymbol{x}; \boldsymbol{\phi}\right)} =: \hat{p}(\boldsymbol{x}; \theta, \boldsymbol{\phi})$$

  where $\left\{\boldsymbol{z}^{(k)}\right\}_{k=1}^{K} \sim q(\boldsymbol{z} \mid \boldsymbol{x}; \boldsymbol{\phi})$.

- Numerical stable IS estimator in log space

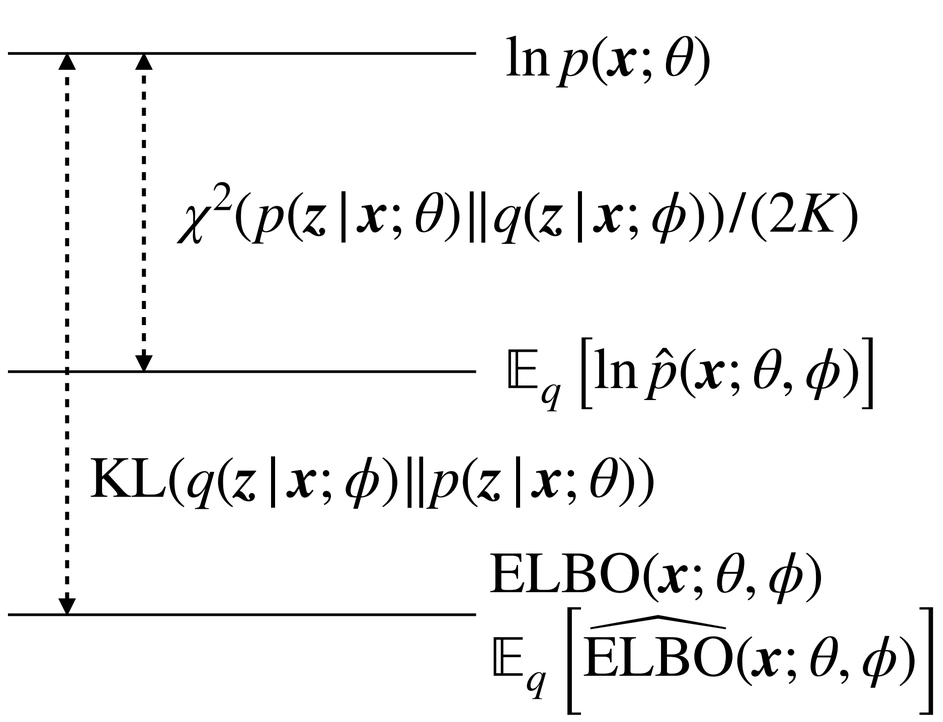
$$\ln \hat{p}(\boldsymbol{x}; \theta, \boldsymbol{\phi}) = \operatorname{logsumexp}\left[\ln p\left(\boldsymbol{x}, \boldsymbol{z}^{(k)}; \theta\right) - \ln q\left(\boldsymbol{z}^{(k)} \middle| \boldsymbol{x}; \boldsymbol{\phi}\right)\right] - \ln K$$

- And its gradient estimator

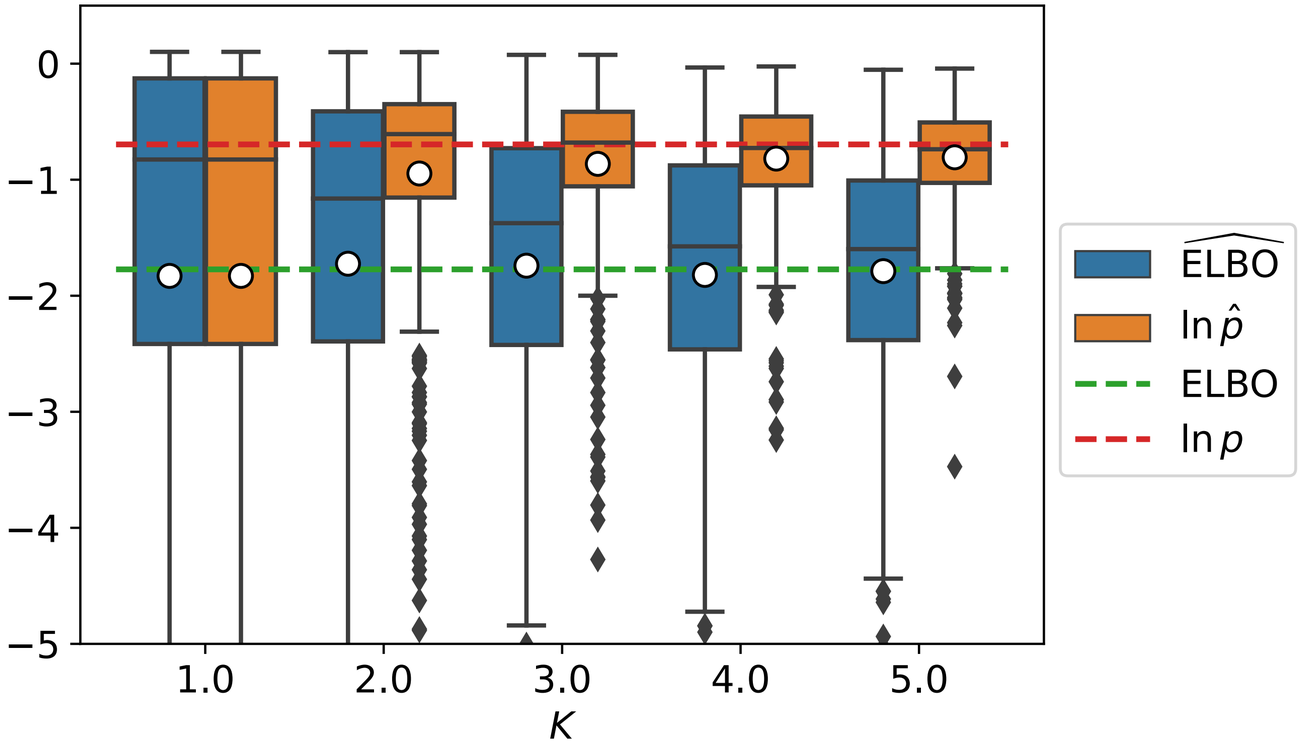
$$\frac{\partial \ln p(\boldsymbol{x}; \theta)}{\partial \theta} \approx \frac{\partial}{\partial \theta} \ln \hat{p}(\boldsymbol{x}; \theta, \boldsymbol{\phi})$$

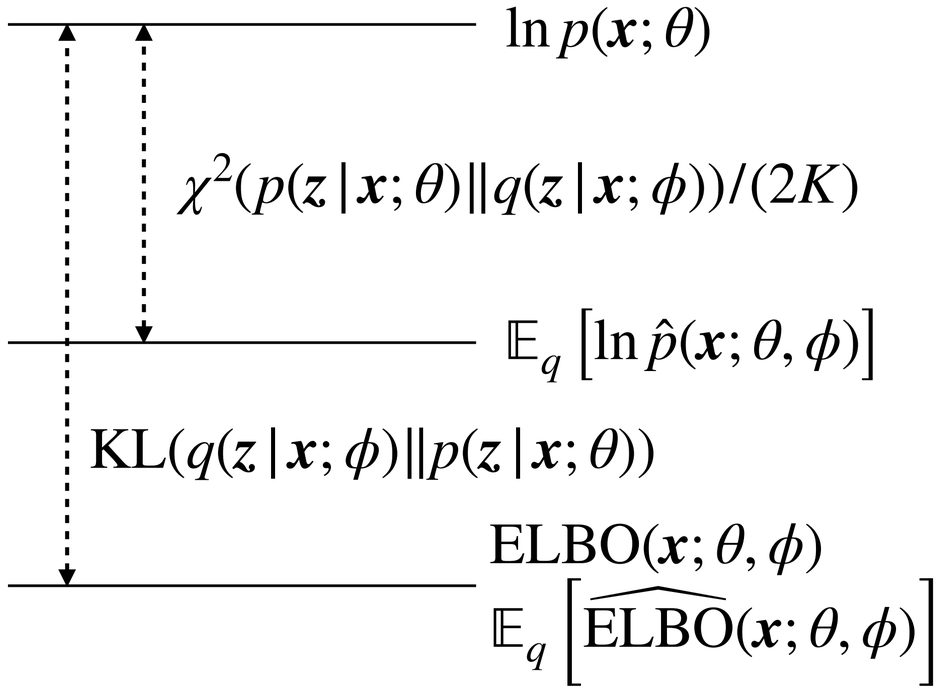
Compare VI and IS

- The bias of the IS estimator:
  $$\mathbb{E}_q\left[\ln \hat{p}(\boldsymbol{x};\theta,\phi) - \ln p(\boldsymbol{x};\theta)\right] \approx -\frac{1}{2K}\chi^2(p(\boldsymbol{z}\,|\,\boldsymbol{x};\theta)\|q(\boldsymbol{z}\,|\,\boldsymbol{x};\phi))$$
- The bias of the IS estimator $\to 0$ as $K \to \infty$.
- Particularly, when $K = 1$,
  $\widehat{\mathrm{ELBO}}(\boldsymbol{x};\theta,\phi) = \ln \hat{p}(\boldsymbol{x};\theta,\phi)$.
- So, compared with $\widehat{\mathrm{ELBO}}(\boldsymbol{x};\theta,\phi)$, $\ln \hat{p}(\boldsymbol{x};\theta,\phi)$ is an asymptotically tighter lower bound.



Validate this by numerical simulations





Since IS estimates $\ln p(\boldsymbol{x};\theta)$ better than VI with a large K, we can use IS to learn θ.

3 VIS as the best way of doing IS

Choosing the optimal proposal distribution

- Remember the bias of the IS estimator is $-\chi^2(p(\boldsymbol{z}\,|\,\boldsymbol{x};\theta)\|q(\boldsymbol{z}\,|\,\boldsymbol{x};\phi))/(2K)$.
- The smaller the absolute bias, the better the estimator.
- In fact, the effectiveness of of the estimator $\mathrm{Var}_q\left[\hat{p}(\boldsymbol{x};\theta,\phi)\right] = \dfrac{p(\boldsymbol{x};\theta)^2}{K}\chi^2(p(\boldsymbol{z}\,|\,\boldsymbol{x};\theta)\|q(\boldsymbol{z}\,|\,\boldsymbol{x};\phi))$
- 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 $p(\boldsymbol{z}\,|\,\boldsymbol{x};\theta)$.
- Similar to the case we encountered in the reverse KL divergence, we don't know $p(\boldsymbol{z}\,|\,\boldsymbol{x};\theta)$, so we derive the following equation.
- $\chi^2(p(\boldsymbol{z}\,|\,\boldsymbol{x};\theta)\|q(\boldsymbol{z}\,|\,\boldsymbol{x};\phi)) = \dfrac{1}{p(\boldsymbol{x};\theta)^2}\displaystyle\int \frac{p(\boldsymbol{x},\boldsymbol{z};\theta)^2}{q(\boldsymbol{z}\,|\,\boldsymbol{x};\phi)}\,\mathrm{d}\boldsymbol{z} - 1 =: \frac{1}{p(\boldsymbol{x};\theta)^2}V(\boldsymbol{x};\theta,\phi) - 1$
- We convert minimizing $\chi^2(p(\boldsymbol{z}\,|\,\boldsymbol{x};\theta)\|q(\boldsymbol{z}\,|\,\boldsymbol{x};\phi))$ to minimizing $V(\boldsymbol{x};\theta,\phi)$.

Estimate and minimize $V(\boldsymbol{x};\theta,\phi)$ in log space

- For numerical stability, $V(\boldsymbol{x};\theta,\phi)$ should be minimized in log space.
- $$\ln V(\boldsymbol{x};\theta,\phi) \approx \operatorname{logsumexp}\left[2\ln p\left(\boldsymbol{x},\boldsymbol{z}^{(k)};\theta\right) - 2\ln q\left(\boldsymbol{z}^{(k)}\middle|\boldsymbol{x};\phi\right)\right] - \ln K = \ln \hat{V}(\boldsymbol{x};\theta,\phi)$$
- The score function gradient estimator of $\ln V(\boldsymbol{x};\theta,\phi)$ is

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

Comparison of VI and VIS

Only changes two lines in the code

	VI	VIS
Sample	Sample $\left\{\boldsymbol{z}^{(k)}\right\}_{k=1}^{K}$ from $q(\boldsymbol{z}\,\vert\,\boldsymbol{x};\boldsymbol{\phi})$	
E-step	Minimize $\mathrm{KL}(q(\boldsymbol{z}\,\vert\,\boldsymbol{x};\boldsymbol{\phi})\Vert p(\boldsymbol{z}\,\vert\,\boldsymbol{x};\boldsymbol{\theta}))$ w.r.t. $\boldsymbol{\phi}$ by maximizing $\widehat{\mathrm{ELBO}}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\phi})$ w.r.t. $\boldsymbol{\phi}$	Minimize $\chi^2(p(\boldsymbol{z}\,\vert\,\boldsymbol{x};\boldsymbol{\phi})\Vert q(\boldsymbol{z}\,\vert\,\boldsymbol{x};\boldsymbol{\theta}))$ w.r.t. $\boldsymbol{\phi}$ by minimizing $\ln \hat{V}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\phi})$ w.r.t. $\boldsymbol{\phi}$
M-step	Maximize $\ln p(\boldsymbol{x};\boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta}$ by maximizing $\widehat{\mathrm{ELBO}}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\phi})$ w.r.t. $\boldsymbol{\theta}$	Maximize $\ln p(\boldsymbol{x};\boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta}$ by maximizing $\ln \hat{p}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\phi})$ w.r.t. $\boldsymbol{\theta}$

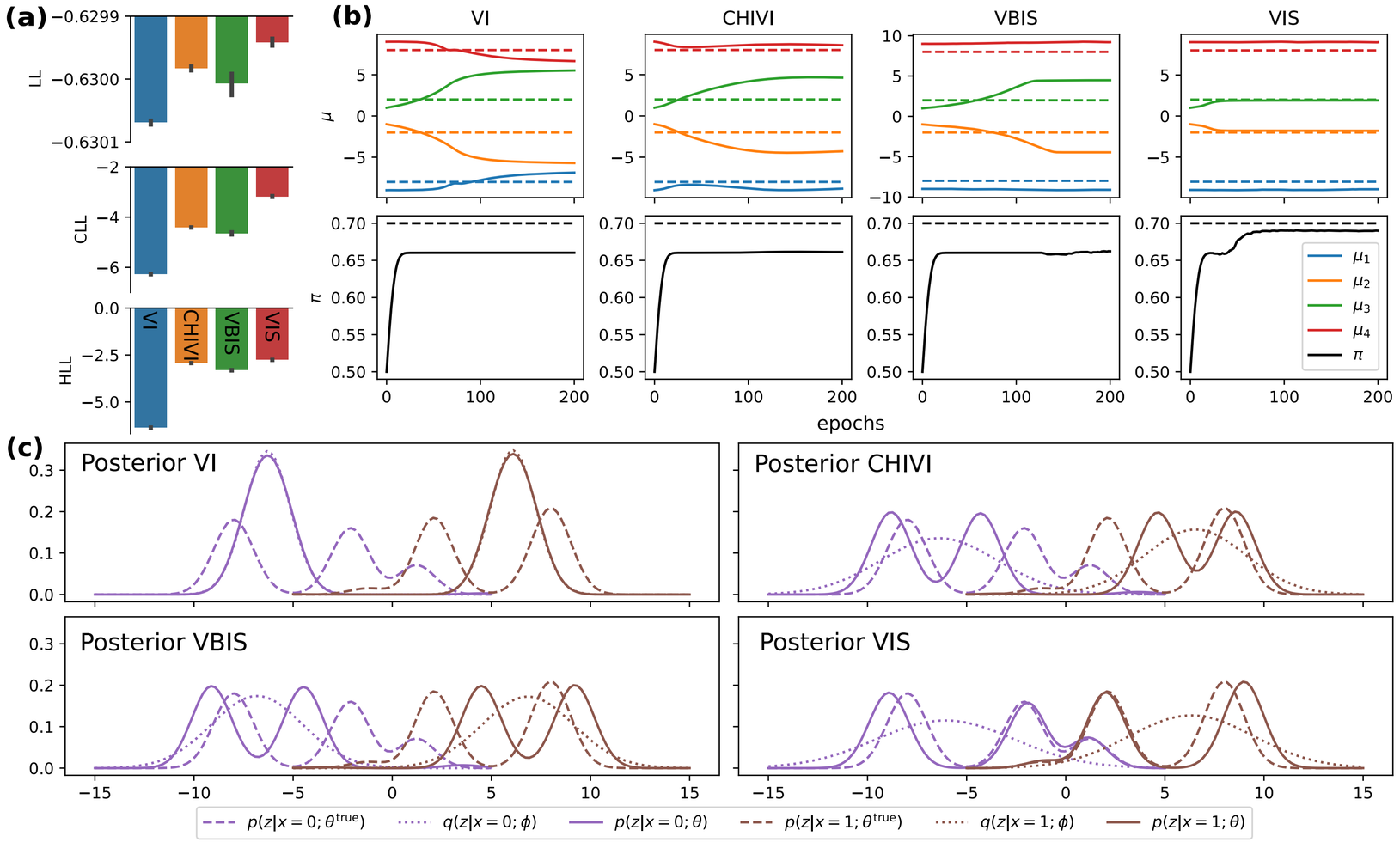
4 Applications to three latent variable models

Mixture model: GMM-Bernoulli

- Model
  - $p(z;\theta) = \sum_{i=1}^{4} \mathcal{N}\left(z;\mu_i,1^2\right)$ with $\pi_1 = \pi_2 = \frac{1-\pi}{2}$, $\pi_3 = \pi_4 = \frac{\pi}{2}$
  - $p(x \mid z;\theta) = \text{Bernoulli}(x;\text{sigmoid}(z))$
  - $\theta = \left\{\pi, \mu_1, \mu_2, \mu_3, \mu_4\right\}$
- Variational/proposal distribution family
  - $q(z \mid x;\phi) = \mathcal{N}\left(z; c_x, \sigma_x^2\right)$ for $x \in \{0,1\}$
  - $\phi = \left\{c_0, c_1, \sigma_0, \sigma_1\right\}$

Mixture model: GMM-Bernoulli

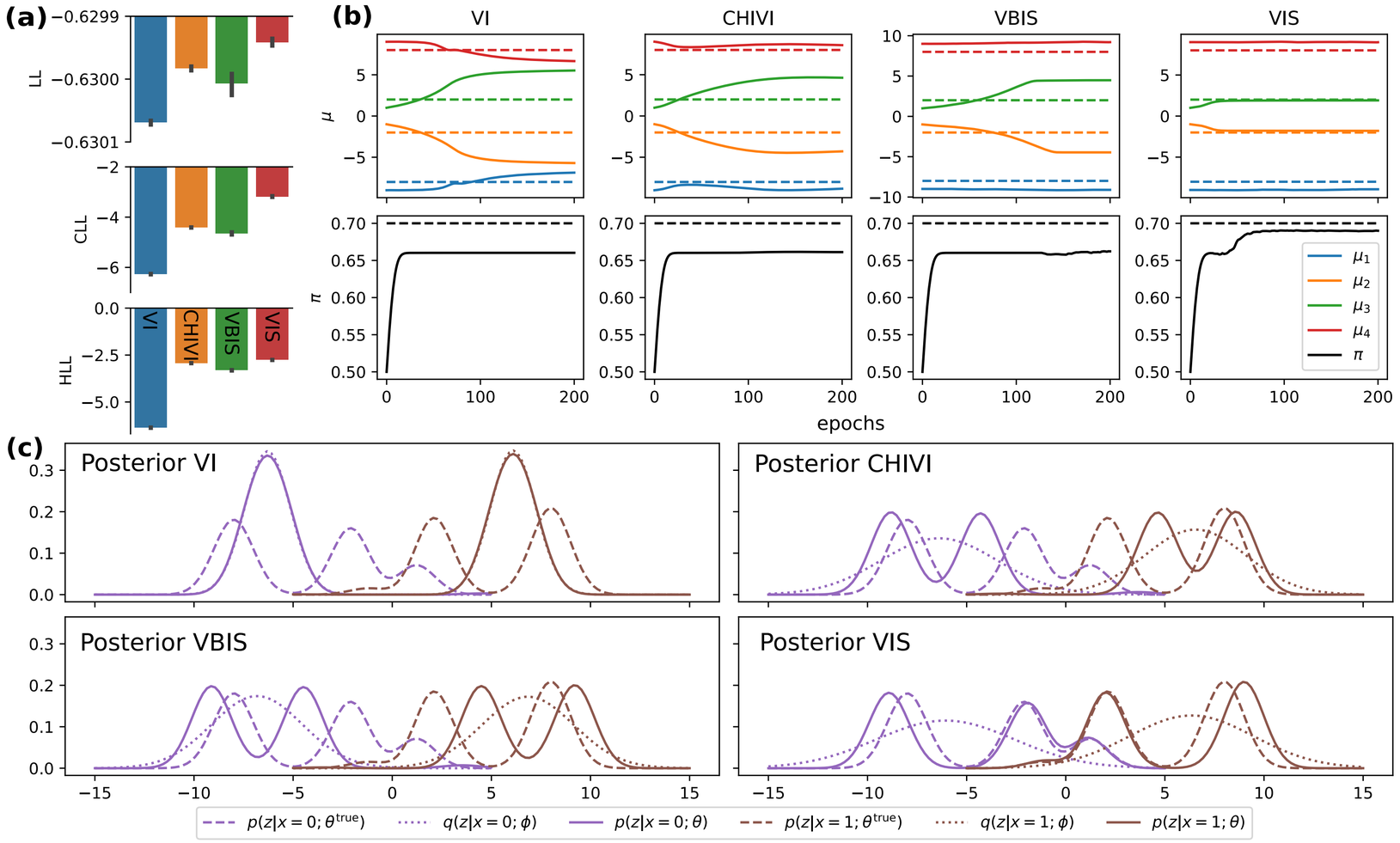
- (a) Quantitative comparison on the test dataset (e.g., test marginal log-likelihood).
- (b) parameter recovery.
- (c) Visualization of the true posterior (dashed), learned posterior (solid), and approximated posterior (dotted).



Mixture model: GMM-Bernoulli

- VI: zero-forcing/mode-seeking behavior of minimizing the reverse KL. Good ELBO, reverse KL is nearly 0, but in fact both $p(z \mid x; \theta)$ and $q(z \mid x; \phi)$ are far from $p(z \mid x; \theta^{\text{true}})$.

- VIS: mass-covering/ mean-seeking behavior of minimizing the forward χ^2. This enlarges the effective support range of $q(z \mid x; \phi)$ for sampling.

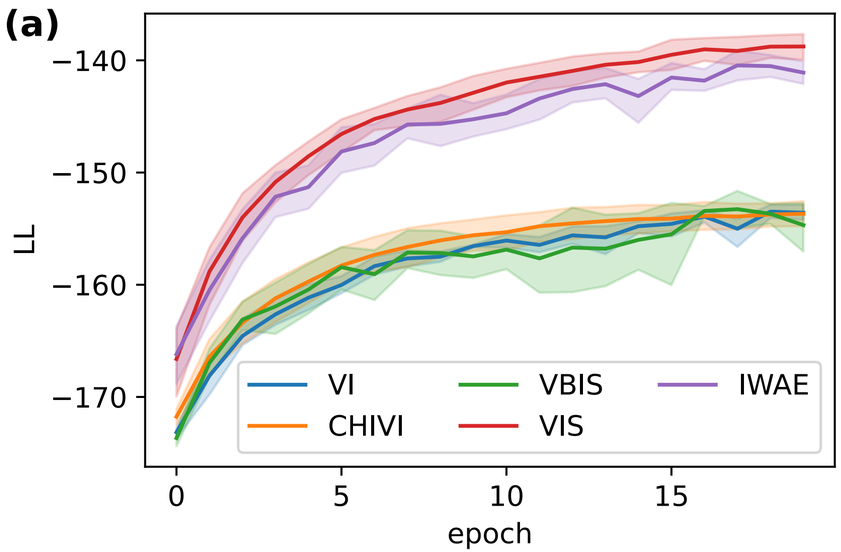


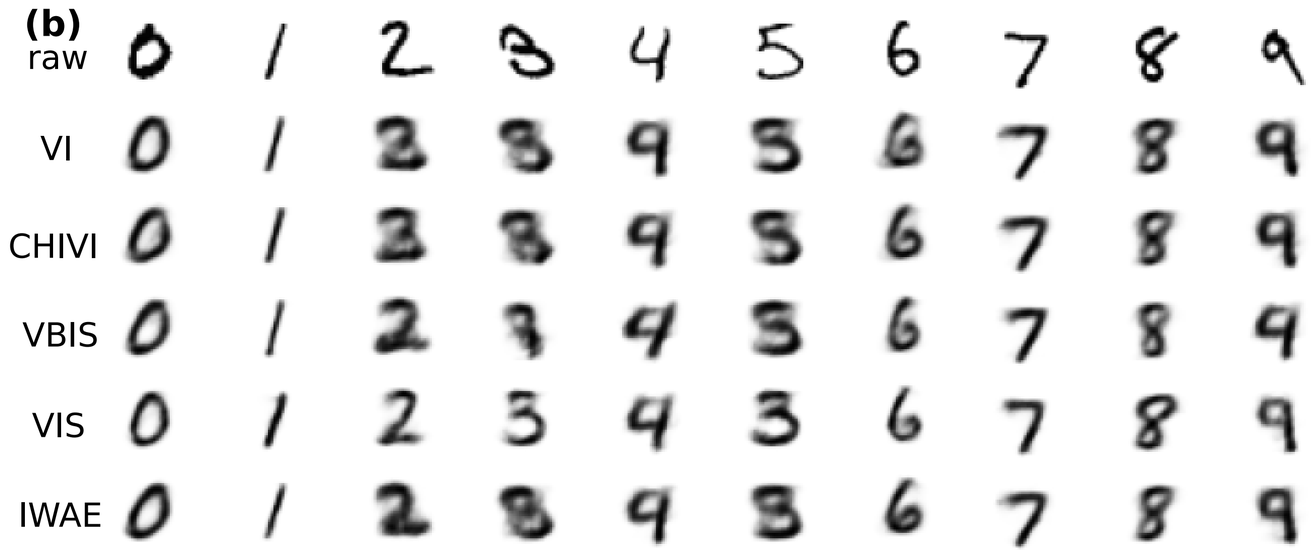
VAE on MNIST

- Model
  - $p(\boldsymbol{z}; \theta) = \mathcal{N}(\boldsymbol{z}; \mathbf{0}, \boldsymbol{I})$
  - $p(\boldsymbol{x} \mid \boldsymbol{z}; \theta) = \text{Bernoulli}(\boldsymbol{x}; \text{sigmoid}(\text{decoder}(\boldsymbol{z})))$
  - θ: decoder's parameters
- Variational/proposal distribution family
  - $q(\boldsymbol{z} \mid \boldsymbol{x}; \boldsymbol{\phi}) = \mathcal{N}\left(\boldsymbol{x}; \boldsymbol{\mu}(\boldsymbol{x}), \text{diag}\, \boldsymbol{\sigma}^2(\boldsymbol{x})\right)$, where $\boldsymbol{\mu}(\boldsymbol{x}), \boldsymbol{\sigma}^2(\boldsymbol{x}) = \text{encoder}(\boldsymbol{x})$
  - $\boldsymbol{\phi}$: encoder's parameters

VAE on MNIST

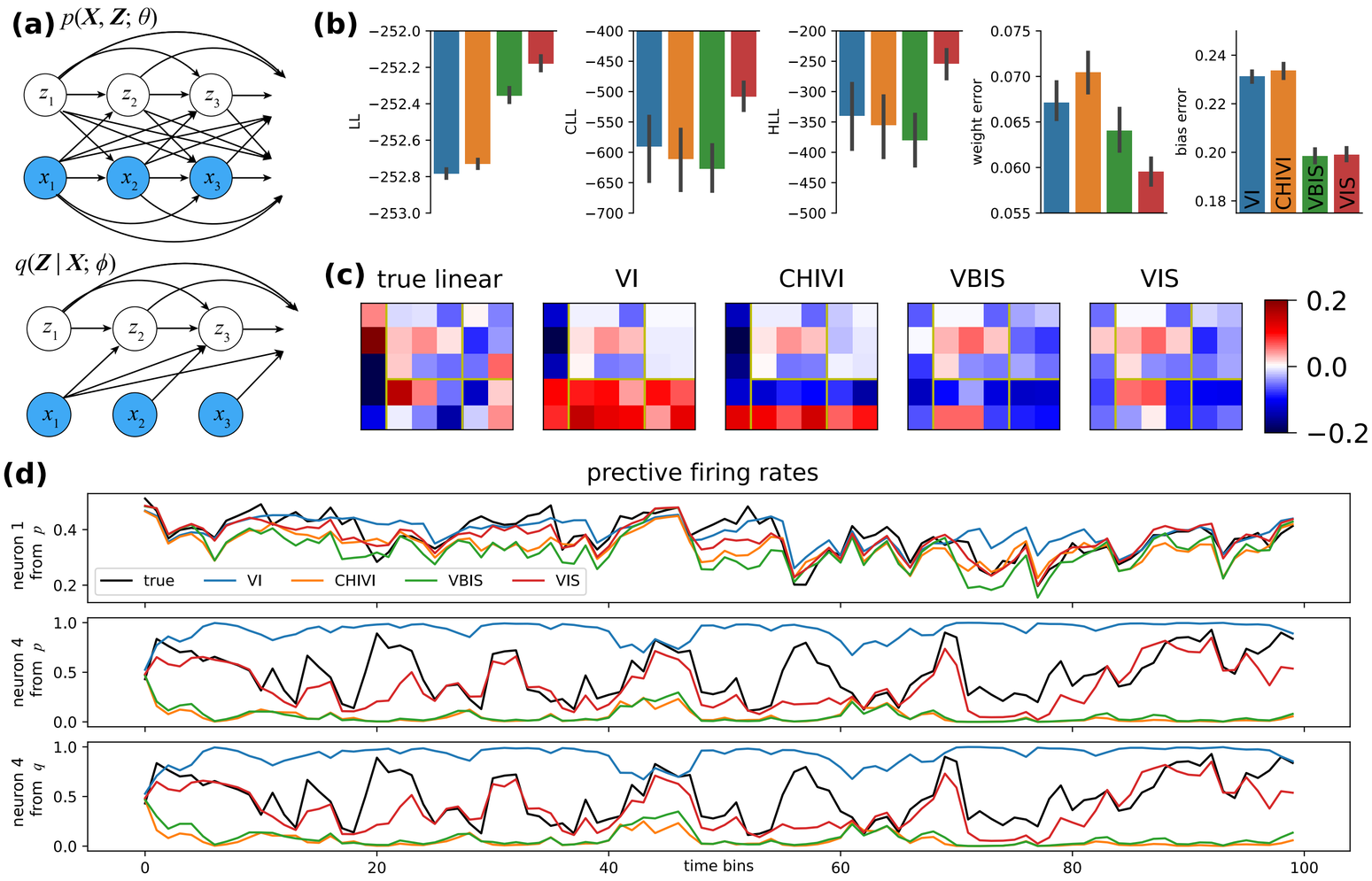
- (a) Convergence curve of the marginal log-likelihood on the test set.
- (b) Examples of raw images and the reconstructed images by different methods.





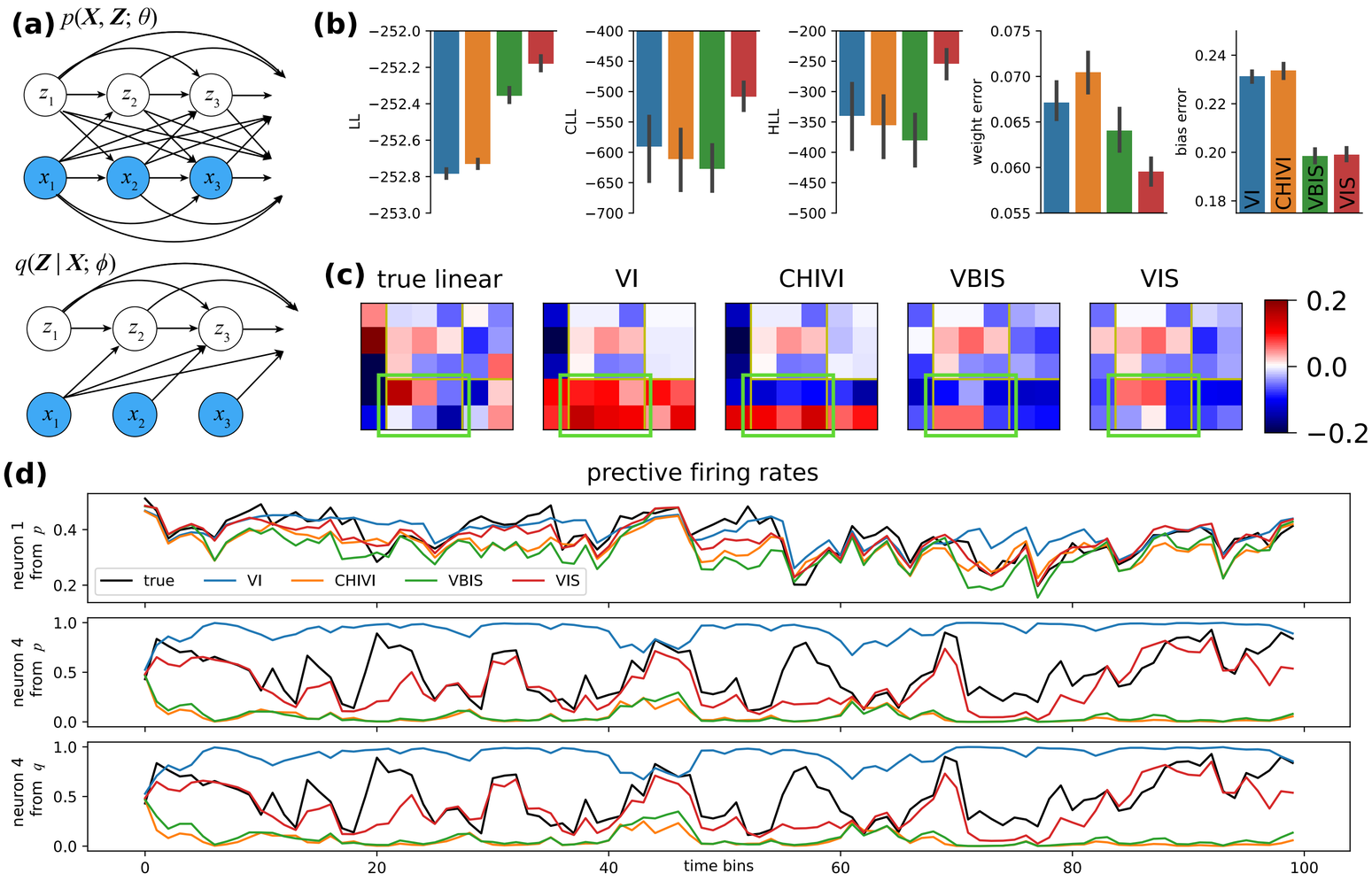
Partially observable GLM (synthetic)

- This is a very hard problem, since $p(\boldsymbol{x}, z; \theta)$ cannot be explicitly factored as $p(\boldsymbol{x} \mid z; \theta)p(z; \theta)$ (e.g., mixture model, HMM, PLDS, VAE, etc).
- V of N neurons are visible and the remaining H neurons are hidden.
- $\boldsymbol{X}$ are spike trains from visible neurons, and $\boldsymbol{Z}$ are spike trains form hidden neurons.
- $\theta = \left\{\boldsymbol{b} \in \mathbb{R}^N, \boldsymbol{W} \in \mathbb{R}^{N \times N}\right\}$, where $w_{n \leftarrow n'}$ represents the influence from neuron n' to neuron n.



Partially observable GLM (synthetic)

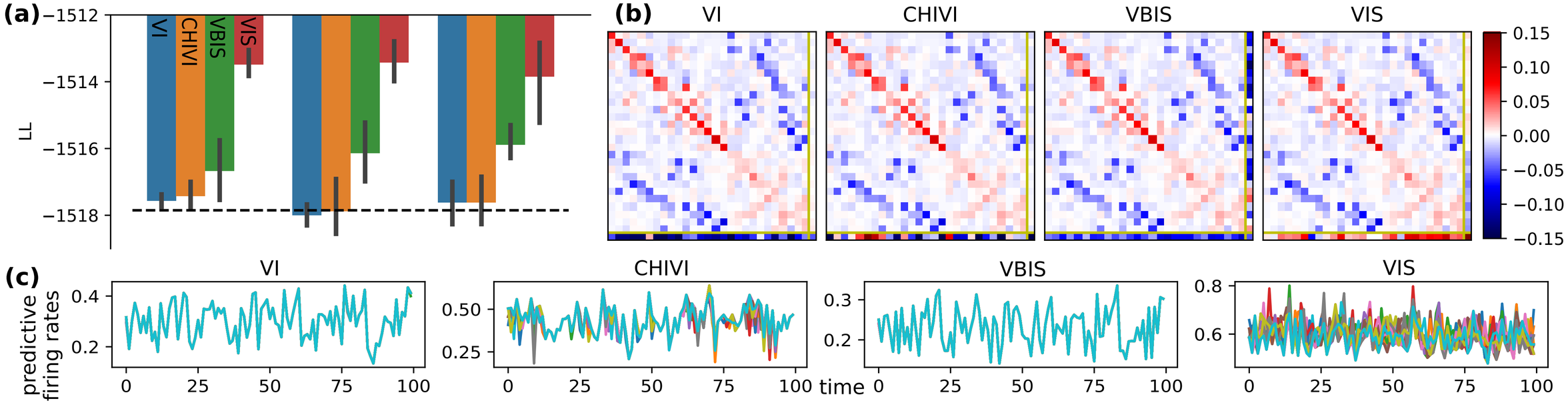
- (b) Quantitative comparison on the test dataset; weight and bias error.
- (c) The estimated weight $\boldsymbol{W}$ and the bias $\boldsymbol{b}$ by different methods compared with the true. The visible-to-hidden block learned by VIS is significantly better than others.
- (d) Predictive firing rates from different methods compared with the true. VIS matches the true the best.



Partially observable GLM (retinal ganglion cell)

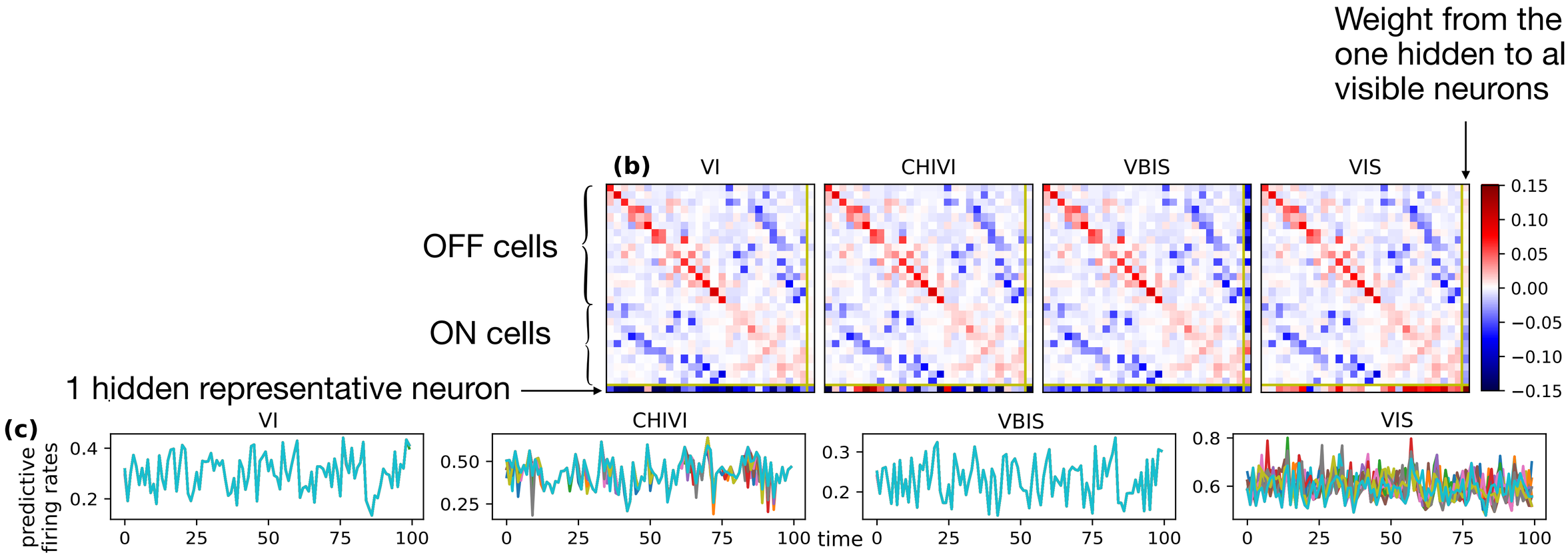
- $V = 27$ basal ganglion neurons are recorded while a mouse is performing a visual test for about 20 mins (Pillow & Scott, 2012). Neuron 1-16 are OFF cells, neuron 17-27 are ON cells.

- (a) Quantitative comparison on the test dataset, with different numbers of hidden neurons $H \in \{1,2,3\}$.



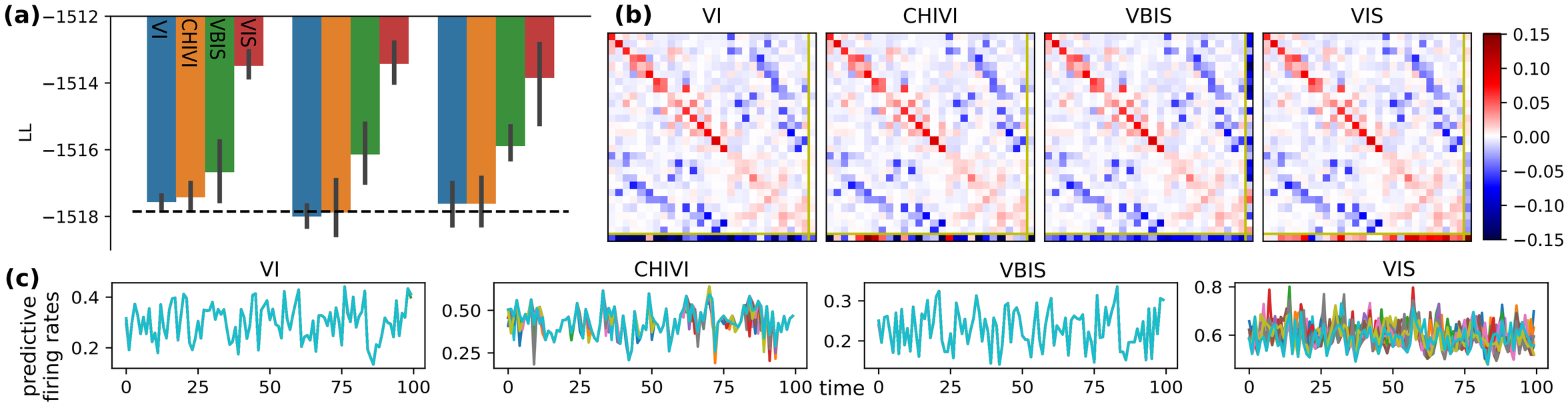
Partially observable GLM (retinal ganglion cell)

- (b) The hidden representative neuron learned by VIS behaves like an OFF cell. The sign of the weight from this hidden neuron to the visible neurons clearly tells us the type of those visible post-synaptic neurons.



Partially observable GLM (retinal ganglion cell)

- (c) 20 randomly sampled predictive firing rates from $q(\boldsymbol{Z} \mid \boldsymbol{X}; \phi)$. The $q(\boldsymbol{Z} \mid \boldsymbol{X}; \phi)$ learned by VIS provides the largest variability, which further improves the effectiveness of learning θ.



Summary

- VIS is the best way of doing importance sampling (IS), from the perspective of statistics.
- VIS has succinct and numerically stable gradient estimator derived in log space.
- By only changing two lines in the code, VIS learns significantly better model parameters, and achieves better test performance.

Thanks for listening!