Semi-Implicit Variational Inference

Mingzhang Yin§
Joint work with Mingyuan Zhou*§

The University of Texas at Austin
§Department of Statistics and Data Sciences
*IROM Department, McCombs School of Business

International Conference on Machine Learning
Stockholm, Sweden, July 11, 2018
Bayesian Inference

- Bayes’ rule:
  \[ P(z \mid X) = \frac{P(X \mid z)P(z)}{P(X)} = \frac{P(X \mid z)P(z)}{\int P(X \mid z)P(z)dz} \]

  Posterior of \( z \) given \( X \) = Conditional Likelihood \( \times \) Prior \( \frac{\text{Marginal Likelihood}}{\text{Marginal Likelihood}} \)

- Two main ways for approximate Bayesian inference:
  - Draw \( z \sim P(z \mid X) \) using Markov chain Monte Carlo (MCMC) based methods such as Gibbs sampling: iteratively sample \( P(z_k \mid X, z \setminus z_k) \)
  - Approximate the posterior \( P(z \mid X) \) with \( Q(z) \), which is straightforward to sample from, using an optimization method such as Laplace approximation and variational inference
Variational inference

- Evidence and ELBO:

\[
\ln P(X) = \int Q(z) \ln \frac{P(X, z)}{Q(z)} \, dz + \int Q(z) \ln \frac{Q(z)}{P(z \mid X)} \, dz \\
= \mathcal{L}(Q) + \text{KL}(Q(z) \mid \mid P(z \mid X)).
\]

- Since \(\text{KL}(Q(z) \mid \mid P(z \mid X)) \geq 0\), minimizing the Kullback-Leibler (KL) divergence from \(P(z \mid X)\) to \(Q(z)\) is the same as maximizing the evidence lower bound:

\[
\min_Q \text{KL}(Q(z) \mid \mid P(z \mid X)) \Leftrightarrow \max_Q \text{ELBO}
\]

\[
\text{ELBO} = \mathcal{L}(Q) = \mathbb{E}_Q[\ln P(X, z)] - \mathbb{E}_Q[\ln Q(z)] \\
= \mathbb{E}_Q[\ln P(X \mid z)] - \text{KL}(Q(z) \mid \mid P(z))
\]

- Variational inference converts the problem of posterior inference into an optimization problem.
Mean-field variational inference

- Mean-field variational inference (VI) factorizes the $Q$ distribution of $z = (z_1, \ldots, z_K)^T$ as

$$Q(z) = \prod_{i=1}^{K} q_{\phi_i}(z_i)$$

- The factorized assumption allows for closed-form coordinate ascent updates:

$$q^*(z_k) = \frac{\exp \left\{ \mathbb{E}_{q(z_{-k})} [\log p(X, z_k, z_{-k})] \right\}}{\int \exp \left\{ \mathbb{E}_{q(z_{-k})} [\log p(X, z_k, z_{-k})] \right\} dz_k}, \quad k = 1, \ldots, K$$

where $z_{-k} = \{z_1, \ldots, z_{k-1}, z_{k+1}, \ldots, z_K\}$.

- However, mean-field VI often clearly underestimates the variance of the posterior, due to the use of KL divergence and two restrictive constraints:
  - $q(z_k)$ are often restricted to the exponential family
  - The dependencies between $z_k$ cannot be captured
Model:

\[ x_i \overset{i.i.d.}{\sim} \text{NB}(r, p), \quad r \sim \text{Gamma}(a, 1/b), \quad p \sim \text{Beta}(\alpha, \beta), \]

Mean-field VI:

\[ Q(r, p) = q(r)q(p) = \text{Gamma}(r; \tilde{a}, \tilde{b})\text{Beta}(p; \tilde{\alpha}, \tilde{\beta}), \]

Mean-field VI underestimates variance (mainly due to the factorized assumption):

![Graph showing comparison between SIVI, MCMC, and Mean-field distributions for r and p.]
“Modern” variational inference

Choose a more flexible $Q_{\phi}(z)$ and infer the variational parameter $\phi$ via (stochastic) gradient descent (by reparameterization or score method)

$$\nabla_{\phi} \mathcal{L}(Q_{\phi}(z)) = \nabla_{\phi} \mathbb{E}_{z \sim Q_{\phi}(z)} \left[ \ln \frac{P(X, z)}{Q_{\phi}(z)} \right]$$

- There are two major flexibilities we want $Q_{\phi}(z)$ have:
  - We wish $Q_{\phi}(z)$ is not restricted to have an analytic density (but should be easy to sample)
  - We wish $Q_{\phi}(z)$ to incorporate dependencies of latent variables

- We also want to maintain computational tractability for a flexible inference distribution

To achieve the computation and accuracy balance, we use the neural network implicit distribution in a hierarchical model.
Implicit distribution

Implicit distribution consists of a source of randomness $q(\epsilon)$ and a deterministic transform $T_\phi : \mathbb{R}^p \to \mathbb{R}^d$

$$z = T_\phi(\epsilon), \ \epsilon \sim q(\epsilon)$$

- When $T_\phi$ is invertible and the dimension is low, the density

$$q_\phi(z) = \frac{\partial}{\partial z_1} \cdots \frac{\partial}{\partial z_d} \int_{T_\phi(\epsilon) \leq z} q(\epsilon) d\epsilon$$

can be calculated using change of variables. But in general $\{T_\phi(\epsilon) \leq z\}$ cannot be calculated and hence the high dimension integral is intractable, making $q_\phi(z)$ become implicit

- Direct inference with implicit distribution can be difficult because of the need to estimate the density ratio $\frac{P(X, z)}{Q_\phi(z)}$
Hierarchical variational family

Capturing the latent variable dependencies plays the key role to accurately estimate the uncertainty.

- One way is to add a hierarchical structure that assumes $z_k$ to be conditional independent but marginally dependent, using

$$q(z \mid \psi) = \prod_{k=1}^{K} q(z_k \mid \psi_k), \quad \psi \sim q_{\phi}(\psi)$$

- Marginalizing $\psi$ out, we can view $z$ as a variable drawn from the distribution family $H$ which we choose as variational family

$$H = \left\{ h_{\phi}(z) : h_{\phi}(z) = \mathbb{E}_{\psi \sim q_{\phi}(\psi)}[q(z \mid \psi)] = \int_{\psi} \left[ \prod_{k=1}^{K} q(z_k \mid \psi_k) \right] q_{\phi}(\psi) d\psi \right\}$$

- It is evident that $q(z \mid \psi) \in Q \subseteq H$, i.e., $H$ is an expansion of the original variational distribution family
Semi-implicit variational inference (SIVI)

- We call the hierarchical model semi-implicit because it requires $q(z | \psi)$ to be explicit while allows $q_\phi(\psi)$ to be implicit, and $h_\phi(z) = \mathbb{E}_{q_\phi(\psi)} q(z | \psi)$ and ELBO is generally not analytic.

- KL convexity and Jensen’s inequality lead to an ELBO lower bound:

$$
\mathcal{L}(q(z | \psi), q_\phi(\psi)) = \mathbb{E}_{\psi \sim q_\phi(\psi)} \mathbb{E}_{z \sim q(z | \psi)} \log \frac{p(x,z)}{q(z | \psi)}
$$

$$
= - \mathbb{E}_{\psi \sim q_\phi(\psi)} \text{KL}(q(z | \psi) \| p(z | x)) + \log p(x)
$$

$$
\leq - \text{KL}(\mathbb{E}_{\psi \sim q_\phi(\psi)} q(z | \psi) \| p(z | x)) + \log p(x)
$$

$$
= \mathcal{L} = \mathbb{E}_{z \sim h_\phi(z)} \log \frac{p(x,z)}{h_\phi(z)}
$$

- Using the concavity of the logarithmic function, we have

$$
\log h_\phi(z) \geq \mathbb{E}_{\psi \sim q_\phi(\psi)} \log q(z | \psi)
$$

and hence an ELBO upper bound:

$$
\bar{\mathcal{L}}(q(z | \psi), q_\phi(\psi)) = \mathbb{E}_{\psi \sim q_\phi(\psi)} \mathbb{E}_{z \sim h_\phi(z)} \log \frac{p(x,z)}{q(z | \psi)} \geq \mathcal{L}
$$

- Note there is a subtle but critical difference between $\mathcal{L}$ and $\bar{\mathcal{L}}$. 

Degeneracy of $\mathcal{L}$

Maximizing the surrogate lower bound $\mathcal{L}$ may lead to degeneracy that $q_\phi(\psi)$ converges to a point mass density:

**Proposition (Degeneracy)**

Let us denote $\psi^* = \arg \max_\psi -\mathbb{E}_{z \sim q(z|\psi)} \log \frac{q(z|\psi)}{p(x,z)}$, then

$$\mathcal{L}(q(z|\psi), q_\phi(\psi)) \leq -\mathbb{E}_{z \sim q(z|\psi^*)} \log \frac{q(z|\psi^*)}{p(x,z)},$$

where the equality is true if and only if $q_\phi(\psi) = \delta_{\psi^*}(\psi)$. 
Asymptotically exact ELBO

- Avoid degeneracy by adding regularization $\mathcal{L}_K = \mathcal{L} + B_K$

\[
B_K = \mathbb{E}_{\psi,\psi^{(1)},...\psi^{(K)} \sim q_\phi(\psi)} \text{KL}(q(z | \psi) || \tilde{h}_K(z)),
\]

(1)

where $\tilde{h}_K(z) = \frac{1}{K+1} [q(z | \psi) + \sum_{k=1}^{K} q(z | \psi^{(k)})]$, $B_K \geq 0$, with $B_K = 0$ if and only if $K = 0$ or $q_\phi(\psi)$ degenerates to a point mass density.

- The Jensen gap can also be narrowed from upper side by $\bar{\mathcal{L}}_k = \bar{\mathcal{L}} - A_k$

\[
A_K = \mathbb{E}_{\psi \sim q_\phi(\psi)} \mathbb{E}_{z \sim h_\phi(z)} \mathbb{E}_{\psi^{(1)},...\psi^{(K)} \sim q_\phi(\psi)} \left[ \log \left( \frac{1}{K} \sum_{k=1}^{K} q(z | \psi^{(k)}) \right) - \log q(z | \psi) \right]
\]

The regularized lower bound $\underline{\mathcal{L}}_K$ is an asymptotically exact ELBO that satisfies $\underline{\mathcal{L}}_0 = \mathcal{L}$ and $\lim_{K \to \infty} \underline{\mathcal{L}}_K = \mathcal{L}$. The regularized upper bound satisfies $\bar{\mathcal{L}}_1 = \bar{\mathcal{L}}$, $\bar{\mathcal{L}}_{K+1} \leq \bar{\mathcal{L}}_K$, and $\lim_{K \to \infty} \bar{\mathcal{L}}_K = \mathcal{L}$.
Algorithm for SIVI

Algorithm 1 Semi-Implicit Variational Inference (SIVI)

**input**: Data \( \{x_i\}_{1:N} \), joint likelihood \( p(x, z) \), explicit variational distribution \( q_\xi(z \mid \psi) \) with reparameterization \( z = f(\epsilon, \xi, \psi) \), \( \epsilon \sim p(\epsilon) \), implicit layer neural network \( T_\phi(\epsilon) \) and source of randomness \( q(\epsilon) \)

**output**: Variational parameter \( \xi \) for the conditional distribution \( q_\xi(z \mid \psi) \), variational parameter \( \phi \) for the mixing distribution \( q(\psi) \)

Initialize \( \xi \) and \( \phi \) randomly

while not converged do

Set \( L_{K_t} = 0 \), \( \rho_t \) and \( \eta_t \) as step sizes, and \( K_t \geq 0 \) as a non-decreasing integer; Sample \( \psi^{(k)} = T_\phi(\epsilon^{(k)}) \), \( \epsilon^{(k)} \sim q(\epsilon) \) for \( k = 1, \ldots, K_t \); take subsample \( x = \{x_i\}_{i=1}^{i_M} \)

for \( j = 1 \) to \( J \) do

Sample \( \psi_j = T_\phi(\epsilon_j) \), \( \epsilon_j \sim q(\epsilon) \)

Sample \( z_j = f(\tilde{\epsilon}_j, \xi, \psi_j) \), \( \tilde{\epsilon}_j \sim p(\epsilon) \)

\( L_{K_t} = L_{K_t} + \frac{1}{J} \left\{ - \log \frac{1}{K_t+1} \left[ \sum_{k=1}^{K_t} q_\xi(z_j \mid \psi^{(k)}) + q_\xi(z_j \mid \psi_j) \right] + \frac{N}{M} \log p(x \mid z_j) + \log p(z_j) \right\} \)

end

\( t = t + 1 \)

\( \xi = \xi + \rho_t \nabla_\xi L_{K_t} (\{\psi^{(k)}\}_{1,K_t}, \{\psi_j\}_{1,J}, \{z_j\}_{1,J}) \)

\( \phi = \phi + \eta_t \nabla_\phi L_{K_t} (\{\psi^{(k)}\}_{1,K_t}, \{\psi_j\}_{1,J}, \{z_j\}_{1,J}) \)

end
Methods to expand variational distribution family

Expand variational family via stochastic and/or deterministic method

- Hierarchical models: eg. Negative Binomial ⇔ Poisson-Gamma hierarchy; Hierarchical variational model (Ranganath et al., 2016)

- Normalizing Flow: transfer simple distribution with a chain of simple invertible mapping $z_t = f_t \circ \cdots \circ f_0(z_0)$ (Rezende and Mohamed, 2015)

- Modeling the dependencies between univariate marginals with copula (Tran et al., 2015)

- Implicit distribution $z = f(\epsilon)$, where $f$ is not invertible; (Tran et al., 2017)

- Our approach: hierarchy with explicit conditional layer, implicit mixing layers (semi-implicit)
Expressiveness of SIVI

\[ h(z) = \mathbb{E}_{\psi \sim q(\psi)} q(z | \psi) \]

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>( z \sim \mathcal{N}(\psi, 0.1) )</td>
<td>( \psi \sim q(\psi) )</td>
</tr>
<tr>
<td>( \mathcal{N}(z; -2, 1) + 0.7\mathcal{N}(z; 2, 1) )</td>
<td></td>
</tr>
<tr>
<td>( \mathcal{N}(\psi; 0.1, 0.1) )</td>
<td>( \psi \sim q(\psi) )</td>
</tr>
<tr>
<td>( 0.5\mathcal{N}(z; -2, I) + 0.5\mathcal{N}(z; 2, I) )</td>
<td></td>
</tr>
<tr>
<td>( \mathcal{N}(\psi; 2/4, 1)\mathcal{N}(\psi; 0, 4) )</td>
<td>( \mathcal{N}(\psi; 1.8, 2) )</td>
</tr>
<tr>
<td>( 0.5\mathcal{N}(z; 0, \begin{bmatrix} 2 &amp; 1.8 \ 2 &amp; -1.8 \end{bmatrix}) + 0.5\mathcal{N}(z; 0, \begin{bmatrix} 2 &amp; -1.8 \ -1.8 &amp; 2 \end{bmatrix}) )</td>
<td></td>
</tr>
</tbody>
</table>
Model:

\[ x_i \overset{i.i.d.}{\sim} \text{NB}(r, p), \ r \sim \text{Gamma}(a, 1/b), \ p \sim \text{Beta}(\alpha, \beta), \]

Mean-field VI:

\[ Q(r, p) = q(r)q(p) = \text{Gamma}(r; \tilde{a}, \tilde{b})\text{Beta}(p; \tilde{\alpha}, \tilde{\beta}), \]

SIVI (both the conditional and mixing \( q \) distributions are reparameterizable):

\[ q(r, p \mid \psi) = \text{Log-Normal}(r; \mu_r, \sigma_r^2)\text{Logit-Normal}(p; \mu_p, \sigma_p^2), \]

\[ \psi = (\mu_r, \mu_p) \sim q(\psi), \]
Figure: Kolmogorov-Smirnov (KS) distance and its corresponding p-value between the marginal posteriors of $r$ and $p$ inferred by SIVI and MCMC. SIVI rapidly improves as $K$ increases.
Score function gradient for conjugate model

If \( q(z | \psi) \) is not reparameterizable, then we introduce a density ratio as

\[
    r_{\xi, \phi}(z, \epsilon, \epsilon^{(1:K)}) = \frac{q_{\xi}(z | T_{\phi}(\epsilon))}{\frac{1}{K+1} [q_{\xi}(z | T_{\phi}(\epsilon)) + \sum_{k=1}^{K} q_{\xi}(z | T_{\phi}(\epsilon^{(k)}))]} 
\]

and approximate the gradient of \( \mathcal{L}_K \) with respect to \( \phi \) as

\[
    \nabla_{\phi} \mathcal{L}_K \approx \frac{1}{J} \sum_{j=1}^{J} \left\{ -\nabla_{\phi} \mathbb{E}_{z \sim q_{\xi}(z | T_{\phi}(\epsilon_j))} \left[ \log \frac{q_{\xi}(z | T_{\phi}(\epsilon_j))}{p(x, z)} \right] + \nabla_{\phi} \log r_{\xi, \phi}(z_j, \epsilon_j, \epsilon^{(1:K)}) \right. \\
    \left. + \left[ \nabla_{\phi} \log q_{\xi}(z_j | T_{\phi}(\epsilon_j)) \right] \log r_{\xi, \phi}(z_j, \epsilon_j, \epsilon^{(1:K)}) \right\},
\]

- The first summation term is equivalent to the gradient of MFVI’s ELBO
- Both the second and third terms correct the restrictions of \( q_{\xi}(z | T_{\phi}(\epsilon_j)) \)
- \( \log r_{\xi, \phi}(z, \epsilon, \epsilon^{(1:K)}) \) in the third term is expected to be small regardless of convergence, effectively mitigating the variance of score function gradient estimation that is usually high in basic black-box VI
Model:

\[ p(n_i, l_i \mid r, p) = r^{l_i} p^{n_i} (1 - p)^r / Z_i, \quad r \sim \text{Gamma}(a, 1/b), \quad p \sim \text{Beta}(\alpha, \beta) \]

Mean-filed VI:

\[ Q(r, p) = q(r)q(p) = \text{Gamma}(r; \tilde{a}, \tilde{b})\text{Beta}(p; \tilde{\alpha}, \tilde{\beta}), \]

SIVI (non-reparameterizable conditional \( q \) distribution but conjugate model):

\[ q(r, p \mid \psi) = \text{Gamma}(r; \psi_1, \psi_2)\text{Beta}(p; \psi_3, \psi_4), \quad \psi = (\psi_1, \psi_2, \psi_3, \psi_4) \sim q(\psi) \]
Bayesian logistic regression (pairwise joint distributions)

\[ y_i \sim \text{Bernoulli}\left( (1 + e^{-x_i' \beta})^{-1} \right), \quad \beta \sim \mathcal{N}(0, \alpha^{-1} I_{v+1}) \]

SIVI: \( q(\beta | \psi) = \mathcal{N}(\psi, \Sigma), \quad \psi \sim q_\phi(\psi) \)

(Blue: MCMC, Red: VI, Green: SIVI):
Bayesian logistic regression

![Figure: Comparing univariate marginals](image)

**Figure:** Comparing posterior covariance matrix

![Figure: Comparing posterior covariance matrix](image)
Bayesian logistic regression (predictive uncertainty)

**Figure**: Comparison of MFVI (red) with a full covariance matrix, MCMC (green on left), and SIVI (green on right) with a full covariance matrix on quantifying predictive uncertainty for Bayesian logistic regression on *waveform*. 
We construct semi-implicit VAE (SIVAE) by using a hierarchical encoder that injects random noise at $M$ different stochastic layers as

$$
\ell_t = T_t(\ell_{t-1}, \epsilon_t, x; \phi), \quad \epsilon_t \sim q_t(\epsilon), \quad t = 1, \ldots, M,
$$

$$
\mu(x, \phi) = f(\ell_M, x; \phi), \quad \Sigma(x, \phi) = g(\ell_M, x; \phi),
$$

$$
q_{\phi}(z | x, \mu, \Sigma) = \mathcal{N}(\mu(x, \phi), \Sigma(x, \phi)),
$$

where $\ell_0 = \emptyset$ and $T_t$, $f$, and $g$ are all deterministic neural networks. Note given data $x_i$, $\mu(x_i, \phi)$, $\Sigma(x_i, \phi)$ are now random variables rather than following vanilla VAE to assume deterministic values. This clearly moves the encoder variational distribution beyond a simple Gaussian form.
### Semi-implicit variational autoencoder

<table>
<thead>
<tr>
<th>Methods</th>
<th>$-\log p(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Results below form Barda et al. (2015)</strong></td>
<td></td>
</tr>
<tr>
<td>VAE + IWAE</td>
<td>= 86.76</td>
</tr>
<tr>
<td>IWAE + IWAE</td>
<td>= 84.78</td>
</tr>
<tr>
<td><strong>Results below form Salimans et al. (2015)</strong></td>
<td></td>
</tr>
<tr>
<td>DLGM + HVI (1 leapfrog step)</td>
<td>= 88.08</td>
</tr>
<tr>
<td>DLGM + HVI (4 leapfrog step)</td>
<td>= 86.40</td>
</tr>
<tr>
<td>DLGM + HVI (8 leapfrog steps)</td>
<td>= 85.51</td>
</tr>
<tr>
<td><strong>Results below form Rezende &amp; Mohamed (2015)</strong></td>
<td></td>
</tr>
<tr>
<td>DLGM+NICE (Dinh et al., 2014) (k = 80)</td>
<td>≤ 87.2</td>
</tr>
<tr>
<td>DLGM+NF (k = 40)</td>
<td>≤ 85.7</td>
</tr>
<tr>
<td>DLGM+NF (k = 80)</td>
<td>≤ 85.1</td>
</tr>
<tr>
<td><strong>Results below form Gregor et al. (2015)</strong></td>
<td></td>
</tr>
<tr>
<td>DLGM</td>
<td>≈ 86.60</td>
</tr>
<tr>
<td>NADE</td>
<td>= 88.33</td>
</tr>
<tr>
<td>DBM 2hl</td>
<td>≈ 84.62</td>
</tr>
<tr>
<td>DBN 2hl</td>
<td>≈ 84.55</td>
</tr>
<tr>
<td>EoNADE-5 2hl (128 orderings)</td>
<td>= 84.68</td>
</tr>
<tr>
<td>DARN 1hl</td>
<td>≈ 84.13</td>
</tr>
<tr>
<td><strong>Results below form Maaløe et al. (2016)</strong></td>
<td></td>
</tr>
<tr>
<td>Auxiliary VAE (L=1, IW=1)</td>
<td>≤ 84.59</td>
</tr>
<tr>
<td><strong>Results below form Mescheder et al. (2017)</strong></td>
<td></td>
</tr>
<tr>
<td>VAE + IAF (Kingma et al., 2016)</td>
<td>≈ 84.9 ± 0.3</td>
</tr>
<tr>
<td>Auxiliary VAE (Maaløe et al., 2016)</td>
<td>≈ 83.8 ± 0.3</td>
</tr>
<tr>
<td>AVB + AC</td>
<td>≈ 83.7 ± 0.3</td>
</tr>
<tr>
<td>SIVI (3 stochastic layers)</td>
<td>= 84.07</td>
</tr>
<tr>
<td>SIVI (3 stochastic layers)+ IW($\tilde{K} = 10$)</td>
<td>= 83.25</td>
</tr>
</tbody>
</table>
Uncertainty estimation is difficult but important in Variational Inference.

One key to get an accurate uncertainty estimation is to construct a flexible variational distribution that can capture the dependencies between latent variables.

Balancing the expressiveness and tractability, semi-implicit variational inference (SIVI) can approach the accuracy of MCMC in quantifying posterior uncertainty, but often pays a lower computational cost and can generate independent posterior samples fast via the inferred stochastic variational inference network.
Thank you!

Welcome to our poster at Hall B # 177
Related inference methods

- VAE: Changing the empirical data distribution leads to degenerated $\mathcal{L}$

$$\mathcal{L}_{\text{VAE}} = \mathbb{E}_{x \sim D(x)} \mathbb{E}_{z \sim q(z|f(x))} \log \frac{p(x, z)}{q(z|f(x))}$$

$$\mathcal{L}_{SIVI} = \mathbb{E}_{\epsilon \sim q(\epsilon)} \mathbb{E}_{z \sim q(z|f(\epsilon))} \log \frac{p(x, z)}{q(z|f(\epsilon))}$$

- Data augmentation: iteratively sample from $p(z|\psi)$ and $p(\psi|z)$ with

$$p(z) = \int p(z, \psi) d\psi$$

- Auxiliary Deep Generative Models (Maaløe et al., 2016): optimize on a less tighter bound

$$\log p(x) \geq \mathbb{E}_{h_\phi(z|x)} \log \frac{p(x, z)}{h_\phi(z|x)} \geq \mathbb{E}_{q_\phi(z, a|x)} \log \frac{p(x, z, a)}{q_\phi(z, a|x)}$$