Variational Inference and Deep Generative Models

Addison Bohannon
US Army Research Laboratory
April 3, 2017

UNCLASSIFIED; APPROVED FOR PUBLIC RELEASE; DISRIBUTION UNLIMITED

U.S. ARMY RDECOM

Overview

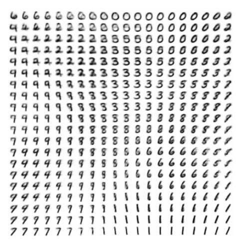
ARL

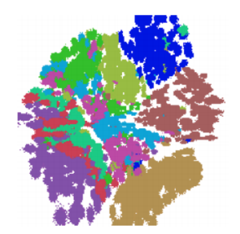
How do we combine variational inference and deep generative modeling into a common algorithm?

Auto-Encoding Variational Bayes

Diederik P. Kingma
Machine Learning Group
Universiteit van Amsterdam
dpkingma@gmail.com

Max Welling
Machine Learning Group
Universiteit van Amsterdam
welling.max@gmail.com





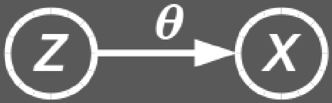
Stochastic Backpropagation and Approximate Inference in Deep Generative Models

Danilo J. Rezende, Shakir Mohamed, Daan Wierstra
{danilor, shakir, daanw}@google.com
Google DeepMind, London

Table of Contents

Table of Contents

Latent variable model



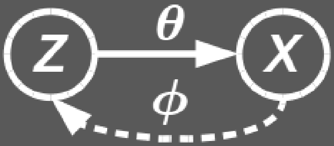
- $X \in \mathbb{R}^n$ is an *observed random variable*
- $Z \in \mathbb{R}^m$ is a *latent random variable*

- Directed probabilistic model
  - $X|Z \sim f_{X|Z}(x|z;\theta)$
- Prior
  - $Z \sim f_Z(z)$
- Posterior density
  - $\frac{f_{X|Z}(x|z;\theta)f_Z(z)}{\int_{\mathbb{R}^m} f_{X|Z}(x|z';\theta)f_Z(z')dz'}$
  - $\frac{f_{X|Z}(x|z;\theta)f_Z(z)}{\frac{1}{N}\sum_{i=1}^N f_{X|Z}(x|z_i;\theta)}$

Table of Contents

Variational inference

What if we can accept an approximation of the true posterior?



$$\left.\begin{aligned}\mathcal{R}(\phi,\theta) &= \mathbb{E}_\theta \mathcal{L}(\phi,\theta)\\ \mathcal{L}(\phi,\theta) &= \mathcal{D}_{KL}\left(q(z|x;\phi)\|f_{Z|X}(z|x;\theta)\right)\end{aligned}\right\} q^* = \arg\min_{q\in\mathcal{F}(\phi)} \mathcal{R}(\phi,\theta)$$

U.S. ARMY RDECOM

Variational inference

ARL

$$
\begin{aligned}
q^* &= \underset{q \in \mathcal{F}(\phi)}{\arg\min}\, \mathcal{R}(\phi, \theta) \\
&= \underset{q \in \mathcal{F}(\phi)}{\arg\min}\, \mathbb{E}_{x \sim f_X} \mathcal{D}_{KL}\left(q(z|x;\phi) \| f_{Z|X}(z|x;\theta)\right) \\
&= \underset{q \in \mathcal{F}(\phi)}{\arg\min}\, \mathbb{E}_{x \sim f_X} \mathbb{E}_{z|x \sim q} \log \frac{q(z|x;\phi)}{f_{Z|X}(z|x;\theta)} \\
&= \underset{q \in \mathcal{F}(\phi)}{\arg\min}\, \mathbb{E}_{x \sim f_X} \mathbb{E}_{z|x \sim q} \log \frac{q(z|x;\phi) f_X(x)}{f_{X|Z}(x|z;\theta) f_Z(z)} \\
&= \underset{q \in \mathcal{F}(\phi)}{\arg\min}\, \mathbb{E}_{x \sim f_X} \mathbb{E}_{z|x \sim q} \log \frac{q(z|x;\phi)}{f_{X|Z}(x|z;\theta) f_Z(z)} + \mathbb{E}_{x \sim f_X} \log f_X(x) \\
&= \underset{q \in \mathcal{F}(\phi)}{\arg\min}\, \mathbb{E}_{x \sim f_X} \mathbb{E}_{z|x \sim q} \log \frac{q(z|x;\phi)}{f_{X|Z}(x|z;\theta) f_Z(z)}
\end{aligned}
$$

Variational inference

Variational inference objective function:

$$q^* = \arg\min_{q \in \mathcal{F}(\phi)} \underbrace{\mathbb{E}_{x \sim f}\, \mathbb{E}_{z|x \sim q} \log \frac{q(z|x;\phi)}{f_{X|Z}(x|z;\theta) f_Z(z)}}_{\text{Log Evidence Lower Bound}}$$

Relationship to autoencoders:

$$\phi^* = \arg\min_{\phi} \mathbb{E}_{x \sim f} \left[\underbrace{-\mathbb{E}_{z|x \sim q} \log f_{X|Z}(x|z;\theta)}_{\text{encoding-decoding loss}} + \underbrace{\mathcal{D}_{KL}(q(z|x;\phi) \| f_Z(z))}_{\text{regularization}} \right]$$

U.S. ARMY RDECOM

Table of Contents

ARL 25

U.S. ARMY RDECOM

Gradient estimation

ARL

We want to use stochastic optimization techniques which require only gradient evaluations:

$$
\begin{aligned}
&\nabla_\phi \mathbb{E}_{x\sim f} \mathbb{E}_{z|x\sim q} \log \frac{q(z|x;\phi)}{f_{X|Z}(x|z;\theta) f_Z(z)} \\
&\neq \mathbb{E}_{x\sim f_X} \mathbb{E}_{z|x\sim q} \nabla_\phi \log \frac{q(z|x;\phi)}{f_{X|Z}(x|z;\theta) f_Z(z)} \\
&= \mathbb{E}_{x\sim f_X} \mathbb{E}_{z|x\sim q} \nabla_\phi \log q(z|x;\phi) \\
&\approx \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} \nabla_\phi \log q(z_j|x_i;\phi), \;\; x_i \sim f_X, z_j|x \sim q
\end{aligned}
$$

Direct computation

We can use the score method to form the gradient as an expectation:

$$
\begin{aligned}
&\nabla_\phi \mathbb{E}_\phi g(x,\phi) \\
&\quad = \nabla_\phi \int g(x,\phi) f(x;\phi) dx \\
&\quad = \int \nabla_\phi \left[g(x,\phi) f(x;\phi) \right] dx \\
&\quad = \int \left(\nabla_\phi g(x,\phi) \right) f(x;\phi) + g(x,\phi) \nabla_\phi f(x;\phi) dx \\
&\quad = \int \left(\nabla_\phi g(x,\phi) \right) f(x;\phi) + g(x,\phi) \left(\nabla_\phi \log f(x;\phi) \right) f(x;\phi) dx \\
&\quad = \mathbb{E}_\phi \left[\nabla_\phi g(x,\phi) + g(x,\phi) \nabla_\phi \log f(x;\phi) \right]
\end{aligned}
$$

Direct computation

This results in a "high-variance" gradient estimator:

$$
\begin{aligned}
&\nabla_\phi \mathbb{E}_{x\sim f}\mathbb{E}_{z|x\sim q} \log \frac{q(z|x;\phi)}{f_{X|Z}(x|z;\theta) f_Z(z)} \\
&= \mathbb{E}_{x\sim f_X}\mathbb{E}_{z|x\sim q}\left[\left(1 + \log \frac{q(z|x;\phi)}{f_{X|Z}(x|z;\theta) f_Z(z)}\right)\nabla_\phi \log q(z|x;\phi)\right] \\
&\approx \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\left[\left(1 + \underbrace{\log \frac{q(z_j|x_i;\phi)}{f_{X|Z}(x_i|z_j;\theta) f_Z(z_j)}}_{\text{correction term}}\right)\nabla_\phi \log q(z_j|x_i;\phi)\right],
\end{aligned}
$$

$$x_i \sim f_X,\, z_j|x \sim q$$

Re-parameterization trick

Alternatively, we could use a change of variable to remove the dependence of $\mathbb{E}_{z|x\sim q}$ on ϕ:

$$\nabla_\phi \mathbb{E}_{x\sim f}\mathbb{E}_{z|x\sim q} \log \frac{q(z|x;\phi)}{f_{X|Z}(x|z;\theta)f_Z(z)}$$

$$= \nabla_\phi \mathbb{E}_{x\sim f_X}\mathbb{E}_{w\sim f} \left(\log \frac{q(z|x;\phi)}{f_{X|Z}(x|z;\theta)f_Z(z)} \circ g \right)(w)$$

$$= \mathbb{E}_{x\sim f_X}\mathbb{E}_{w\sim f} \nabla_\phi \left(\log \frac{q(z|x;\phi)}{f_{X|Z}(x|z;\theta)f_Z(z)} \circ g \right)(w)$$

$$\approx \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} \nabla_\phi \left(\log \frac{q(z|x;\phi)}{f_{X|Z}(x_i|z;\theta)f_Z(z)} \circ g \right)(w_j),$$

$$x_i \sim f_X,\, w_j|x \sim f_W$$

Re-parameterization

Theorem: Change of variable

Let $U, V \subset \mathbb{R}^n$ be open sets and $g : U \to V$ be an invertible map for which $g, g^{-1} \in C^1$. Then, for an absolutely integrable function, $f : V \to \mathbb{R}$,

$$\int_V f(x)dx = \int_U f \circ g(y')J_y g(y')dy'$$

provided that the Jacobian does not vanish on more than a set of measure zero.

U.S. ARMY RDECOM

Re-parameterization

ARL

Corollary: Change of variable (expectation)

Let $W \in \mathbb{R}^n$ be a random variable and $g : \mathbb{R}^n \to \mathbb{R}^n$ be an invertible map for which $g, g^{-1} \in C^1$. Let $Z = g(W)$. Then, for any absolutely integrable function, $h : \mathbb{R}^n \to \mathbb{R}$,

$$\mathbb{E}_{w \sim f_W}(h \circ g)(W) = \mathbb{E}_{z \sim f_z} h(Z)$$

where $f_Z(z') = (f_W \circ g^{-1})(z') J_z g^{-1}(z')$, provided that the Jacobian does not vanish on more than a set of measure zero.

Re-parameterization

Consider a random variable, $W \in \mathbb{R}^m$, and an invertible map, $g_\phi : \mathbb{R}^m \to \mathbb{R}^m$. Then, we want a model distribution:

$$q(z'|x;\phi) = (f_W \circ g_\phi^{-1})(z') J_z g_\phi^{-1}(z').$$

where

$$W \sim f_W(w)$$
$$Z = g_\phi(W)$$

Re-parameterization

- Location-scale family

$$\left.\begin{aligned} W &\sim \mathcal{N}(w; 0, I) \\ Z &= \mu + \Sigma^{\frac{1}{2}} W \end{aligned}\right\} Z \sim \mathcal{N}(z; \mu, \Sigma)$$

- Inverse cumulative distribution function

$$\left.\begin{aligned} W &\sim \mathcal{U}(0, 1) \\ Z &= -\frac{1}{\lambda}\log(1 - W) \end{aligned}\right\} Z \sim \exp(\lambda)$$

- Transformations

$$\left.\begin{aligned} W &\sim \mathcal{N}(z; \mu, \sigma^2) \\ Z &= \exp(W) \end{aligned}\right\} Z \sim \mathcal{N}(\log z; \mu, \sigma^2)$$

Re-parameterization

Given $q(z'|x;\phi) = (f_W \circ g_\phi^{-1})(z') J_z g_\phi^{-1}(z')$:

$$
\begin{aligned}
&\mathbb{E}_{x\sim f}\mathbb{E}_{z|x\sim q} \log \frac{q(z|x;\phi)}{f_{X|Z}(x|z;\theta) f_Z(z)} \\
&\quad = \mathbb{E}_{x\sim f}\mathbb{E}_{z|x\sim q} \log \frac{(f_W \circ g_\phi^{-1})(z) J_z g_\phi^{-1}(z)}{f_{X|Z}(x|z;\theta) f_Z(z)} \\
&\quad = \mathbb{E}_{x\sim f}\mathbb{E}_{w'\sim f} \log \frac{f_W(w') J_{g_\phi(w)}(g_\phi^{-1} \circ g_\phi)(w')}{f_{X|Z}(x|g_\phi(w');\theta)(f_Z \circ g_\phi)(w')} \\
&\quad = \mathbb{E}_{x\sim f}\mathbb{E}_{w'\sim f} \log \frac{f_W(w')}{f_{X|Z}(x|g_\phi(w');\theta)(f_Z \circ g_\phi)(w') J_w g_\phi(w')}
\end{aligned}
$$

U.S. ARMY RDECOM ARL

Re-parameterization

Now, we can move the gradient inside of both expectations for gradient estimation:

$$
\begin{aligned}
&\nabla_\phi \mathbb{E}_{x\sim f} \mathbb{E}_{w'\sim f} \log \frac{f_W(w')}{f_{X|Z}(x|g_\phi(w');\theta)(f_Z \circ g_\phi)(w') J_w g_\phi(w')} \\
&= \mathbb{E}_{x\sim f_X} \mathbb{E}_{w'\sim f} \nabla_\phi \log \frac{f_W(w')}{f_{X|Z}(x|g_\phi(w');\theta)(f_Z \circ g_\phi)(w') J_w g_\phi(w')} \\
&= -\mathbb{E}_{x\sim f_X} \mathbb{E}_{w'\sim f} \nabla_\phi \log \left[f_{X|Z}(x|g_\phi(w');\theta)(f_Z \circ g_\phi)(w') J_w g_\phi(w') \right] \\
&= -\frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} \nabla_\phi \log \left[f_{X|Z}(x_i|g_\phi(w_j);\theta)(f_Z \circ g_\phi)(w_j) J_w g_\phi(w_j) \right], \\
&\quad x_i \sim f_X, w_j \sim f_W
\end{aligned}
$$

Table of Contents

Non-linear models

We incorporate conditional dependence and deep non-linear functions into the statistical model, $q(z|x;\phi) = q(z;\phi(x;\psi))$, through the parameters

$$\phi(X;\psi) = \sigma(b_r + A_r\sigma(\cdots\sigma(b_1 + A_1X)))$$
$$\psi = \{b_i, A_i | i = 1,\ldots,r\}$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is a non-linear function applied element-wise.

We can do the same for $f_{X|Z}(x|z;\theta) = f_{X|Z}(x;\theta_\lambda(z))$:

$$\theta(Z;\lambda) = \sigma(c_r + D_r\sigma(\cdots\sigma(c_1 + D_1Z)))$$
$$\lambda = \{c_i, D_i | i = 1,\ldots,r\}$$

U.S. ARMY RDECOM

Non-linear models

ARL

Example: supervised learning

Given iid observations $\{(X_1, Z_1), \ldots, (X_N, Z_N)\}$ and an approximate distribution,

$$Z_1 | X_1 \sim \mathcal{N}(z; \mu(X; A_1, A_2, b_1, b_2), I)$$
$$\mu(X_1) = b_2 + A_2 \sigma(b_1 + A_1 X_1)$$

the maximum likelihood estimate is

$$A_1, A_2, b_1, b_2 = \underset{A_1, A_2, b_1, b_2}{\arg\min} \frac{1}{2N} \sum_{i=1}^{N} \| Z_i - \mu(X; A_1, A_2, b_1, b_2) \|^2$$

U.S. ARMY RDECOM ARL

Table of Contents

U.S. ARMY RDECOM

Example

ARL

Variational inference objective

$$\phi^*, \theta^* = \arg\min_{\phi,\theta} -\mathbb{E}_{x\sim f_X, w\sim f_W} \Big[\log f_{X|Z}(x;\theta) + \log(f_Z \circ g_\phi)(w) + \log J_w g_\phi(w)\Big]$$

For an observed random variable $X \in \mathbb{R}^n$, consider a latent variable model

$$X|Z \sim \mathcal{N}(x; \mu, \gamma I)$$
$$Z \sim \mathcal{N}(z; 0, I)$$

with $Z \in \mathbb{R}^m$,

and an approximate inference model, q,

$$Z|X \sim \mathcal{N}(z|\nu, \Sigma)$$
$$\Sigma^{\frac{1}{2}} = diag(\sigma)$$

Example

Variational inference objective

$$\phi^*, \theta^* = \arg\min_{\phi,\theta} -\mathbb{E}_{x\sim f_X, w\sim f_W} \left[\log f_{X|Z}(x;\theta) + \log(f_Z \circ g_\phi)(w) + \log J_w g_\phi(w)\right]$$

For an observed random variable $X \in \mathbb{R}^n$, consider a latent variable model

$$X|Z \sim \mathcal{N}(x; \mu, \gamma I)$$
$$Z \sim \mathcal{N}(z; 0, I)$$

with $Z \in \mathbb{R}^m$,

and an approximate inference model, $(f_W \circ g^{-1})(z) J_z g^{-1}(z)$,

$$W \sim \mathcal{N}(w; 0, I)$$
$$g(W) = \nu + \Sigma^{\frac{1}{2}} W$$
$$J_w g(W) = |\Sigma^{\frac{1}{2}}| = \prod_{j=1}^{m} \sigma_j$$

Example

Variational inference objective

$$\nu^*, \sigma^*, \mu^* = \arg\min_{\nu,\sigma,\mu} -\mathbb{E}_{x \sim f_X, w \sim f_W} \Big[\log f_{X|Z}(x; \mu) + \log(f_Z \circ g_{\nu,\sigma})(w) + \log J_w g_{\nu,\sigma}(w)\Big]$$

For an observed random variable $X \in \mathbb{R}^n$, consider a latent variable model

$$X|Z \sim \mathcal{N}(x; \mu, \gamma I)$$
$$Z \sim \mathcal{N}(z; 0, I)$$

with $Z \in \mathbb{R}^m$,

and an approximate inference model, $(f_W \circ g^{-1})(z) J_z g^{-1}(z)$,

$$W \sim \mathcal{N}(w; 0, I)$$
$$g(W) = \nu + \Sigma^{\frac{1}{2}} W$$
$$J_w g(W) = |\Sigma^{\frac{1}{2}}| = \prod_{j=1}^{m} \sigma_j$$

U.S. ARMY RDECOM

Example

ARL

$$
\begin{aligned}
&-\mathbb{E}_{x\sim f_X, w\sim f_W}\left[\log f_{X|Z}(x;\mu) + \log(f_Z \circ g_{\nu,\sigma})(w) + \log J_w g_{\nu,\sigma}(w)\right] \\
&= -\mathbb{E}_{x\sim f_X, w\sim f_W}\left[\log f_{X|Z}(x;\mu) + \log f_Z(\nu + \Sigma^{\frac{1}{2}} w) + \log \prod_{j=1}^{m} \sigma_j\right] \\
&= \mathbb{E}_{x\sim f_X}\left[\mathbb{E}_{w\sim f_W}\left[\frac{1}{2\gamma}\|x-\mu\|^2 + \frac{1}{2}\|\nu + \Sigma^{\frac{1}{2}} w\|^2\right] - \sum_{j=1}^{m} \log \sigma_j\right] + C \\
&= \mathbb{E}_{x\sim f_X}\left[\frac{1}{2\gamma}\mathbb{E}_{w\sim f_W}\|x-\mu\|^2 + \frac{1}{2}\|\nu\|^2 + \frac{1}{2}\|\sigma\|^2 - \sum_{j=1}^{m} \log \sigma_j\right] + C
\end{aligned}
$$

U.S. ARMY RDECOM

Example

ARL

Variational inference objective

$$\nu^*, \sigma^*, \mu^* = \operatorname*{arg\,min}_{\nu,\sigma,\mu} -\mathbb{E}_{x\sim f_X, w\sim f_W}\left[\frac{1}{2\gamma}\mathbb{E}_{w\sim f_W}\|x-\mu\|^2 + \frac{1}{2}\|\nu\|^2 + \frac{1}{2}\|\sigma\|^2 - \sum_{j=1}^{m}\log \sigma_j\right]$$

U.S. ARMY RDECOM

Example

ARL

Now, we can include non-linear statistical models, $\mu = \mu_\lambda(z)$, $\nu = \nu_\psi(x)$, and $\sigma = \sigma_\psi(x)$.

Variational inference objective

$$\psi^*, \lambda^* = \arg\min_{\psi,\lambda} \mathbb{E}_{x \sim f_X} \left[\frac{1}{2\gamma} \mathbb{E}_{w \sim f_W} \left\| x - \mu_\lambda \left(\nu_\psi(x) + \Sigma_\psi^{\frac{1}{2}}(x) w \right) \right\|^2 + \frac{1}{2} \|\nu_\psi(x)\|^2 + \frac{1}{2} \|\sigma_\psi(x)\|^2 - \sum_{j=1}^{m} \log \sigma_{\psi,j}(x) \right]$$

Example

Finally, we can estimate gradients by finite sampling:

$$\nabla_{\psi,\lambda}\mathbb{E}_{x\sim f_X}\left[\frac{1}{2\gamma}\mathbb{E}_{w\sim f_W}\left\|x-\mu_\lambda\left(\nu_\psi(x)+\Sigma_\psi^{\frac{1}{2}}(x)w\right)\right\|^2 + \frac{1}{2}\|\nu_\psi(x)\|^2+\frac{1}{2}\|\sigma_\psi(x)\|^2-\sum_{j=1}^{m}\log\sigma_{\psi,j}(x)\right]$$

$$\approx \frac{1}{NM}\sum_{i=1}^{N}\sum_{i=1}^{M}\nabla_{\psi,\lambda}\left[\frac{1}{2\gamma}\left\|x_i-\mu_\lambda\left(\nu_\psi(x_i)+\Sigma_\psi^{\frac{1}{2}}(x_i)w_j\right)\right\|^2 + \frac{1}{2}\|\nu_\psi(x_i)\|^2+\frac{1}{2}\|\sigma_\psi(x_i)\|^2-\sum_{j=1}^{m}\log\sigma_{\psi,j}(x_i)\right]$$

U.S. ARMY RDECOM

Example

ARL

- MNIST data set
- $60,000$ training examples of handwritten digits
- $X \in \mathbb{R}^{28\times 28}$, $Z \in \mathbb{R}^{10}$
- Stochastic gradient descent (Adam) $(\alpha_K = \frac{0.005}{1+K})$
- NVIDIA Quadro M3000M (4GB)
- TensorFlow

Figure: t-distributed stochastic neighbor embedding generated in TensorFlow

Table of Contents

Significance

How do I evaluate $f_X(x)$? (likelihood)

- $f_Z\left(\mathbb{E}_{z|x\sim q}\left[Z|X=x\right]\right)$

How do I generate realizations of X? (sampling)

- $f_{X|Z}(x|z;\theta),\ \ z \sim f_Z(z)$

How do I generate realizations of X like x_i? (characterizing)

- $f_{X|Z}(x|Z=\mathbb{E}_{z|x\sim q}\left[Z|X=x_i\right]+\delta;\theta)$

References

[1] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[2] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," *arXiv preprint arXiv:1401.4082*, 2014.

[3] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *arXiv preprint arXiv:1601.00670*, 2016.

[4] G. Casella and R. L. Berger, *Statistical inference*. Duxbury Pacific Grove, CA, 2002, vol. 2.

Autoencoder formulation

$$
\begin{aligned}
&\arg\min_{\phi,\theta} \mathbb{E}_{x\sim f_X} \mathcal{D}_{KL}(q(z|x;\phi)\|p(z|x;\theta)) + \mathcal{D}_{KL}(f_X(x)\|p(x;\theta)) \\
&= \arg\min_{\phi,\theta} \mathbb{E}_{x\sim f_X} \Big[\mathcal{D}_{KL}(q(z|x;\phi)\|p(x|z;\theta)p(z)) \\
&\qquad\qquad + \log p(x;\theta) + \log \frac{f_X(x)}{p(x;\theta)}\Big] \\
&= \arg\min_{\phi,\theta} \mathbb{E}_{x\sim f_X} \left[\mathcal{D}_{KL}(q(z|x;\phi)\|p(x|z;\theta)p(z)) + \log f_X(x)\right] \\
&= \arg\min_{\phi,\theta} \mathbb{E}_{x\sim f_X} \mathcal{D}_{KL}(q(z|x;\phi)\|p(x|z;\theta)p(z))
\end{aligned}
$$