DL HW4

Mehdi Jamalkhah

September 2024

1 Variational Inference

1.1

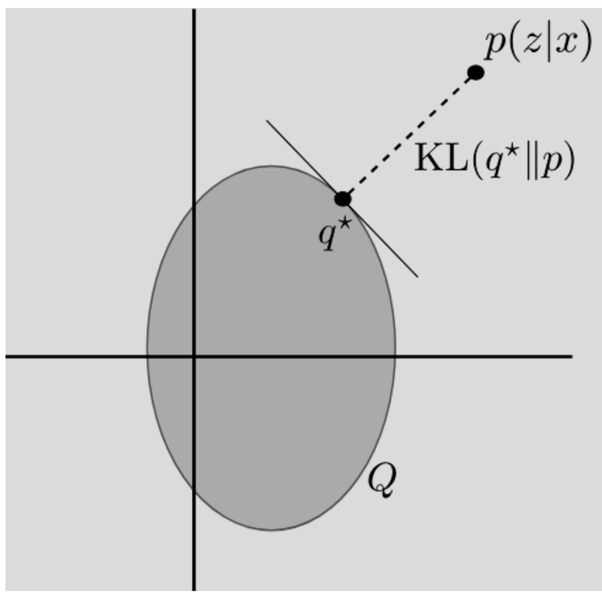

Figure 1: Abstraction of the variational inference problem. We seek to find an approximate posterior distribution that minimizes the KL divergence. Figure credit: Jeffrey Regier

$$
\begin{aligned}
q^* &= \underset{q \in Q}{\text{argmin}}\ \mathrm{KL}(q(z)||p(z|x)) \\
&= \underset{q \in Q}{\text{argmin}}\ \mathbb{E}_q[log(q(z)) - log(p(z|x))] \\
&= \underset{q \in Q}{\text{argmin}}\ \mathbb{E}_q[log(q(z)) - log(p(z,x))] + log(p(x)) \\
&= \underset{q \in Q}{\text{argmin}}\ \mathbb{E}_q[log(q(z)) - log(p(z,x))] \\
&= \underset{q \in Q}{\text{argmin}}\ \mathbb{E}_q[log(q(z)) - log(p(x|z) - log(p(z)))] \\
&= \underset{q \in Q}{\text{argmax}}\ \mathbb{E}_q[log(p(x|z))] - \mathrm{KL}(q(z)||p(z)) = \mathrm{ELBO}
\end{aligned}
$$

$$
\begin{aligned}
\log\, p(x) &= \mathbb{E}_q[\log\, p(x)] \\
&= \mathbb{E}_q\left[\log\, \frac{p(x|z)p(z)}{p(z|x)}\right] \\
&= \mathbb{E}_q\left[\log\, \left(\frac{p(x|z)p(z)}{p(z|x)} \times \frac{q(z)}{q(z)}\right)\right] \\
&= \mathbb{E}_q[\log\, p(x|z)] - \mathbb{E}_q\left[\log\, \frac{q(z)}{p(z)}\right] + \mathbb{E}_q\left[\log\, \frac{q(z)}{p(z|x)}\right] \\
&= \mathbb{E}_q[\log\, p(x|z)] - \mathrm{KL}(q(z)||p(z)) + \mathrm{KL}(q(z)||p(z|x)) \\
&= \mathrm{ELBO} + \mathrm{KL}(q(z)||p(z|x))
\end{aligned}
$$

1.2

$$
\psi, \theta = \psi, \theta - \eta \nabla_{\psi,\theta} \sum_{i=1}^{N} \mathrm{ELBO}(q_{\psi_i}, x_i, \theta)
$$

1.3

Stochastic VI

$$
\psi, \theta = \psi, \theta - \eta \frac{N}{B} \nabla_{\psi,\theta} \sum_{i=1}^{B} \mathrm{ELBO}(q_{\psi_i}, x_i, \theta)
$$

Amortized VI

$$
\phi, \theta = \phi, \theta - \eta \frac{N}{B} \nabla_{\phi,\theta} \sum_{i=1}^{B} \mathrm{ELBO}(q_{\phi}(.|x_i), x_i, \theta)
$$

1.4

$$
\begin{aligned}
p_\theta(x|z) &= \frac{1}{(\sqrt{2\pi}\sigma)^D} e^{-\frac{1}{2\sigma^2}||x - f_\theta(z)||^2} \\
\log\, p_\theta(x|z) &= C - \frac{1}{2\sigma^2}||x - f_\theta(z)||^2
\end{aligned}
$$

where D is the dimentionality of the x.

$$
\begin{aligned}
\mathrm{KL}(q(z)||p(z)) &= \frac{1}{2}(tr(\mathrm{diag}(\sigma_\phi^2(x))) + \mu_\phi(x)^T \mu_\phi(x) - k - \log\, \det\, \mathrm{diag}(\sigma_\phi^2(x))) \\
&= \frac{1}{2}\left(\mathrm{sum}(\sigma_\phi^2(x)) + ||\mu_\phi(x)||^2 - k - \log\, \mathrm{mul}(\sigma_\phi^2(x))\right)
\end{aligned}
$$

where k is the dimentionality of the z.

$$
\begin{aligned}
L &= -\mathbb{E}_q[\log\, p(x|z)] + \mathrm{KL}(q(z)||p(z)) \\
&= \frac{1}{2\sigma^2}||x - f_\theta(z)||^2 + \frac{1}{2}\left(\mathrm{sum}(\sigma_\phi^2(x)) + ||\mu_\phi(x)||^2 - k - \log\, \mathrm{mul}(\sigma_\phi^2(x))\right)
\end{aligned}
$$

1.5

$$
\begin{aligned}
\mathrm{ELBO} &= -\mathrm{KL}(q(z)||p(z|x)) \\
&= -\mathbb{E}_q[log(q(z)) - log(p(z|x))] \\
&= -\mathbb{E}_q[log(q(z)) - log(p(z,x))] + log(p(x)) \\
&= -\mathbb{E}_q[log(q(z)) - log(p(z,x))] \\
&= -\mathbb{E}_{\prod_i q_i(z_i)}[log(\prod_i q_i(z_i)) - log(p(z,x))] \\
&= -\mathbb{E}_{\prod_i q_i(z_i)}[\sum_i log\ q_i(z_i) - log(p(z,x))] \\
&= -\int \prod_i q_i(z_i) \left[\sum_i log\ q_i(z_i) + log(p(z,x))\right] dz \\
&= -\int \left[\prod_i q_i(z_i)\right] \sum_i log\ q_i(z_i) dz + \int \left[\prod_i q_i(z_i)\right] log(p(z,x)) dz \\
&= -\int q_j(z_j) \int \left[\prod_{i\neq j} q_i(z_i)\right] \sum_i log\ q_i(z_i) dz + \int q_j(z_j) \int \left[\prod_{i\neq j} q_i(z_i)\right] log(p(z,x)) dz \\
&= -\int q_j(z_j) log\ q_j(z_j) \int \prod_{i\neq j} q_i(z_i) dz - \int q_j(z_j) \int \left[\prod_{i\neq j} q_i(z_i)\right] \sum_{i\neq j} log\ q_i(z_i) dz + \int q_j(z_j) \mathbb{E}_{z_{-j}}[log(p(z,x))] dz_j \\
&= -\int q_j(z_j) log\ q_j(z_j) dz_j - \int \left[\prod_{i\neq j} q_i(z_i)\right] \sum_{i\neq j} log\ q_i(z_i) dz_{-j} + \int q_j(z_j) \mathbb{E}_{z_{-j}}[log(p(z,x))] dz_j \\
&= -\int q_j(z_j) log\ q_j(z_j) dz_j + \int q_j(z_j) \mathbb{E}_{z_{-j}}[log(p(z,x))] dz_j + C \\
&= -\int q_j(z_j) \left[log\ q_j(z_j) - \mathbb{E}_{z_{-j}}[log(p(z,x))]\right] dz_j + C
\end{aligned}
$$

$$
\begin{aligned}
Lagrangian &= -\int q_j(z_j) \left[log\ q_j(z_j) - q_j(z_j)\mathbb{E}_{z_{-j}}[log(p(z,x))]\right] dz_j - \sum_i \lambda_i \int q_i(z_i) dz_i \\
\frac{\partial Lagrangian}{\partial q_j} &= \mathbb{E}_{z_{-j}}[log(p(z,x))] - log\ q_j(z_j) - 1 - \lambda_j = 0 \\
log\ q_j(z_j) &= \mathbb{E}_{z_{-j}}[log(p(z,x))] + C \\
q_j(z_j) &\propto exp[\mathbb{E}_{z_{-j}}[log\ p(z,x)]]
\end{aligned}
$$

2 Diffusion Models

2.1

$$
\begin{aligned}
q(z_t|x) &= \mathcal{N}(a_t x, \sigma_t^2 I) \\
&= a_t x + \mathcal{N}(0, \sigma_t^2 I) \\
&= Dist(0, a_t^2) + \mathcal{N}(0, \sigma_t^2 I) \\
\mathrm{Var}_q &= a_t^2 + \sigma_t^2 = 1 \\
a_t &= \sqrt{1 - \sigma_t^2}
\end{aligned}
$$

2.2

$$
q(z_s, z_t|x) = q(z_s|z_t, x)q(z_t|x) = q(z_s|z_t)q(z_t|x)
$$

2.3

$$
\begin{aligned}
z_s &\sim \mathcal{N}(a_s x, \sigma_s^2 I) \\
z_t &\sim \mathcal{N}(a_t x, \sigma_t^2 I) \\
\frac{a_t}{a_s} z_s &\sim \mathcal{N}(a_t x, \frac{a_t^2}{a_s^2} \sigma_s^2 I) \\
\mathcal{N}(0, (\sigma_t^2 - \frac{a_t^2}{a_s^2} \sigma_s^2) I) + \frac{a_t}{a_s} z_s &\sim \mathcal{N}(a_t x, \sigma_t^2 I) \\
\mathcal{N}(0, (\sigma_t^2 - \frac{a_t^2}{a_s^2} \sigma_s^2) I) + \frac{a_t}{a_s} z_s &\sim z_t \\
\mathcal{N}(\frac{a_t}{a_s} z_s, (\sigma_t^2 - \frac{a_t^2}{a_s^2} \sigma_s^2) I) &\sim z_t \\
\mathcal{N}(a_{t|s} z_s, \sigma_{t|s}^2 I) &\sim z_t
\end{aligned}
$$

where,

$$
a_{t|s} = \frac{a_t}{a_s},\ \sigma_{t|s}^2 = (\sigma_t^2 - \frac{a_t^2}{a_s^2} \sigma_s^2)
$$

2.4

$$
\begin{aligned}
q(z_s|z_t,x) &\propto q(z_t|z_s)q(z_s|x) \\
&\propto \mathcal{N}(a_{t|s}z_s, \sigma_{t|s}^2 I)\mathcal{N}(a_s x, \sigma_s^2 I) \\
&\propto exp(-\frac{1}{2}(\frac{z_t - a_{t|s}z_s}{\sigma_{t|s}})^2 - \frac{1}{2}(\frac{z_s - a_s x}{\sigma_s})^2) \\
&\propto exp(-\frac{1}{2}\frac{(\sigma_s^2 a_{t|s}^2 + \sigma_{t|s}^2)z_s^2 + (z_t a_{t|s}\sigma_s^2 + a_s x \sigma_{t|s}^2)z_s}{\sigma_{t|s}^2\sigma_s^2}) \\
&\propto exp(-\frac{1}{2}\frac{(\sigma_t^2)z_s^2 + (z_t a_{t|s}\sigma_s^2 + a_s x \sigma_{t|s}^2)z_s}{\sigma_{t|s}^2\sigma_s^2}) \\
&\propto exp(-\frac{1}{2}\frac{z_s^2 + \frac{z_t a_{t|s}\sigma_s^2 + a_s x \sigma_{t|s}^2}{\sigma_t^2}z_s}{\frac{\sigma_{t|s}^2\sigma_s^2}{\sigma_t^2}}) \\
&\propto exp(-\frac{1}{2}(\frac{z_s - \mu_Q(z_t,x;s,t)}{\sigma_Q^2(s,t)})^2) \\
&\propto \mathcal{N}(\mu_Q(z_t,x;s,t), \sigma_Q^2(s,t))
\end{aligned}
$$

where,

$$
\begin{aligned}
\mu_Q(z_t,x;s,t) &= \frac{(a_{t|s}\sigma_s^2)}{\sigma_t^2}z_t + \frac{(a_s\sigma_{t|s}^2)}{\sigma_t^2}x \\
\sigma_Q^2(s,t) &= \frac{\sigma_{t|s}^2\sigma_s^2}{\sigma_t^2}
\end{aligned}
$$

2.5

$$
\begin{aligned}
D_{\mathrm{KL}}(q(z_s|z_t,x)||p_\theta(z_s|z_t)) &= \frac{1}{2}(d + \frac{1}{\sigma_Q^2(s,t)}||\mu_Q(z_t,x;s,t) - \mu_\theta(z_t;s,t)||_2^2 - d + \log(1)) \\
&= \frac{1}{2\sigma_Q^2(s,t)}||\mu_Q(z_t,x;s,t) - \mu_\theta(z_t;s,t)||_2^2
\end{aligned}
$$

2.6

$$
\begin{aligned}
D_{\mathrm{KL}}(q(z_s|z_t,x)||p_\theta(z_s|z_t)) &= \frac{1}{2\sigma_Q(s,t)}||\mu_Q(z_t,x;s,t) - \mu_\theta(z_t;s,t)||_2^2 \\
&= \frac{1}{2\sigma_Q(s,t)}||\frac{(a_{t|s\sigma_s^2})}{\sigma_t^2}z_t + \frac{(a_s\sigma_{t|s}^2)}{\sigma_t^2}x - \frac{(a_{t|s\sigma_s^2})}{\sigma_t^2}z_t + \frac{(a_s\sigma_{t|s}^2)}{\sigma_t^2}\hat{x}_\theta(z_t,t)||_2^2 \\
&= \frac{(a_s^2\sigma_{t|s}^4)}{2\sigma_t^4\sigma_Q^2(s,t)}||x - \hat{x}_\theta(z_t,t)||_2^2 \\
&= \frac{1}{2}\gamma||x - \hat{x}_\theta(z_t,t)||_2^2
\end{aligned}
$$

where,

$$
\begin{aligned}
\gamma &= \frac{a_s^2\sigma_{t|s}^4}{\sigma_t^4\sigma_Q^2(s,t)} \\
&= \frac{a_s^2\sigma_{t|s}^2}{\sigma_t^2\sigma_s^2} \\
&= \frac{a_s^2(\sigma_t^2 - \frac{a_t^2}{a_s^2}\sigma_s^2)}{\sigma_t^2\sigma_s^2} \\
&= \frac{a_s^{@}\sigma_t^2 - a_t^2\sigma_s^2}{\sigma_t^2\sigma_s^2} \\
&= \frac{a_s^2}{\sigma_s^2} - \frac{a_t^2}{\sigma_t^2} \\
&= \text{SNR}(s) - \text{SNR}(t)
\end{aligned}
$$

2.7

$$
\begin{aligned}
L_\infty(x) &= \lim_{T\to\infty} L_T(x) \\
&= \lim_{T\to\infty} \frac{T}{2}\mathbb{E}[(SNR(s(i)) - SNR(t(i)))||x - \hat{x}_\theta(z_{t(i)}; t(i))||_2^2] \\
&= \lim_{T\to\infty} \frac{T}{2}\mathbb{E}[(SNR(t - \frac{1}{T}) - SNR(t))||x - \hat{x}_\theta(z_{t(i)}; t(i))||_2^2] \\
&= -\frac{1}{2}\mathbb{E}[\lim_{T\to\infty} T(SNR(t) - SNR(t - \frac{1}{T}))||x - \hat{x}_\theta(z_{t(i)}; t(i))||_2^2] \\
&= -\frac{1}{2}\mathbb{E}[SNR'(t)||x - \hat{x}_\theta(z_{t(i)}; t(i))||_2^2]
\end{aligned}
$$

3 Score Matching

3.1

$$
\begin{aligned}
x_{t+1} &= x_t + \delta\nabla_x \log p(x_t) \\
\text{No Update: } x_t &= x_t + \delta\nabla_x \log p(x_t) \\
\nabla_x \log p(x_t) &= 0
\end{aligned}
$$

At point x_t, $p(x)$ reaches a peak because the gradient is zero at this location.

Maximum likelihood typically identifies the global maximum, whereas this method may converge on a local maximum.

3.2

There are two reasons for the presence of a noise term. First, it helps the algorithm escape local maxima, especially if the starting point is close to one. Second, the noise encourages the discovery of flatter maxima.

3.3

$$\nabla_x \text{log} q(x) = \frac{q'(x)}{q(x)} = \frac{\frac{1}{M}\sum_{i=1}^{M} \frac{-2(x-x^{(i)})}{2\sigma^2} K(x|x^{(i)})}{\frac{1}{M}\sum_{i=1}^{M} K(x|x^{(i)})}$$
$$= \frac{q'(x)}{q(x)} = \frac{\sum_{i=1}^{M} \frac{1}{\sigma^2}(x^{(i)} - x) K(x|x^{(i)})}{\sum_{i=1}^{M} K(x|x^{(i)})}$$

3.4

If we lack sufficient data in a particular region, this method will not produce an accurate density estimate for that area, which may prevent us from correctly identifying the peak of the density.

3.5

$$J_1(\theta) = \mathbb{E}_{q(x)}[\frac{1}{2}||s_\theta(x)||^2] - g(\theta) + C$$

$$\begin{aligned}
g(\theta) &= \mathbb{E}_{q(x)}[\left\langle s(x), \frac{\partial \text{log } q(x)}{\partial x} \right\rangle] \\
&= \int_x q(x) \left\langle s(x), \frac{\partial \text{log } q(x)}{\partial x} \right\rangle dx \\
&= \int_x q(x) \left\langle s(x), \frac{\frac{\partial q(x)}{\partial x}}{q(x)} \right\rangle dx \\
&= \int_x \left\langle s(x), \frac{\partial q(x)}{\partial x} \right\rangle dx \\
&= \int_x \left\langle s(x), \frac{\partial}{\partial x} \int_{x_0} q_0(x_0) q(x|x_0) dx_0 \right\rangle dx \\
&= \int_x \left\langle s(x), \int_{x_0} q_0(x_0) \frac{\partial q(x|x_0)}{\partial x} dx_0 \right\rangle dx \\
&= \int_x \left\langle s(x), \int_{x_0} q_0(x_0) q(x|x_0) \frac{\partial \text{log } q(x|x_0)}{\partial x} dx_0 \right\rangle dx \\
&= \int_x \int_{x_0} q_0(x_0) q(x|x_0) \left\langle s(x), \frac{\partial \text{log } q(x|x_0)}{\partial x} \right\rangle dx dx_0 \\
&= \int_x \int_{x_0} q(x, x_0) \left\langle s(x), \frac{\partial \text{log } q(x|x_0)}{\partial x} \right\rangle dx dx_0 \\
&= \mathbb{E}_{q(x,x_0)} \left[\left\langle s(x), \frac{\partial \text{log } q(x|x_0)}{\partial x} \right\rangle \right]
\end{aligned}$$

So,

$$J_1(\theta) = J_2(\theta) + C$$

3.6

$$
\begin{aligned}
J_2(\theta) &= \mathbb{E}_{q(x,x_0)}[\frac{1}{2}||s_\theta(x) - \nabla_x \log\, q(x|x_0)||^2] \\
&= \mathbb{E}_{q(x,x_0)}[\frac{1}{2}||s_\theta(x) - \frac{1}{\sigma^2}(x - x_0)||^2]
\end{aligned}
$$

Algorithm 1 Diffusion Training

for $i \leftarrow 1$ to m **do**
    $x_1 = x^{(i)}$
    for $t \leftarrow 1$ to T **do**
        $x_{t+1} = x_t + \lambda s_\theta(x_t) + \sqrt{2\lambda}\epsilon$
        $\theta = \theta - \eta \nabla_\theta J_2(\theta)$
    end for
end for

Algorithm 2 Diffusion Evaluating

$x_1 \sim \mathcal{N}(0, I)$
for $t \leftarrow 1$ to T **do**
    $x_{t+1} = x_t + \lambda s_\theta(x_t) + \sqrt{2\lambda}\epsilon$
    $\theta = \theta - \eta \nabla_\theta J_2(\theta)$
end for
return x_T