$$\sum_i \log p(x;\theta) = \sum_i \log\left[\sum_z p(x,z;\theta)\right]$$

when z is high-dim e.g. $\{0,1\}^{30}$, the summation/integral is intractable

Need approximation

· **Naive Monte Carlo**

$$P_\theta(x) = \sum_{z\in Val(z)} P_\theta(x,z) = |Val(z)| \sum_{z\in Val(z)} \frac{1}{|Val(z)|} P_\theta(x,z)$$

$$= |Val(z)| \cdot E_{z\sim \text{uniform}(Val(z))}\left[P_\theta(x,z)\right]$$

$$\approx |Val(z)| \cdot \frac{1}{k}\sum_{z^{(k)}} P_\theta(x,z)$$

randomly sample $z^{(1)} \dots z^{(k)} \sim \text{uniform}(Val(z))$

approximate $\sum_{z\in Val(z)} P_\theta(x,z) \approx |Val(z)| \cdot \frac{1}{k}\sum_{z^{(k)}} P_\theta(x,z^{(k)})$

problems: most $P_\theta(x,z)$ will be very low (most image completion does not make sense)

some $P_\theta(x,z)$ is very high, but is rarely sampled by uniform sampling

Need a clever way to select $z^{(j)}$ to reduce the variance of the estimator

· **Importance Sampling**

$$p_\theta(x) = \sum_{z\in Val(z)} P_\theta(x,z) = \sum_{z\in Val(z)} \frac{q(z)}{q(z)}\cdot P_\theta(x,z)$$

$$= E_{z\sim q}\left[\frac{P_\theta(x,z)}{q(z)}\right]$$

$$\approx \frac{1}{k}\sum_k \frac{P_\theta(x,z^{(k)})}{q(z^{(k)})}$$

randomly sample $z^{(1)} \dots z^{(k)} \sim q(z)$

approximate $P_\theta(x) \approx \frac{1}{k}\sum_k \frac{P_\theta(x,z^{(k)})}{q(z^{(k)})}$

$$E_{z\sim q}\left[\frac{1}{k}\sum_{j=1}^{k}\frac{P_\theta(x,z^{(j)})}{q(z^{(j)})}\right] = E_{z\sim q}\left[\frac{P_\theta(x,z)}{q(z)}\right]$$

$$= \sum_z q(z)\frac{P_\theta(x,z)}{q(z)}$$

$$= P_\theta(x)$$

unbiased estimator

$$Var_{z\sim q}\left[\frac{1}{k}\sum_{j=1}^{k}\frac{P_\theta(x,z^{(j)})}{q(z^{(j)})}\right] = \frac{1}{M} Var_{z\sim q}\left(\frac{P_\theta(x,z)}{q(z)}\right)$$

$$= \frac{1}{M}\left(E_{z\sim q}\left[\left(\frac{P_\theta(x,z)}{q(z)}\right)^2\right] - E_{z\sim q}\left[\frac{P_\theta(x,z)}{q(z)}\right]^2\right)$$

$$= \frac{1}{M}\left(E_{z\sim q}\left[\left(\frac{P_\theta(x,z)}{q(z)}\right)^2\right] - P_\theta(x)^2\right)$$

$$\log(P_\theta(x)) \approx \log\left(\frac{1}{k}\sum_k \frac{P_\theta(x,z^{(k)})}{q(z^{(k)})}\right)$$

$$\overset{k=1}{=} \log\left(\frac{P_\theta(x,z^{(k)})}{q(z^{(k)})}\right)$$

when $\frac{P_\theta(x,z)}{q(z)} \approx P_\theta(x)$, the var is small

$q(z) = P_\theta(z|x)$

$$\log\left[E_{z\sim q}\left[\frac{P_\theta(x,z)}{q(z)}\right]\right] \neq E_{z\sim q}\left[\log\left(\frac{P_\theta(x,z)}{q(z)}\right)\right]$$

unless $\frac{P_\theta(x,z)}{q(z)}$ is constant $\forall z$

· **Variational Inference**

suppose $q(z)$ is any distribution over z

$$\log p(x;\theta) = \log\left[\sum_z q(z)\cdot\frac{p(x,z;\theta)}{q(z)}\right]$$

$$\geq \sum_z q(z)\cdot\log\frac{p(x,z;\theta)}{q(z)}$$

equality holds if $q(z) = p_\theta(z|x)$

$$= \sum_z q(z) \cdot \log P(x,z;\theta) - \sum_z q(z) \cdot \log q(z)$$

$$= \underbrace{\sum_z q(z) \cdot \log P(x,z;\theta) + H(q)}_{} \quad \text{lower bound}$$

$$\log P_\theta(x;\theta) = \sum_z q(z) \log P_\theta(x;\theta)$$

$$= \sum_z q(z) \left[\log P_\theta(x;\theta) + \log q(z) - \log q(z) \right]$$

$$= \sum_z q(z) \left[\log \frac{P_\theta(x,z;\theta)}{P_\theta(z|x;\theta)} + \log q(z) - \log q(z) \right]$$

$$= \sum_z q(z) \cdot \left[\log \frac{q(z)}{P_\theta(z|x;\theta)} + \log P_\theta(x,z;\theta) - \log q(z) \right]$$

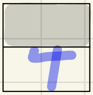
$$= KL\left(q(z) \,\|\, P_\theta(z|x;\theta)\right) + \text{lower bound}$$

the closer $q(z)$ is to $P_\theta(z|x;\theta)$, the closer the lower bound is to $\log P(x;\theta)$

variational Inference: choose a parametric form $q(z;\phi)$, and make $q(z;\phi)$ as close as possible to $P(z|x;\theta)$

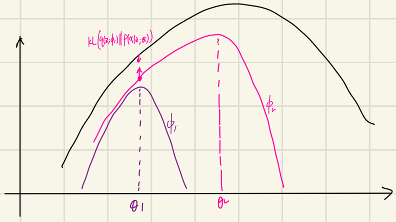
eg

Assume $P(x^{bottom}, z^{top};\theta)$ has high prob if (x,z) looks like real digitals



suppose we choose a binomial distribution

$$q(z^{top}) = \prod_i (\phi_i)^{z_i^{top}} (1-\phi_i)^{1-z_i^{top}}$$



lower bounder $\sum_z q(z;\phi) \log p(x,z;\theta) + H(q;\phi)$ $\phi=\phi_1$

lower bounder $\sum_z q(z;\phi) \log p(x,z;\theta) + H(q;\phi)$ $\phi=\phi_1$

- Lower Bound for entire dataset

$$\ell(\theta) = \sum_i \log P(x_i;\theta) = \sum_i \log \sum_{z_i} P(x_i,z_i;\theta)$$

$$= \sum_i \log \sum_{z_i} q_i(z_i;\phi_i) \cdot \frac{P(x_i,z_i;\theta)}{q_i(z_i;\phi_i)}$$

$$\begin{cases} \geq \sum_i \sum_{z_i} q_i(z_i;\phi_i) \cdot \log \frac{P(x_i,z_i;\theta)}{q_i(z_i;\phi_i)} \\ = \sum_i \left[\sum_{z_i} q_i(z_i;\phi_i) \cdot \log p(x_i,z_i;\theta) + H(q) + KL\left(q_i(z_i;\phi_i) \,\|\, P(z|x_i;\theta)\right) \right] \end{cases}$$

let $\theta^* = \arg\max_\theta \ell(\theta)$ $\quad (\hat\theta, \hat\phi) = \arg\max_{\theta,\phi} \sum_i \sum_{z_i} q_i(z_i;\phi_i) \cdot \log \frac{p(x_i, z_i;\theta)}{q_i(z_i;\phi_i)}$

$$\sum_i \sum_{z_i} q_i(z_i;\hat\phi_i) \cdot \log \frac{p(x_i, z_i;\hat\theta)}{q_i(z_i;\hat\phi_i)} \le \sum_i \log \sum_{z_i} p(x_i, z_i;\hat\theta) = \ell(\hat\theta) \le \ell(\theta^*)$$

for each x_i, need to find a ϕ_i such that $KL(q_i(z_i;\phi_i) \| p(z_i|x_i;\theta))$ is small
(can be expensive)

- Learning via Stochastic Variational Inference

the lower bound $\mathcal{L}(x_i;\theta,\phi_i) = \sum_{z_i} q_i(z_i;\phi_i) \cdot \log \frac{p(x_i,z_i;\theta)}{q_i(z_i;\phi_i)} = E_{z_i \sim q_i}\left[\log \frac{p(x_i,z_i;\theta)}{q_i(z_i;\phi_i)}\right]$

optimize $\sum_i \mathcal{L}(x_i;\theta,\phi_i)$ with stochastic gradient descent

initialize $\theta, \phi^1 \cdots \phi^m$
randomly sample x^i from D
optimize $\mathcal{L}(x_i;\theta,\phi_i)$ w.r.t. ϕ_i
  compute $\nabla_{\phi_i} \mathcal{L}(x_i;\theta,\phi_i)$
  apply gradient ascent until $\phi_i \approx \arg\max_{\phi_i} \mathcal{L}(x_i;\theta,\phi_i)$
compute $\nabla_\theta \mathcal{L}(x_i;\theta,\phi_i)$
update θ with gradient ascent, go to ②

$$\max_{\phi_i} \sum_{z_i} q_i(z_i;\phi_i) \cdot \log \frac{p(x_i,z_i;\theta)}{q_i(z_i;\phi_i)}$$
$$= \max_{\phi_i} \sum_{z_i} \left[q_i(z_i;\phi_i) \cdot \log \frac{p(z_i|x_i;\theta) \cdot p(x_i;\theta)}{q_i(z_i;\phi_i)} \right]$$
$$= \max_{\phi_i} \sum_{z_i} q_i(z_i;\phi_i) \log \frac{p(z_i|x_i;\theta)}{q_i(z_i;\phi_i)} - \sum_{z_i} q_i(z_i;\phi_i) \cdot p(x_i;\theta)$$
$$= \max_{\phi_i} \sum_{z_i} q_i(z_i;\phi_i) \log \frac{p(z_i|x_i;\theta)}{q_i(z_i;\phi_i)}$$
$$= \min_{\phi_i} KL(q_i(z_i;\phi_i) \| p(z_i|x_i;\theta))$$

to evaluate the bound, sample $z_i^{(1)} \cdots z_i^{(k)} \sim q_i(z_i;\phi_i)$

approximate $\mathcal{L}(x_i;\theta,\phi_i) = E_{z_i\sim q_i}\left[\log \frac{p(x_i,z_i;\theta)}{q_i(z_i;\phi_i)}\right] \approx \frac{1}{k}\sum_k \log \frac{p(x_i, z_i^{(k)};\theta)}{q_i(z_i^{(k)};\phi_i)}$

(Assume $q(\cdot)$ is easy to evaluate and sample from)

$$\nabla_\theta E_{z_i\sim q_i}\left[\log p(x_i,z_i;\theta) - \log q_i(z_i;\phi_i)\right] = E_{z_i\sim q_i}\left[\nabla_\theta \log p(x_i,z_i;\theta)\right] \approx \frac{1}{k}\sum_k \nabla_\theta \log p(x_i,z_i;\theta)$$

evaluating ∇_{ϕ_i} is hard

- Computing $\nabla_{\phi_i} E_{z_i\sim q_i}[-\log q_i(z_i;\phi_i)]$ when z_i is continuous

$$E_{z_i\sim q_i}[-\log q_i(z_i;\phi_i)] = -\int_{z_i} q_i(z_i;\phi_i) \log q_i(z_i;\phi_i)\, dz_i$$

suppose q_i is $\mathcal{N}(u_i, \sigma_i^2 I)$, to sample from q_i, there're 2 equivalent ways

① sample $z_i \sim \mathcal{N}(u_i, \sigma_i^2 I)$

② sample $t_i \sim \mathcal{N}(0, I)$, $z_i = u_i + t_i \cdot \sigma_i = g(t_i; u_i, \sigma_i)$