# 1) Set Up

- Assume that $x_{1:n}$ are observations and $z_{1:m}$ are hidden variables
  - May include the "parameters" ($\alpha$ are hyperparameters)
- Interested in **posterior** distribution

$$p(z \mid x, \alpha) = \underbrace{\frac{p(z, x \mid \alpha)}{\int_z p(z, x \mid \alpha)}}_{\text{links data \& model}}$$

# 2) Motivation

Difficult to compute posterior for many interesting models

Consider the Bayesian mixture of Gaussians,

1. Draw $\mu_k \sim \mathcal{N}(0, \tau^2)$ for $k = 1, \dots, K$
2. For $i = 1, \dots, n$
   (a) Draw $z_i \sim \text{Mult}(\pi)$;
   (b) Draw $x_i \sim \mathcal{N}(\mu_{z_i}, \sigma^2)$

Supressing fixed parameters, the posterior distribution is

EZ

$$p(\mu_{1:K}, z_{1:n} \mid x_{1:n}) = \frac{\prod_{k=1}^{K} \mathbb{P}(\mu_k) \prod_{i=1}^{n} \mathbb{P}(z_i) \mathbb{P}(x_i \mid z_i, \mu_{1:K})}{\int_{\mu_{1:K}} \sum_{z_{1:n}} \prod_{k=1}^{K} \mathbb{P}(\mu_k) \prod_{i=1}^{n} \mathbb{P}(z_i) \mathbb{P}(x_i \mid z_i, \mu_{1:K})}$$

NOT quite so simple

## 3) Main Idea

To pick a family of distributions over the latent variables with its own **variational parameters**,

$$q(z_{1:m} | \nu)$$

Then, find the setting of parameters that makes $q$ close to the posterior of interest.

Use $q$ w/ the fitted parameters as a proxy for the posterior
- i.e to form predictions about future data

Typically, true posterior is **NOT** in the variational family.

## 4) Kullback-Leibler Divergence

Measures closeness of two distributions

The KL divergence for variational inference is

$$KL(q \| p) = \mathbb{E}_q \left[ \log \frac{q(Z)}{\mathbb{P}(Z|x)} \right]$$

3 cases

1. $q$ is high and $p$ is high ⇒ :)
2. $q$ is high and $p$ is low ⇒ :| (price to pay)
3. $q$ is low ⇒ don't care

# 5) The evidence lower bound (ELBO)

Can't minimize KL divergence exactly, but we can minimize a function that is equal to it up to a constant. This is the ELBO

Recall Jensen's inequality as applied to probability distributions. When $f$ is concave,

$$f(\mathbb{E}(X)) \geq \mathbb{E}[f(X)]$$

Use Jensen's inequality on the log probability of the observations,

$$\log(\mathbb{P}(x)) = \log\left(\int_z \mathbb{P}(x,z)\right)$$

$$= \log\left(\int_z \mathbb{P}(x,z)\frac{q(z)}{q(z)}\right)$$

$$= \log\left(E_q\left[\frac{\mathbb{P}(x,Z)}{q(z)}\right]\right)$$

Note: this is Entropy!

This is the ELBO $\longrightarrow$ $\geq \mathbb{E}_q[\log \mathbb{P}(x,z)] - \mathbb{E}_q[\log q(Z)]$

Choose a family of variational distributions such that expectations are computable

Then, maximize ELBO to find the parameters that gives as tight a bound as possible on the marginal probability of $x$.

What does this have to do w/ KL divergence of posterior?

Fist, note that

$$\mathbb{P}(z|x) = \frac{\mathbb{P}(z,x)}{\mathbb{P}(x)}$$

Now use this in KL divergence,

$$KL\left(q(z) \,||\, \mathbb{P}(z|x)\right) = \mathbb{E}_q\left[\log \frac{q(z)}{\mathbb{P}(z|x)}\right]$$

(linearity of expectation) $= \mathbb{E}_q\left[\log q(z)\right] - \mathbb{E}_q\left[\log \mathbb{P}(z|x)\right]$

(Bayes Rule) $= \mathbb{E}_q\left[\log q(z)\right] - \left(\mathbb{E}_q\left[\log \mathbb{P}(z,x)\right] - \mathbb{E}_q\left[\log \mathbb{P}(x)\right]\right)$

$$= -\underbrace{\left(\mathbb{E}_q\left[\log \mathbb{P}(z|x)\right] - \mathbb{E}_q\left[\log q(z)\right]\right)}_{\text{ELBO}} + \underbrace{\log \mathbb{P}(x)}_{\text{Log Marginal probability of } x}$$

Thus minimizing KL divergence is same as maximizing ELBO

And, the difference between the ELBO and the KL-divergence is the log-normalizer which is what ELBO bounds.

# 6) Mean Field Variational Inference

Assume that the variational family factorizes

$$q(z_1, \ldots, z_m) = \prod_{i=1}^{m} q(z_i) \leftarrow$$ each variable is independent. (suppressing parameters $\nu_j$)

This is more general than it initially appears - the hidden variable can be grouped and the distribution of each group factorizes

Typically, this family does **NOT** contain the true posterior b/c the hidden variables are dependent

$$p(z_{1:m} \mid x_{1:n}) = p(x_{1:n}) \prod_{j=1}^{m} p(z_j \mid z_{1:(j-1)}, x_{1:n})$$