Foreword

Variational Inference (VI) first starts as a handy tool in Bayesian inference to approximate intractable posterior. Now, its usage goes beyond Bayesian inference, and we can see VI everywhere in classic learning, and deep learning-anywhere needs an approximation. After this lecture, we should:

- Understand the fundamentals of VI (why “variational” here?)
- See how models use VI (from Topic to Diffusion)
- Use VI for your problem

Motivation

The fundamental problem in Bayesian statistics is finding a model p that fits observed data. Given the model, we can even generate new data that looks like the observed one (hence, p is also a generative model). Formally, given a dataset of N data points $X = \{x_1, x_2, ..., x_N\}$, we want to search for a model p parameterised by θ s.t. $p(X|\theta)$ (also written as $p_\theta(X)$) is maximised. If the data is independent and identically distributed (i.i.d), the optimisation problem reads $\theta^* = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{N} p(X_i|\theta)$. For computing convenience, we may want to maximise the log-likelihood $\log p(X|\theta)$, leading to an equivalent problem:

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} \log p(x_i|\theta) \tag{1}$$

In addition to the parameters, we need to know the form of the model (Gaussian distribution, neural networks,...) to compute p. The fitting problem is easy for a simple model like Gaussian, where the optimal parameters to fit X are $\theta = \{\mu, \Sigma\}$ where $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ and $\Sigma = \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)(x_i - \mu)^T$ (see Fig. 1 below as an example).

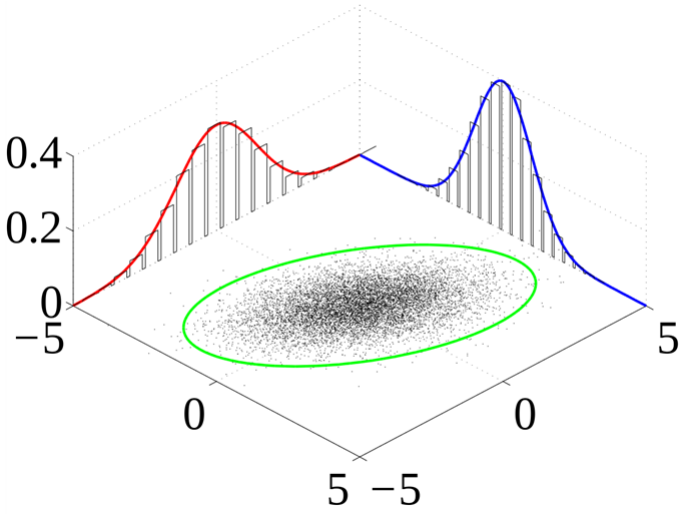

Figure 1: Fitting 2d points with 2d Gaussian distribution. Image source: Wikipedia

Fitting challenges. Fitting is only easy for simple models (e.g. Gaussian, low dimensional) while real data (images) are high dimensional, likely non Gaussian. Hence, to generate real data, we need to seek for complicated $p_\theta(x)$. *But how to make p complicated?*

Latent Variable Models (LVM)

One way to make complicated distribution is to assume the existence of latent variables in the generative model of the data. Let's start with one latent variable z . The generation procedure for each data point x is then:

- Someone (God) samples z from a prior $p(z)$ parameterized by α
- Then he/she generates x from a conditional distribution model parameterized by θ: $p(x|z;\theta)$ or $p_\theta(x|z)$. The probabilistic graphical model for this single-latent model is given in Fig. 2.

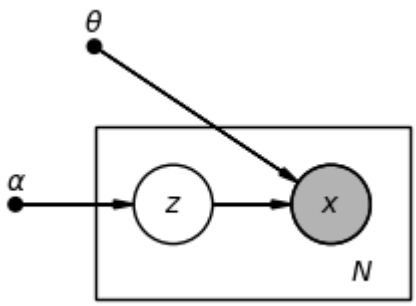

Figure 2: Propabilistic graphic model of a single-latent model

The distribution of the latent z is ok to be simple (e.g. some fixed Gaussian), but z can be high dimensional, which make the marginal $p(x) = \int p(x|z)p(z)dz$ complicated and hard to compute.

Problems in LVM

Fitting. As the marginal distribution $p(x)$ is complicated, the fitting problem (Eq. 1) also becomes more challenging. For example, consider a simple LVM, mixture of univariate Gaussians, which reads:

- Sample $\mu_k \sim \mathcal{N}(0, \tau^2)$ for $k = 1, 2, .., K$
- For each data point x_i:
  - Sample $z_i \sim \text{Mult}(\pi)$
  - Sample $x_i \sim \mathcal{N}(\mu_{z_i}, \sigma^2)$

also manifested by its PGM in Fig. 3.

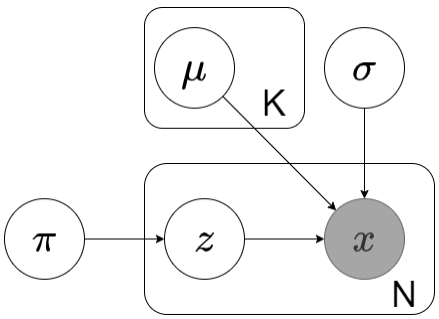

Figure 3: Propabilistic graphic model of a mixture of Gaussian model.

The latent variable z_i indicates which cluster the data point x_i belongs to. In this case, the marginal reads:

$$p(X) = \int_{\mu_{1:K}, z_{1:N}} p(x_{i:N}, z_{i:N}, \mu_{1:K}) \tag{2}$$

$$= \int_{\mu_{1:K}} \sum_{z_{1:N}} \prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{N} p(x_i|z_i, \mu_{1:K}, \sigma) p(z_i|\pi) \tag{3}$$

Although it is possible to find the closed-form for products of the conditional and integrals over $\mu_{1:K}$, there are N^K ways to assign N points to any of K clusters, and thus, N^K terms needed to compute. It is intractable as K and N increase. Therefore, for LVM, in general, it is better not to directly optimize the likelihood through computing the marginal. Instead, we can rewrite the log-likelihood as $\sum_{i=1}^{N} \log \int p_\theta(x_i|z_i)p(z)dz = \sum_{i=1}^{N} \log \mathbb{E}_p[p_\theta(x_i|z)]$ and use MCMC sampling to estimate the expectation. However, this straightforward solution is undesirable due to:

1. Sampling from the uninformative prior $p(z)$ with initially random θ mostly leads to $p_\theta(x_i|z) \sim 0$, which makes learning with gradient extremely slow.

2. Optimising $\log$ of expectation leads to biased MCMC approximation, which makes learning with gradients cannot correctly converge.

These two issues can be fixed with the help of variational inference, as we will see later.

Inference. In LVM, another vital problem besides fitting is inference, where we want to infer the latent variable given values of the observables, a.k.a, finding the posterior $p(z|x)$. The knowledge of the latent will be helpful for:

- Understand the data:
    - Which cluster do the data belong to?
    - Which topics does the document have?
- Use in downstream tasks (Bayesian estimator, predictive distribution)
- *Solve the fitting problem!*

Unfortunately, the inference is also intractable since in general $p(z|x) = \frac{p(z,x)}{p(x)}$, which involves the intractable marginal in the denominator.

One good news is if we can estimate the posterior, we can reuse the result to fix the issue 1 mentioned above. By employing important sampling, we can sample from the posterior $p(z|x)$ instead of the prior $p(x)$ to estimate the expectation of the conditional. Compared to the prior, samples from the posterior is more informative to produce non-zero $p_\theta(x|z)$ and hasten the learning of θ. **Hence, the core problem of LVM is to estimate the posterior** $p(z|x)$.

Exercises

1. Derive the optimal solution for the fitting problem using Gaussian distribution model.
2. Prove Eq. 3.
3. Compute the bias of $\log$ of expectation.
4. Rewrite the log-likelihood using important sampling