Variational Inference

Variational Inference recap

Given observation x and latent variables z, we want to approximate the posterior distribution:

$$p(z|x) = \frac{p(x, z)}{p(x)}.$$

Variational Inference recap

Variational inference provides an optimization-based alternative to sampling algorithms.

- Choose variational approximation $q_\psi(z)$, paramterized by ψ.
- Minimize KL-divergence, which is equivalent to maximizing ELBO:

$$\psi^* = \max_\psi \mathbb{E}_{q_\psi(z)}[\log p(x, z)] - H(q_\psi)$$

Variational Inference recap

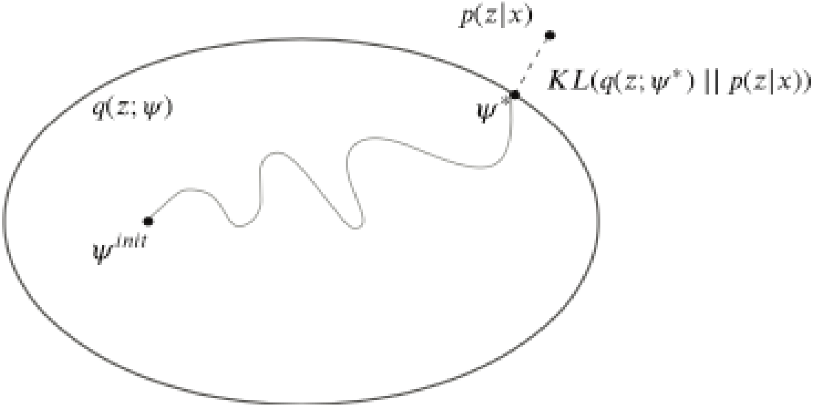

Figure 1: Section 10.1 from PML2

Coordinate ascent variational inference

For $z = z_{1:J}$ and mean-field VI:

$$q(z) = \prod_{j=1}^{J} q_j(z_j).$$

In this case, we can update the variational distribution for j with the rest fixed:

$$q_j^* \propto \exp(\mathbb{E}_{-z_j}[\log p(x, z)]).$$

- We need to compute the expectation wrt the Markov blanket of z_j.
- Possible for exponential family.

Latent Dirichlet Allocation

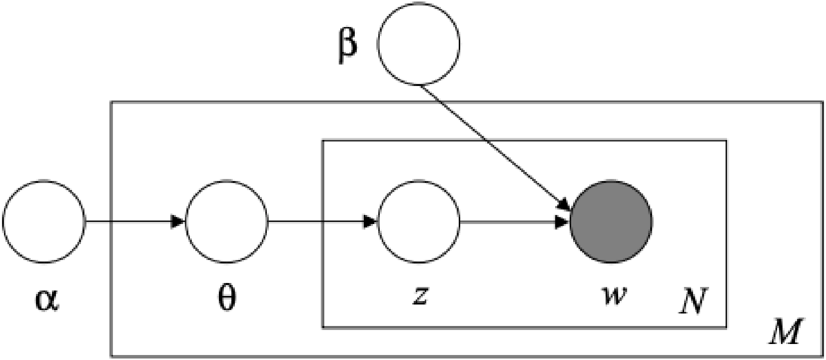

Figure 2: Blei et al (2003)

Latent Dirichlet Allocation: Topic modeling

- Data: M documents in a corpus.
- Each document is represented as a mixture of latent topics.

Latent Dirichlet Allocation: Topic modeling

- Data: M documents in a corpus.
- Each document is represented as a mixture of latent topics.

Sample a probability distribution over the topics for each document:

$$\theta_m \sim \text{Dirichlet}(\alpha).$$

Latent Dirichlet Allocation: Topic modeling

- Data: M documents in a corpus.
- Each document is represented as a mixture of latent topics.

Sample a probability distribution over the topics for each document:

$$\theta_m \sim \text{Dirichlet}(\alpha).$$

Choose the number of words $N \sim \text{Poisson}(\lambda)$.

Latent Dirichlet Allocation: Topic modeling

- Data: M documents in a corpus.
- Each document is represented as a mixture of latent topics.

Sample a probability distribution over the topics for each document:

$$\theta_m \sim \text{Dirichlet}(\alpha).$$

Choose the number of words $N \sim \text{Poisson}(\lambda)$.

For each word $n = 1, ..., N$:

- Select a topic $z_{m,n} \sim \text{Multinomial}(\theta)$,
- Generate a word $w_{m,n} \sim p(w|z_n, \beta)$, a probability distribution over words for a given topic z_n parameterized by β.

Latent Dirichlet Allocation: Topic modeling

$$\theta_m \sim \text{Dirichlet}(\alpha) \quad (1)$$

$$N_m \sim \text{Poisson}(\lambda) \quad (2)$$

$$z_{m,n} \sim \text{Multinomial}(\theta) \quad (3)$$

$$w_{m,n} \sim p(\cdot|z_n, \beta). \quad (4)$$

Dirichlet prior on θ and the Multinomial distribution over the topics $z_{m,n}$ are in the exponential family and are conjugate distributions.

Latent Dirichlet Allocation: Topic modeling

The posterior distribution:

$$p(z, \theta | w, \alpha, \beta) = \frac{p(z, w|\beta, \theta)p(\theta|\alpha)}{p(w|\alpha, \beta)} \tag{5}$$

$$= \frac{\prod_{m=1}^{M} \prod_{n=1}^{N_m} p(w_{m,n}, z_{m,n}|\beta, \theta)p(\theta|\alpha)}{p(w|\alpha, \beta)}. \tag{6}$$

Latent Dirichlet Allocation: Topic modeling

The marginal likelihood:

$$p(w|\alpha, \beta) = \prod_{m=1}^{M} \int p(\theta_m|\alpha) \left(\prod_{n=1}^{N_m} \sum_{z_{m,n}} p(z_{m,n}|\theta_m) p(w_{m,n}|z_{m,n}, \beta) \right) d\theta_m \tag{7}$$

Latent Dirichlet Allocation: Topic modeling

Variational approximation:

$$q_{\gamma,\phi}(\theta, z) = \prod_{m=1}^{M} q_{\gamma_m}(\theta_m) \prod_{n=1}^{N_m} q_{\phi_n}(z_{m,n}). \tag{8}$$

$$\gamma^*, \phi^* = \min_{\gamma,\phi} D_{KL}(q_{\gamma,\phi}(\theta, z)||p(\theta, z|w, \alpha, \beta)).$$

How many parameters do we have? $M \times K + \sum_{m=1}^{M} N_m \times K$.

Latent Dirichlet Allocation: Topic modeling

Use CAVI: for each document m, q_{γ_m} is Dirichlet and $q_{\phi_{n,k}}$ is Multinomial for $n = 1, ..., N_m$. The parameter updates:

$$\phi_{n,k} \propto \beta_{k,w_n} \exp(\mathbb{E}_q[\log(\theta_{n,k})|\gamma_m]) \tag{9}$$

$$\gamma_{m,k} = \alpha_k + \sum_{n=1}^{N_m} \phi_{n,k}. \tag{10}$$

The closed form expectation can be derived:

$$\mathbb{E}_q[\log(\theta_k)|\gamma] = \Psi(\phi_k) - \Psi(\sum_{j=1}^{K} \phi_j),$$

Ψ is the digamma function (the first derivative of $\log \Gamma$ function).

Latent Dirichlet Allocation: Topic modeling

What about the parameters α, β?

These can be updated using Variational Expectation-Maximization algorithm.

- Variational EM was proposed by Neal and Hinton (1998).
- In the E-step, maximize the ELBO with respect to variational parameters.
- In the M-step, maximize ELBO wrt α, β.

Latent Dirichlet Allocation: Topic modeling

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Figure 3: Blei et al (2003).

Latent Dirichlet Allocation: Topic modeling

- 16,000 documents from Associated Press newswire stories.
- Trained LDA model with 100-topics.
- Top words from each identified topics are shown above (note: topic labeling is manually done post inference).

Gradient-based optimization for VI

CAVI requires being able to compute the exact expectation (exponential family).

- To make VI applicable to broader settings, we want to utilize gradient-based learning.
- The major challenge is that we need to compute the gradient of the ELBO, which involves computing the expectation:

$$\nabla_\gamma \mathbb{E}_{q_\gamma(z)}[\log p(x, z) - \log q_\gamma(z)].$$

Stochastic optimization

Robbins and Munro (1951) showed that convergence is possible using only an unbiased estimator of the gradient.

Let $L(x, z, \gamma) = \log p(x, z) - \log q_\gamma(z)$.

$$\nabla_\gamma \mathbb{E}_{q_\gamma(z)}[L(x, z, \gamma)] = \nabla_\gamma \int q_\gamma(z) L(x, z, \gamma) dz \tag{11}$$

$$= \int q_\gamma(z) \nabla_\gamma L(x, z, \gamma) dz - \int L(x, z, \gamma) \nabla_\gamma q_\gamma(z) dz \tag{12}$$

Note: we invoke Leibniz theorem to interchange derivative and integral.

Stochastic optimization

We can sample $z_n \sim q_\gamma$ to approximate the first integral.

$$\int q_\gamma(z) \nabla_\gamma L(x, z, \gamma) dz \approx \frac{1}{N} \sum_{n=1}^{N} \nabla_\gamma L(x, z, \gamma).$$

How do we approximate the second integral?

$$\int L(x, z, \gamma) \nabla_\gamma q_\gamma(z) dz = ??$$

REINFORCE (Score function estimator)

Score function: $\nabla_\gamma \log q_\gamma(z)$.

$$\nabla_\gamma q_\gamma(z) = q_\gamma(z) \nabla_\gamma \log q_\gamma(z).$$

$$\mathbb{E}_{q_\gamma(z)}[L(x,z,\gamma) \nabla_\gamma \log q_\gamma(z)] = \int L(x,z,\gamma) q_\gamma(z) \frac{\nabla_\gamma q_\gamma(z)}{q_\gamma(z)} dz \quad (13)$$

$$= \int L(x,z,\gamma) \nabla_\gamma q_\gamma(z) dz. \quad (14)$$

REINFORCE (Score function estimator)

So we can sample $z_n \sim q_\gamma(z)$ and estimate the second integrand:

$$\int L(x, z, \gamma) \nabla_\gamma q_\gamma(z) \approx \frac{1}{N} \sum_{n=1}^{N} L(x, z_n, \gamma) \nabla_\gamma \log q_\gamma(z_n). \tag{15}$$

REINFORCE (Score function estimator)

REINFORCE estimator is known to have high variance. We can reduce the variance by using

- Control variates,
- Rao-Blackwellization,

where applicable.

Reparameterization trick

Suppose the latent variable z is Normally distributed with $\gamma = (\mu, \sigma^2)$.

Then, we can sample $z_n \sim q_\gamma$ by first sampling $\epsilon_n \sim \text{Normal}(0, 1)$ and $z_n = \mu + \sigma \cdot \epsilon_n$.

Then,

$$\mathbb{E}_{q_\gamma(z)}[L(x, z, \gamma)] = \mathbb{E}_{q(\epsilon)}[L(x, g(\gamma, \epsilon))].$$

$$\begin{aligned} \nabla_\gamma \mathbb{E}_{q(\epsilon)}[L(x, g(\gamma, \epsilon))] &= \mathbb{E}_{q(\epsilon)}[\nabla_\gamma L(x, g(\gamma, \epsilon))] \\ &\approx \frac{1}{N} \sum_{n=1}^{N} \nabla_\gamma L(x, g(\gamma, \epsilon_n)). \end{aligned}$$

Reparameterization trick

Generally, the idea is to reparameterize $z = g(\gamma, \epsilon)$ where $\epsilon \sim q_0$ is free of the variational parameters γ.

Examples:

- Normal: $z = \mu + \sigma\epsilon$, $\epsilon \sim$ Normal$(0, 1)$.
- Exponential: $z \sim$ Exp(λ) then $z = -\frac{1}{\lambda}\log(\epsilon)$, $\epsilon \sim$ Uniform$(0, 1)$.
- Gumbel-softmax trick for discrete z.

Reparameterization trick

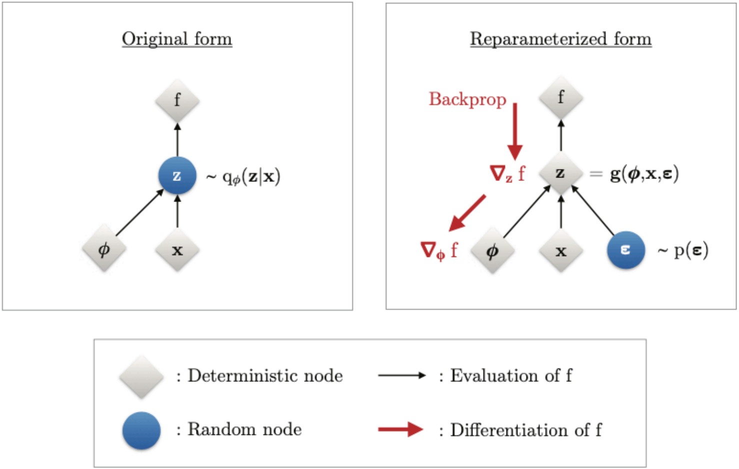

Figure 4: Figure 10.4 PML2

Variational Auto-Encoders (VAE)

Auto-encoder is a neural network trained to learn low-dimensional embedding of the input data by reconstruction.

- Input: $x \in \mathcal{X}$.
- Output: $\tilde{x} \in \mathcal{X}$.

By attempting to compress the input into a lower embededding, the neural network architecture learns the essential features of the data.

Variational Auto-Encoders (VAE)

There are two components of an auto-encoder:

- Encoder: $q_\psi : \mathcal{X} \to \mathcal{Z}$.
- Decoder: $p_\theta : \mathcal{Z} \to \mathcal{X}$.

ψ, θ denote the parameters of encoder and decoder neural networks.

The original auto-encoder minimizes reconstruction error (loss function L):

$$\theta^*, \psi^* = \min_{\theta,\psi} \sum_i L(x_i, p_\theta(q_\psi(x_i)))$$

Variational Auto-Encoders (VAE)

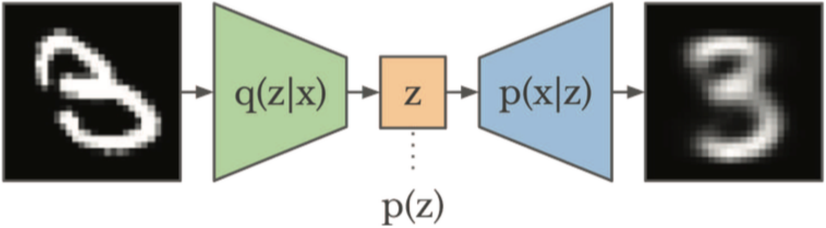

Figure 5: PML2

Variational Auto-Encoders (VAE)

"Auto-Encoding Variational Bayes" by Kingma and Welling (2013) proposed probabilistic formulation of the auto-encoder (over 40,000 citations).

- The data is generated given latent z: $p_\theta(x|z)$.
- Prior on the latent: $p(z) = N(0, I)$.
- Approximate the posterior $p_\theta(z|x) \propto p_\theta(x|z)p(z)$.

Variational Auto-Encoders (VAE)

In the original paper: $q_\psi(z|x) = \prod q_\psi(z_d|x)$, where

$$q_\psi(z_d|x) = N(\mu_{\psi,d}(x), \sigma^2_{\psi,d}(x)).$$

- μ_ψ, σ_ψ represent transformation of outputs from a neural network parameterized by ψ.

Variational Auto-Encoders (VAE)

Reparameterization trick is used to separate the parameters ψ from the randomness in z:

$$z_d = \mu_{\psi,d}(x) + \sigma_{\psi,d}(x)\epsilon_d,$$

where $\epsilon_d \sim N(0, 1)$.

Note: in the original VAE fomrulation, only one ϵ sample is taken.

Variational Auto-Encoders (VAE)

Maximize ELBO as the loss:

$$\psi^*, \theta^* = \max_{\psi,\theta} \mathbb{E}_{q_\psi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\psi(z|x)||p(z)).$$

- The first term aims to minimize the reconstruction error, commonly use binary cross-entropy:

$$L(x, \tilde{x}) = -\sum_{i=1}^{N} x_i \log \hat{x}_i + (1 - x_i) \log(1 - \hat{x}_i).$$

- The second term serves to regularize the neural network parameters.
- The gradient optimization allows gradients to flow, allowing optimization of both ψ, θ.

Variational Auto-Encoders (VAE)

Demo on Colab.