S&DS 365 / 665 Intermediate Machine Learning

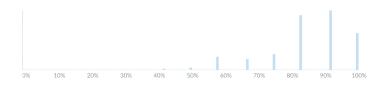
### **Variational Inference**

October 7

#### Reminders

- Assignment 2 due Wednesday
- Midterm week from today, in class
- Material up to and including today's class
- Cheat sheet: One side of 8.5x11 sheet, handwritten
- Multiple review sessions scheduled (see Canvas)
- Quiz 3 scores posted; can discuss in review sessions
- Practice midterms posted





### **For Today**

- Variational inference: The ELBO
- Derivations and examples
- Next time: Variational autoencoders (VAEs)



## **Recall: Inverting generative models**

#### Template for generative model:

- Choose Z
- ② Given z, generate (sample) X

We often want to invert this:

- Given x
- What is Z that generated it?

# Inverting models

#### Bayesian setup:

- **1** Choose  $\theta$
- 2 Given  $\theta$ , generate (sample) X

#### Posterior inference:

- Given x
- **2** What is  $\theta$  that generated it?

### Approximate inference

If we have a random vector  $Z \sim p(Z \mid x)$ , we might want to compute the following:

- marginal probabilities  $\mathbb{P}(Z_i = z \mid x)$
- marginal means  $\mathbb{E}(Z_i = z \mid x)$
- most probable assignments  $z^\star = \operatorname{\mathsf{arg}} \max_{Z} \mathbb{P}(\{Z_i = z_i\} \,|\, x)$
- maximum marginals  $z_i^* = \arg \max_{z_i} \mathbb{P}(Z_i = z_i \mid x)$
- joint probability  $\mathbb{P}(Z \mid x)$
- joint mean  $\mathbb{E}(Z \mid x)$

Each of these quantities is intractable to calculate exactly, in general.

7

### Gibbs sampler

#### Generic algorithm:

- Sample one component holding others fixed
- 2 Iterate

### Gibbs sampler

#### Generic algorithm:

- Choose index *j* randomly
- **2** Sample  $\theta_i$  fixing  $\theta_i$ ,  $i \neq j$  and x
- 3 Iterate

### Gibbs sampler

#### Generic algorithm:

- Choose index *j* randomly
- ② Sample  $\theta_i$  from  $\mathbb{P}(\theta_i | \theta_i, i \neq j, x)$ .
- 3 Iterate

When run for sufficiently long, this will draw a sample from the posterior  $\mathbb{P}(\theta \mid x)$ .

8

#### Variational methods

- Gibbs sampling is stochastic approximation
- Variational methods iteratively refine deterministic approximations
- Variational and Markov chain approximations originated in physics

### Stochastic approximation

#### Gibbs sampler

Iterate until converged:

- **1** Choose vertex  $s \in V$  at random
- 2 Sample  $z_s$  holding others fixed

$$egin{aligned} heta_{ extbf{s}} &= ext{sigmoid} \left(eta_{ extbf{s}} + \sum_{t \in extbf{N(s)}} eta_{st} z_t
ight) \ Z_{ extbf{s}} \mid heta_{ extbf{s}} \sim ext{Bernoulli}( heta_{ extbf{s}}) \end{aligned}$$

You should verify that this is the correct conditional distribution of  $Z_s$  holding the others fixed.

 $sigmoid(u) = 1/(1 + e^{-u})$ 

1(

### **Deterministic approximation**

#### Mean field variational algorithm

Iterate until converged:

- **1** Choose vertex  $s \in V$  at random
- 2 Update mean  $\mu_s$  holding others fixed

$$\mu_{s} = \operatorname{sigmoid}\left(\beta_{s} + \sum_{t \in N(s)} \beta_{st} \mu_{t}\right)$$

1:

### Deterministic vs. stochastic approximation

- The z<sub>s</sub> variables are random
- The  $\mu_s$  variables are deterministic
- The Gibbs sampler convergence is in distribution
- The mean field convergence is numerical
- The Gibbs sampler approximates the full distribution
- The mean field algorithm approximates the mean of each node

### **Example 2: A finite mixture model**

Fix two distributions  $F_0$  and  $F_1$ , with densities  $f_0(x)$  and  $f_1(x)$ , and form the mixture model

$$heta \sim \mathrm{Beta}(lpha,eta) \ X \, | \, heta \sim heta F_1 + (1- heta) F_0.$$

The likelihood for data  $x_1, \ldots, x_n$  is

$$p(x_{1:n}) = \int_0^1 \text{Beta}(\theta \mid \alpha, \beta) \prod_{i=1}^n (\theta f_1(x_i) + (1-\theta) f_0(x_i)) d\theta.$$

Our goal is to approximate the posterior  $p(\theta \mid x_{1:n})$ 

### Stochastic approximation

#### Gibbs sampler

- **1** Sample  $Z_i | \theta, x_{1:n}$  for i = 1, ..., n
- **2** Sample  $\theta \mid z_{1:n}, x_{1:n}$

The first step is carried out by sampling

$$Z_i = \begin{cases} 1 & \text{with probability } \propto \theta f_1(x_i) \\ 0 & \text{with probability } \propto (1 - \theta) f_0(x_i) \end{cases}$$

### Stochastic approximation

#### Gibbs sampler

- **1** Sample  $Z_i | \theta, x_{1:n}$  for i = 1, ..., n
- **2** Sample  $\theta \mid z_{1:n}, x_{1:n}$

The second step is carried out by sampling

$$\theta \sim \text{Beta}\left(\sum_{i=1}^{n} z_i + \alpha, n - \sum_{i=1}^{n} z_i + \beta\right).$$

Posterior over  $\theta$  is approximated as *mixture* of Beta distributions; number of components is n+1

### Variational inference: Strategy

- We'd like to compute  $p(\theta, z | x)$ , but it's too complicated.
- Strategy: Approximate as  $q(\theta, z)$  that has a "nice" form
- q is a function of variational parameters, optimized for each x.
- Maximize a lower bound on p(x).

#### Variational inference: The ELBO

The ELBO is the following lower bound on  $\log p(x)$ :

$$\log p(x) = \int \sum_{z} q(z,\theta) \log p(x) d\theta$$

$$= \sum_{z} \int q(z,\theta) \log \left( \frac{p(x,z,\theta) q(z,\theta)}{p(z,\theta \mid x) q(z,\theta)} \right) d\theta$$

$$= \sum_{z} \int q(z,\theta) \log \left( \frac{p(x,z,\theta)}{q(z,\theta)} \right) d\theta + \sum_{z} \int q(z,\theta) \log \left( \frac{q(z,\theta)}{p(z,\theta \mid x)} \right) d\theta$$

$$\geq \sum_{z} \int q(z,\theta) \log \left( \frac{p(x,z,\theta)}{q(z,\theta)} \right) d\theta$$

$$= H(q) + \mathbb{E}_{q}(\log p(x,z,\theta))$$

We maximize this over the parameters of q



#### Variational inference: The ELBO

The inequality above uses concavity of the logarithm:

$$\log\left(\sum_{\alpha} w_{\alpha} x_{\alpha}\right) \geq \sum_{\alpha} w_{\alpha} \log x_{\alpha}$$

So, if  $q_{\alpha} \geq 0$  and  $p_{\alpha} \geq 0$  sum (or integrate) to one, then

$$0 = \log \left( \sum_lpha p_lpha 
ight) = \log \left( \sum_lpha q_lpha rac{p_lpha}{q_lpha} 
ight) \geq \sum_lpha q_lpha \log \left( rac{p_lpha}{q_lpha} 
ight)$$

Therefore

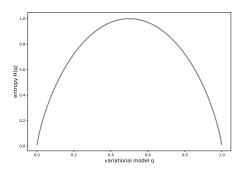
$$\sum_{lpha} q_lpha \log \left(rac{q_lpha}{p_lpha}
ight) \geq 0$$

17

### Variational inference: The ELBO

The ELBO is  $H(q) + \mathbb{E}_q(\log p)$ 

The entropy term H(q) encourages q to be spread out:



The cross-entropy  $\mathbb{E}_q \log p$  tries to match q to p

### **Example 2: A finite mixture model**

Fix two distributions  $F_0$  and  $F_1$ , with densities  $f_0(x)$  and  $f_1(x)$ , and form the mixture model

$$heta \sim \mathrm{Beta}(lpha,eta) \ X \, | \, heta \sim heta F_1 + (1- heta) F_0.$$

The likelihood for data  $x_1, \ldots, x_n$  is

$$p(x_{1:n}) = \int_0^1 \text{Beta}(\theta \mid \alpha, \beta) \prod_{i=1}^n (\theta f_1(x_i) + (1-\theta) f_0(x_i)) d\theta.$$

Our goal is to approximate the posterior  $p(\theta \mid x_{1:n})$ 

### Variational approximation

Our variational approximation is

$$q(z,\theta) = q(\theta \mid \gamma_1, \gamma_2) \prod_{i=1}^n q_i^{z_i} (1-q_i)^{(1-z_i)}$$

where  $q(\theta \mid \gamma_1, \gamma_2)$  is a Beta $(\gamma_1, \gamma_2)$  distribution, and  $0 \le q_i \le 1$  are n free parameters.

Need to maximize ELBO  $H(q) + \mathbb{E}_q \log p$ 

Let's sketch part of the calculation

### Variational approximation

First, we have

$$\log p(x,\theta,z) = \log p(\theta \mid \alpha,\beta) + \sum_{i=1}^{n} \left\{ \log \left( \theta^{z_i} f_1(x_i) \right) + \log \left( \theta^{1-z_i} f_0(x_i) \right) \right\}$$

Next we use identities such as

$$\mathbb{E}_q \log \theta = \psi(\gamma_1) - \psi(\gamma_1 + \gamma_2)$$

for the digamma function  $\psi(\cdot)$ .

After some calculus and algebra, we end up with the following algorithm (see the notes on the course web page for more detail)

### Variational algorithm for mixture

#### Variational inference

Iterate the following steps for variational parameters  $q_{1:n}$  and  $(\gamma_1, \gamma_2)$ :

**1** Holding  $q_i$  fixed, set  $\gamma = (\gamma_1, \gamma_2)$  to

$$\gamma_1 = \alpha + \sum_{i=1}^n q_i$$
  $\gamma_2 = \beta + n - \sum_{i=1}^n q_i$ 

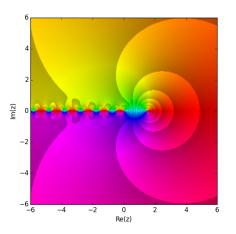
2 Holding  $\gamma_1$  and  $\gamma_2$  fixed, set  $q_i$  to

$$q_i = \frac{f_1(x_i) \exp \psi(\gamma_1)}{f_1(x_i) \exp \psi(\gamma_1) + f_0(x_i) \exp \psi(\gamma_2)}$$

After convergence, approximate posterior distribution over  $\theta$  is

$$\widehat{p}(\theta \mid x_{1:n}) = \mathsf{Beta}(\theta \mid \gamma_1, \gamma_2)$$

### **Digamma function**



 $\psi(\mathbf{x})$  is the digamma function  $\label{eq:psi} \text{https://en.wikipedia.org/wiki/Digamma\_function}$ 

### **Deterministic approximation**

- Convergence is numerical, not stochastic
- Posterior is approximated as a single Beta
- Very similar algorithm is used for topic models

### **Example 3: More general mixtures**

$$heta \sim \mathsf{Dirichlet}(lpha_1, \dots, lpha_k)$$
 $X \mid heta \sim heta_1 F_1 + \dots + heta_k F_k$ 

The likelihood for single data point x is

$$p(x) = \int \mathsf{Dirichlet}(\theta \,|\, \alpha_1, \dots, \alpha_k) \left( \sum_{j=1}^k \theta_j f_j(x) \right) \, d\theta.$$

When distributions  $F_j$  are learned, this is a "topic model." Variational inference is one of the most useful ways of training topic models

- Notes on variational methods: Please read Sections 1–4
- The other sections are more advanced / specialized material
- Next up: Variational autoencoders
- VAEs not on midterm, but please review basics of variational methods discussed so far, including the ELBO



Q: What is the best *q* we could use?

Q: What is the best *q* we could use?

A: The true posterior  $q(z, \theta \mid x) = p(z, \theta \mid x)$ 

Q: What is the best q we could use?

A: The true posterior  $q(z, \theta \mid x) = p(z, \theta \mid x)$ 

Why? Because this maximizes the ELBO. Mathematically,

$$\sum_{z} \int q(z,\theta) \log \left( \frac{q(z,\theta)}{p(z,\theta \mid x)} \right) d\theta = 0$$

in this case, so the ELBO inequality is an equality.

Q: How does the ELBO regularize?

Q: How does the ELBO regularize?

A: The entropy term favors distributions that are "spread out"

Q: How does the ELBO regularize?

A: The entropy term favors distributions that are "spread out"

Why? For discrete distributions the maximum entropy distribution is uniform. For Gaussian, the entropy is  $\log \sigma^2$  which favors  $\sigma^2$  large.

Q: Is the ELBO easy to maximize?

Q: Is the ELBO easy to maximize?

A: No.

Q: Is the ELBO easy to maximize?

A: No.

Why? In general it is non-convex, and the solution depends on where we start an iterative algorithm. This is unlike the Gibbs sampler, which converges to the right thing if we wait long enough.

#### **Next class: Variational autoencoders**

- Variational autoencoders are generative models that are trained using variational inference
- The "decoder" is a neural net that generates from a latent variable
- The "encoder" approximates the posterior distribution with another neural network trained using variational inference

### **Summary**

- Gibbs sampling makes stochastic approximations
- Variational methods make deterministic approximations
- General recipe: Maximize ELBO over variational parameters
- Gives a powerful approach to generative modeling