

# Solutions

10-418/10-618 ML for Structured Data  
Fall 2022

Practice Problems 2

December 8, 2022

Time Limit: – minutes

Name: \_\_\_\_\_

Andrew ID: \_\_\_\_\_

---

## Instructions:

- Verify your name and Andrew ID above.
  - This exam contains 33 pages (including this cover page).  
The total number of points is 87.
  - Clearly mark your answers in the allocated space. If you have made a mistake, cross out the invalid parts of your solution, and circle the ones which should be graded.
  - Look over the exam first to make sure that none of the 33 pages are missing.
  - No electronic devices may be used during the exam.
  - Please write all answers in pen or *darkly* in pencil.
  - You have – minutes to complete the exam. Good luck!
-

## Instructions for Specific Problem Types

For “Select One” questions, please fill in the appropriate bubble completely:

**Select One:** Who taught this course?

- ☒ Matt Gormley
- ☐ Marie Curie
- ☐ Noam Chomsky

If you need to change your answer, you may cross out the previous answer and bubble in the new answer:

**Select One:** Who taught this course?

- ☒ Matt Gormley
- ☐ Marie Curie
- ☒ Noam Chomsky

For “Select all that apply” questions, please fill in all appropriate squares completely:

**Select all that apply:** Which are scientists?

- ☒ Stephen Hawking
- ☒ Albert Einstein
- ☒ Isaac Newton
- ☐ I don't know

Again, if you need to change your answer, you may cross out the previous answer(s) and bubble in the new answer(s):

**Select all that apply:** Which are scientists?

- ☒ Stephen Hawking
- ☒ Albert Einstein
- ☒ Isaac Newton
- ☒ I don't know

For questions where you must fill in a blank, please make sure your final answer is fully included in the given space. You may cross out answers or parts of answers, but the final answer must still be within the given space.

**Fill in the blank:** What is the course number?

10-418

10-~~4~~618

# 1 Short Questions

1. (1 point) **Multiple Choice:** Suppose we have a factor graph that is a tree and we hope to do inference. What is the best approach we've learned?
- ☐ Variational Inference
  - ☐ Belief Propagation
  - ☐ DAgger
  - ☐ Loopy Belief Propagation

Belief Propagation

2. (1 point) **Multiple Choice:** What flaw in structured prediction as search does DAgger hope to fix?
- ☐ DAgger fixes the assumption that the output space is a small set of actions.
  - ☐ DAgger fixes the assumption that we have access to a reliable expert.
  - ☐ DAgger fixes the assumptions that the distribution over test and train states is the same.

The distribution over test and train states.

3. (1 point) **True or False:** The Naive Bayes Classifier and a Hidden Markov Model are both examples of Bayesian Networks.
- ☐ True
  - ☐ False

True

4. (1 point) **True or False:** Bayesian Networks and Markov Random Fields can express the same types of distributions. The only reason to pick one over the other is computational efficiency.
- ☐ True
  - ☐ False

False

5. (1 point) **True or False:** Markov Random Fields model  $p(\mathbf{y}|\mathbf{x})$  and Conditional Random Fields model  $p(\mathbf{y}, \mathbf{x})$ .
- ☐ True
  - ☐ False

False. It's the exact opposite.

6. (1 point) **Fill in the blank:** *The minimum Bayes risk (MBR) decoder for \_\_\_\_\_ on a conditional random field (CRF) is  $h(\mathbf{x}) = \arg \max_{\hat{\mathbf{y}}} p(\hat{\mathbf{y}}|\mathbf{x})$ .*

- ☐ Hamming loss  $\ell(\hat{\mathbf{y}}, \mathbf{y}) = \sum_v (1 - \mathbb{1}(\hat{y}_v = y_v))$
- ☐ 0 – 1 loss  $\ell(\hat{\mathbf{y}}, \mathbf{y}) = 1 - \mathbb{1}(\hat{\mathbf{y}} = \mathbf{y})$
- ☐ cross entropy loss  $\ell(\hat{\mathbf{y}}, \mathbf{y}) = \sum_v y_v \log(\hat{y}_v)$
- ☐ None of the above

0 – 1 loss  $\ell(\hat{\mathbf{y}}, \mathbf{y}) = 1 - \mathbb{1}(\hat{\mathbf{y}} = \mathbf{y})$

## 2 Monte Carlo Methods and Markov Chains

1. (1 point) **True or False:** Monte Carlo methods can be used to generate samples from a distribution  $p(\mathbf{x}) = \frac{\tilde{p}(\mathbf{x})}{Z}$  when the partition function  $Z$  is unknown.

☐ True

☐ False

True.

2. Suppose we employ Rejection Sampling to draw samples from  $p(\mathbf{x})$  where  $\mathbf{x} \in \mathbb{R}^M$  using a proposal distribution  $q(\mathbf{x})$ .

- (a) (1 point) **True or False:** Many samples will be rejected if the proposal distribution is proportional to the true distribution (i.e.  $q(\mathbf{x}) \propto p(\mathbf{x})$ ). Assume  $kq(\mathbf{x}) \geq p(\mathbf{x})$  with  $k$  chosen to give as tight an upper bound as possible.

☐ True

☐ False

False

- (b) (1 point) **True or False:** The sampling procedure will scale poorly as the dimensionality  $M$  increases.

☐ True

☐ False

True

3. (1 point) Suppose you are given a first order Markov Chain over a series of random variables  $Y_1, Y_2, Y_3, \dots$ . Let  $P(Y_t)$  be the marginal probability for variable  $Y_t$ .

**True or False:** The equilibrium distribution of the Markov Chain is  $P^*(Y)$  if and only if  $\lim_{t \rightarrow \infty} P(Y_t) = P^*(Y)$ .

☐ True

☐ False

True

### 3 Gibbs Sampling

1. Consider the following joint probabilities for two random variables,  $X$  and  $Y$ .

$$\begin{bmatrix} P(X=0, Y=0) & P(X=0, Y=1) \\ P(X=1, Y=0) & P(X=1, Y=1) \end{bmatrix} = \begin{bmatrix} 0.1 & 0.2 \\ 0.5 & 0.2 \end{bmatrix}$$

Given that this is our target distribution, you wish to design a Gibbs sampler to draw samples from it. Below, define the four “full conditionals” required to build the Gibbs sampler.

- (a) (1 point) **Numerical Answer:** What is the full conditional  $P(X=0|Y=0)$

- (b) (1 point) **Numerical Answer:** What is the full conditional  $P(X=1|Y=0)$

- (c) (1 point) **Numerical Answer:** What is the full conditional  $P(X=0|Y=1)$

- (d) (1 point) **Numerical Answer:** What is the full conditional  $P(X=1|Y=1)$

$$\begin{bmatrix} 0.1/0.6 & 0.2/0.4 \\ 0.5/0.6 & 0.2/0.4 \end{bmatrix}$$

2. Consider a new distribution over four random variables  $P(Z_1, Z_2, Z_3, Z_4)$  for which you build a Gibbs sampler.

- (a) (1 point) **Numerical Answer:** Suppose you begin with the sample  $[z_0, z_1, z_2, z_3] = [0, 0, 0, 0]$ . How many sampling steps are needed to transition to the sample  $[1, 1, 1, 1]$ ?

4

- (b) (1 point) **Short Answer:** Describe or name a (closely related) sampling method that can transition from  $[0, 0, 0, 0]$  to  $[1, 1, 1, 1]$  in a smaller number of steps.

---



---

blocked Gibbs sampling

## 4 Metropolis-Hastings

1. Now let's sample from a density  $\pi(x)$ ,  $x \in \mathbb{R}$ , using the Metropolis-Hastings algorithm. The proposal, at time  $n$  of the algorithm (with  $X_0 = 0$ ) is

$$Y = X_{n-1} + \sigma Z_n$$

with  $Z_n \sim N(0, 1)$  (normal distribution zero mean, unit variance),  $\sigma$  a known scaling factor and  $Z_n$  independent of all other variables for each time point  $n$ .

**Hint:** Recall that the density of a Gaussian with mean  $\mu$  and variance  $\sigma^2$  is:

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- (a) (2 points) **Derivation:** Write down the proposal density (i.e. given  $X_{n-1} = x$  for  $x$  fixed).

$$q(y|x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-y)^2}{2\sigma^2}\right)$$

- (b) (2 points) **Derivation:** Suppose that for the true density  $\pi(\cdot)$  we have:

$$\pi(x) \propto (1+x^2)^{-1}$$

What is the acceptance probability of the Metropolis-Hastings algorithm associated with the proposal in (a).

$$\min\left(\frac{(1+y^2)^{-1}}{(1+x^2)^{-1}}, 1\right)$$

- (c) (2 points) **Short Answer:** How does the performance of the sampler change as a function of  $\sigma$ ? (For simplicity, you can simply describe what happens with large and small values of

$\sigma$

.)

---

---

---

$\sigma$  too large, probability of rejection is very high. Too small, it will proceed with very small steps and therefore take a long time to cover the whole sample space

## 5 Bayesian Inference

1. Suppose we wish to take a Bayesian approach to linear regression. We have data consisting of feature vectors  $\mathbf{x} \in \mathbb{R}^M$  and outputs  $y^{(i)} \in \mathbb{R}$ . We define our probability model as follows:

$$p(y|\mathbf{x}) = \mathcal{N}(y|\mathbf{w}^T \mathbf{x}, \sigma^2)$$

We also impose a Gaussian prior over the parameters:

$$p(\mathbf{w}) = \prod_{m=1}^M \mathcal{N}(w_m|0, \tau^2)$$

Above,  $\mathcal{N}$  is the pdf of a univariate Gaussian:

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- (a) (2 points) **Short Answer:** Consider the distribution  $p(\mathbf{w}|\mathbf{x}, \mathbf{y})$ . To what family would this distribution belong?

**Gaussian**

- (b) (1 point) **Select all that apply:** Suppose we are working with a prior such as the Laplace distribution. Which of the following methods could be used to approximate  $p(\mathbf{w}|\mathbf{x}, \mathbf{y})$ ?

- ☐ Variational Inference
- ☐ MCMC
- ☐ Structured SVM
- ☐ None of the above

**VI and MCMC.**

2. (1 point) **Select One:** Select the best description of the goal of Bayesian inference.
- ☐ To find a prior over parameters.
  - ☐ To find a MAP estimate of parameters.
  - ☐ To find an MLE estimate of parameters.
  - ☐ To find a posterior distribution over parameters.

**To find a posterior distribution.**

## 6 Deep Learning

1. (1 point) **True or False:** Models that combine aspects of deep learning and graphical models typically result in inference-related intractabilities that render such models useless to the machine learning practitioner.

☐ True

☐ False

False.

2. (2 points) **Numerical Answer:** Suppose you are given the grayscale 3x3 image in Figure 1 and the parameters of a 2x2 convolution in Figure 2. (Note: the pixel values of a grayscale image range from 0.0 (black) to 1.0 (white).)

0.5	0.0	1.0
1.0	1.0	1.0
1.0	0.0	0.0

Figure 1: Image

1	0
1	1

Figure 2: Convolution

You apply the convolution with stride 1 and no padding to the image to produce a 2x2 convolved image. Write the pixel values of the convolved image into Figure 3.


Figure 3: Convolved Output

2.5	2.0
2.0	1.0

3. (1 point) **Select one:** One of the key insights in ResNet was the use of residual connections. Suppose you are building a neural network. One of the layers has input  $\mathbf{x}$  and produces output  $\mathbf{y} = f(\mathbf{x})$ , where  $f : \mathbb{R}^M \rightarrow \mathbb{R}^M$ . Which of the following is an augmentation of this layer with a residual connection? Below  $g(\mathbf{x}) = \mathbf{W}\mathbf{x} + b$  is linear function of  $\mathbf{x}$  with  $g(\mathbf{x}) \in \mathbb{R}^M$ ,  $\lambda \in \mathbb{R}$  is a hyperparameter.

☐  $\mathbf{y} = \lambda f(\mathbf{x}) + (1 - \lambda)g(\mathbf{x})$

☐  $\mathbf{y} = f(\mathbf{x}) + \mathbf{x}$

☐  $\mathbf{y} = (\mathbf{x}^T \mathbf{x})f(\mathbf{x})$

☐  $\mathbf{y} = \lambda f(\mathbf{x}) + (1 - \lambda)\mathbf{y}$

☐ None of the above

$\mathbf{y} = f(\mathbf{x}) + \mathbf{x}$

## 7 Variational Inference

1. (1 point) **Select all that apply:** Which of the multivariate Gaussian distributions in Figure 4 can be perfectly recovered by a mean-field approximation?

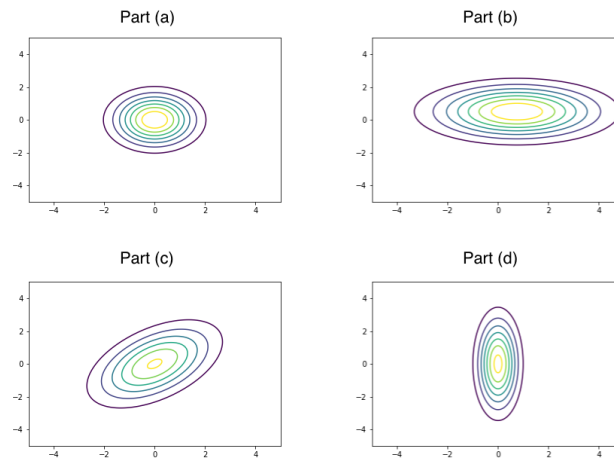


Figure 4: 2d Gaussian distributions with different means and covariance matrices.

- ☐ Gaussian in Part (a)
- ☐ Gaussian in Part (b)
- ☐ Gaussian in Part (c)
- ☐ Gaussian in Part (d)
- ☐ None of the above

(a), (b), and (d). Not (c).

2. Assume that we have a ground truth distribution  $p(x, y)$  over variables  $x, y$  and we hope to approximate it with the mean field approximation  $q(x, y) = q(x)q(y)$ . Instead of minimizing  $D_{KL}(q||p)$  we opt to minimize  $D_{KL}(p||q)$ .

Recall that the KL-divergence between two discrete distributions  $p$  and  $q$  is as follows:

$$D_{KL}(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

- (a) (3 points) **Derivation:** Show that the KL-divergence  $D_{KL}(p(x, y)||q(x, y))$  can be written as a sum of KL-divergences between marginals of  $p$  and  $q$  plus some constant (by constant, we mean a term involving  $p$  only).

$$D_{KL}(p||q) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{q(x)q(y)} \quad (1)$$

$$= \sum_{x,y} p(x,y) \log p(x,y) - \sum_{x,y} p(x,y) \log q(x) - \sum_{x,y} p(x,y) \log q(y) \quad (2)$$

$$= \sum_{x,y} p(x,y) \log p(x,y) - \sum_x p(x) \log q(x) \sum_y p(y|x) - \sum_y p(y) \log q(y) \sum_x p(x|y) \quad (3)$$

$$= \sum_{x,y} p(x,y) \log p(x,y) - \sum_x p(x) \log q(x) - \sum_y p(y) \log q(y) \quad (4)$$

$$= \text{constant} + \sum_x (p(x) \log p(x) - p(x) \log q(x)) + \sum_y (p(y) \log p(y) - p(y) \log q(y)) \quad (5)$$

$$= \text{constant} + \sum_x p(x) \log \frac{p(x)}{q(x)} + \sum_y p(y) \log \frac{p(y)}{q(y)} \quad (6)$$

$$= \text{constant} + KL(p(x)||q(x)) + KL(p(y)||q(y)) \quad (7)$$

- (b) (2 points) **Short Answer:** Argue that the equality you derived above implies that  $KL(p(x,y) | q(x,y))$  is minimized when  $q(x) = p(x)$  and  $q(y) = p(y)$ .

The argmin of  $KL(p||q)$  is when  $p$  equals  $q$ . Thus we achieve the minimum of the sum of these two KL's of marginals when  $q(x) = p(x)$  and  $q(y) = p(y)$ .

3. (2 points) **Short Answer:** Suppose we wanted to use variational inference to model a probability distribution over real-world photographs. If we only want to generate realistic images, would it be better to optimize the standard KL-divergence,  $D_{KL}(p||q)$ , or the "reverse" KL-divergence,  $D_{KL}(q||p)$ ? Explain why.

---

---

---

---

---

Minimizing the standard KL-divergence incentivizes placing probability mass everywhere that the true distribution places mass. The reverse KL-divergence incentivizes not putting probability anywhere that the true distribution doesn't place mass. For this task we would want to use the reverse KL.

## 8 Expectation Maximization

These questions were never actually used on a 10-418/10-618 exam, but they are still a nice review of the basics around EM.

1. Let's see how to run expectation maximization for a simple problem! Assume you have two biased dice, **Red** and **Blue**, which are two-sided (i.e. they can only roll values  $\{0, 1\}$ ), with parameters  $\theta_{red}$  and  $\theta_{blue}$ . Here,  $\theta_{red} = P(\text{Red} = 1)$  and  $\theta_{blue} = P(\text{Blue} = 1)$ . Consider the following dice-throw procedure:

1. Choose a dice from (**Red**, **Blue**)
2. Toss the chosen dice and record the observation.

This procedure is run  $i$  times. Assume variables  $z_{i1}$  and  $z_{i2}$  record which dice was chosen in step 1 at the  $i^{th}$  toss. If  $z_{i1} = 1$ , the red dice is chosen. If  $z_{i2} = 1$ , the blue dice is chosen. Binary variables  $o_i$  record the outcome of the toss. If the dice lands on 0,  $o_i = 0$  and if the dice lands on 1,  $o_i = 1$ .

- (a) (2 points) **Numerical Answer:** Suppose you run this procedure 5 times, resulting in the sequence of dice choices **{Red, Red, Blue, Blue, Red}** and observations **{1, 0, 1, 1, 0}**. Compute the MLE estimates for  $\theta_{red}$  and  $\theta_{blue}$ .

Parameter	Estimate
$\theta_{red}$	
$\theta_{blue}$	

Parameter	Estimate
$\theta_{red}$	1/3
$\theta_{blue}$	1

- (b) (4 points) **Numerical Answer:** Now assume that we do not observe the  $z_{i1}$  and  $z_{i2}$  variables. This means that at each toss, we do not know which dice was used to make the toss. In this scenario, we need to use expectation maximization to estimate parameters. In the E-step, we compute the expected values of latent variables  $z_{i1}, z_{i2}$  after fixing parameters  $\theta_{red}, \theta_{blue}$ . For our model, the E-step estimates are as follows:

$$\mathbb{E}_{p(\mathbf{Z}|\mathbf{O})}[z_{i1}] = \frac{\theta_{red}^{o_i}(1 - \theta_{red})^{(1-o_i)}}{\theta_{red}^{o_i}(1 - \theta_{red})^{(1-o_i)} + \theta_{blue}^{o_i}(1 - \theta_{blue})^{(1-o_i)}} \quad (8)$$

$$\mathbb{E}_{p(\mathbf{Z}|\mathbf{O})}[z_{i2}] = \frac{\theta_{blue}^{o_i}(1 - \theta_{blue})^{(1-o_i)}}{\theta_{red}^{o_i}(1 - \theta_{red})^{(1-o_i)} + \theta_{blue}^{o_i}(1 - \theta_{blue})^{(1-o_i)}} \quad (9)$$

$$(10)$$

Assume  $\theta_{red} = \frac{1}{4}, \theta_{blue} = \frac{2}{3}$ . For the sequence of observations  $\{1, 0\}$ , compute the E-step estimates for  $z_{i1}$  and  $z_{i2}$

Parameter	Estimate
$z_{01}$	
$z_{02}$	
$z_{11}$	
$z_{12}$	

Parameter	Estimate
$z_{01}$	3/11
$z_{02}$	8/11
$z_{11}$	9/13
$z_{12}$	4/13

- (c) (2 points) **Numerical Answer:** After computing E-step estimates for the latent variables  $z_{i1}$  and  $z_{i2}$ , in the M-step, we use the estimates to find values for parameters  $\theta_{red}, \theta_{blue}$  which maximize the likelihood of the data. For our model, M-step parameter estimates are computed as follows:

$$\theta_{red} = \frac{\sum_i \mathbb{E}_{p(\mathbf{z}|\mathbf{O})}[z_{i1}]o_i}{\sum_i \mathbb{E}_{p(\mathbf{z}|\mathbf{O})}[z_{i1}]} \quad (11)$$

$$\theta_{blue} = \frac{\sum_i \mathbb{E}_{p(\mathbf{z}|\mathbf{O})}[z_{i2}]o_i}{\sum_i \mathbb{E}_{p(\mathbf{z}|\mathbf{O})}[z_{i2}]} \quad (12)$$

Given E-step estimates  $z_{i1} = \{4/5, 1/3\}$  and  $z_{i2} = \{1/5, 2/3\}$  and the observation sequence  $\{0, 1\}$ , compute the parameter estimates for  $\theta_{red}$  and  $\theta_{blue}$ .

Parameter	Estimate
$\theta_{red}$	
$\theta_{blue}$	

Parameter	Estimate
$\theta_{red}$	5/17
$\theta_{blue}$	10/13

## 9 Learning with Partial Observations

1. (2 points) **Short answer:** Suppose we have a tree-shaped factor graph defining a CRF over 7 output variables conditioned on 3 input variables. At training time, you only observe the values of 5 of the output variables. To train you optimize the marginal likelihood of the observed output variables given the inputs using SGD. How many times must you run belief propagation to compute each stochastic gradient? **Briefly justify your answer.**

---

---

---

2. Once for the numerator (clamped factor graph), once for the denominator (unclamped factor graph).

2. Suppose you wish to run Variational EM for a true distribution  $p_\alpha(\mathbf{x}, \mathbf{z})$  with a variational approximation  $q_\theta(\mathbf{z})$ .

- (a) (1 point) **Select all that apply:** Which of the following describe the Variational E-Step?

- ☐ keep  $\theta$  fixed and update  $\alpha$
- ☐ keep  $\alpha$  fixed and update  $\theta$
- ☐ run variational inference to minimize  $KL(q_\theta || p_\alpha)$
- ☐ improve  $E_{q_\theta}[\log p_\alpha(\mathbf{x}, \mathbf{z})]$  by adjusting  $\alpha$
- ☐ None of the above

B and C

- (b) (1 point) **Select all that apply:** Which of the following describe the Variational M-Step?

- ☐ keep  $\theta$  fixed and update  $\alpha$
- ☐ keep  $\alpha$  fixed and update  $\theta$
- ☐ run variational inference to minimize  $KL(q_\theta || p_\alpha)$
- ☐ improve  $E_{q_\theta}[\log p_\alpha(\mathbf{x}, \mathbf{z})]$  by adjusting  $\alpha$
- ☐ None of the above

A and D

3. Answer the following questions based on your understanding of variational expectation maximization (EM) algorithm.

(a) (1 point) **True or False:** Variational EM will always converge to the global optimum.

☐ True

☐ False

False

(b) (1 point) **True or False:** EM is the special case of Variational EM, where the true distribution is contained within the variational family.

☐ True

☐ False

True

## 10 Deep Generative Models

Some of these questions are out-of-scope: We did not cover Boltzmann Machines or Sigmoid Belief Networks or Contrastive Divergence.

1. (1 point) **True or False:** In a variational autoencoder the decoder  $p_{\phi}(\mathbf{x} \mid \mathbf{z})$  can always be characterized as a Gaussian where the mean and variance are given by a neural network with parameters  $\phi$ .

☐ True

☐ False

False

2. (1 point) **True or False:** The purpose of the reparameterization trick in a variational autoencoder is to obtain a polynomial factor speedup of the backpropagation algorithm.

☐ True

☐ False

False

3. (1 point) **Select all that apply:** For which of the following models can we in general compute the partition function in polynomial time?

☐ Restricted Boltzmann Machine

☐ Boltzmann Machine

☐ Sigmoid Belief Network

☐ Deep Boltzmann Machine

☐ None of the above

Only the Sigmoid Belief Network. It's a Bayesian network and so the partition function is 1.0.

4. Suppose we wish to learn a latent variable model  $p(\mathbf{v}, \mathbf{h}) = p(\mathbf{v} \mid \mathbf{h})p(\mathbf{h})$  where  $p(\mathbf{h}) \sim \mathcal{N}(\mathbf{0}, \Sigma)$  and  $p(\mathbf{v} \mid \mathbf{h}) \sim \mathcal{N}(\text{MLP}(\mathbf{h}, \theta), \mathbf{I})$ .  $\text{MLP}(\mathbf{h}, \theta)$  is a feed-forward neural network with  $D$  input units, three hidden layers, and  $M$  output units. In the training data we only observe the visible variables  $\mathbf{v} \in \mathbb{R}^M$ , not the hidden variables  $\mathbf{h} \in \mathbb{R}^D$ . Now you want to learn the parameters  $\theta$  and  $\Sigma$ .

- (a) (2 points) **Short answer:** You try training by contrastive divergence. Will this approach work? If so, briefly describe one iteration of training. If not, explain why not.

No, this would not work. You would need to compute the full conditional  $p(\mathbf{h}|\mathbf{v})$  in order to run the one step Gibbs sampler, but the full conditional is intractable to compute. because of the MLP.

Another valid solution would be that this does work and we can use contrastive divergence but instead of using a Gibbs sampler, we have to swap in some other suitable MCMC approach (e.g. Metropolis Hastings).

- (b) (2 points) **Short answer:** You try training by variational EM. Will this approach work? If so, briefly describe a suitable variational approximation. If not, explain why not.

Yes, this works fine and is much like any variational autoencoder. You could use a variational approximation  $q(\mathbf{v}, \mathbf{h}) = \prod_{i=1}^D q(v_i; \lambda_i) \prod_{j=1}^M q(h_j; \tau_j)$  where each  $q(v_i; \lambda_i)$  and  $q(h_j; \tau_j)$  is a Gaussian.

## 11 MAP Inference and MILP

1. (1 point) **True or False:** The simplex algorithm returns a global optimum of an integer linear program.

- ☐ True  
☐ False

**False.** Branch-and-bound is the standard algorithm used to solve integer linear programs.

2. (1 point) **Select all that apply:** Which of the following algorithms can be used to perform MAP inference?

- ☐ Integer Linear Programming  
☐ Belief Propagation  
☐ Variational Inference  
☐ None of the above

**All three (ILP, Belief Propagation, Variational Inference) can be used to perform MAP inference.**

3. (1 point) **True or False:** In general, MAP inference is NP-hard and marginal inference is #P-hard.

- ☐ True  
☐ False

**True**

4. Suppose we define a Bayesian Network over two random variables,  $X_1$  and  $X_2$  (as shown in figure 5) to generate binary strings of length 2. That is,  $X_1 \in \{0, 1\}$  and  $X_2 \in \{0, 1\}$ .

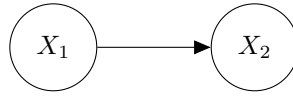


Figure 5: Bayesian Network

We further define a set of binary node indicator variables  $x_{i,c}$  for  $i \in \{1, 2\}$  and  $c \in \{0, 1\}$ . If  $x_{i,0} = 1$ , variable  $X_i$  is assigned to value 0. If  $x_{i,1} = 1$ , variable  $X_i$  is assigned to value 1. For this model, answer the following questions:

- (a) (2 points) **Short Answer:** Write a constraint to ensure that each variable  $X_i$  takes on only one value. (Note: Such constraints can also be called sanity constraints since they ensure that you do not get invalid configurations.)

$$\sum_c x_{i,c} = 1, \forall i \in \{1, 2\} \quad (13)$$

- (b) (3 points) **Short Answer:** Now we define a set of binary edge indicator variables  $y_{c,d}$  with  $c \in \{0, 1\}$  and  $d \in \{0, 1\}$  to represent the joint assignment of  $X_1$  and  $X_2$  in the Bayesian Network. That is,  $y_{c,d} = 1$  if  $X_1 = c$  and  $X_2 = d$ ,  $y_{c,d} = 0$  otherwise. Define any necessary sanity constraints and constraints to ensure that node and edge indicator variables maintain consistent assignments.

We will have a set of 4 binary indicator variables corresponding to 4 possible configurations i.e.  $y_{00}, y_{01}, y_{10}, y_{11}$ . These variables need to be subjected to the following constraints:

$$y_{00} + y_{01} + y_{10} + y_{11} = 1 \quad (14)$$

$$y_{ab} \leq x_{1,a} \quad (15)$$

$$y_{ab} \leq x_{2,b}, \forall a \in \{0, 1\}, b \in \{0, 1\} \quad (16)$$



- (c) (3 points) **Short Answer:** Assume that node  $X_1$  is governed by a Bernoulli distribution with parameter  $p$  (i.e.  $P(X_1 = 1) = p$ ). The conditional distribution over  $X_2$  given  $X_1$  is governed by the transition probability matrix  $Q$ , where  $P(X_2 = b|X_1 = a) = q_{ab}$  and

$$Q = \begin{bmatrix} q_{00} & q_{01} \\ q_{10} & q_{11} \end{bmatrix}$$

Given these probabilities, write the MAP inference problem as an ILP. (You need not repeat the constraints from previous parts, defining the objective is enough.)

Solve the objective:

$$\max \log p x_{1,1} + \log(1-p) x_{1,0} + \sum_{a=0}^1 \sum_{b=0}^1 y_{ab} * \log q_{ab} \quad (17)$$

## 12 Structured Perceptron and Structured SVM

- Let's look at building structured perceptron and structured SVM models for the POS tagging task (i.e. assigning a part-of-speech tag to every word in a given sequence). Assume that the set of POS tags for our task is  $\mathcal{Y} = \{N, V\}$ . Throughout this question, we will stick with the first-order Markov assumption i.e. the tag  $y_i$  for a word  $x_i$  is *only dependent* on the tag  $y_{i-1}$  for the previous word  $x_{i-1}$ . Additionally assume that each word  $x_i$  only influences the assignment of its own tag  $y_i$ .

A linguist defines some feature functions for you  $\phi_{wt}$  and  $\phi_{tt}$ . The linguist merely informs you that  $\phi_{wt}(x_i, y_i) \in \mathbb{R}$  is a feature value for when tag  $y_i$  is assigned to word  $x_i$  and that  $\phi_{tt}(y_{i-1}, y_i) \in \mathbb{R}$  is the feature value for when tag  $y_i$  follows tag  $y_{i-1}$  in the sequence.

You build a linear model using these feature functions. Assume that weights  $W_{wt}(x_i, y_i)$  and  $W_{tt}(y_{i-1}, y_i)$  are the weights corresponding to word-tag and tag pair assignments. Given this setup, the task of computing the best POS tag sequence can be formulated as a MAP inference problem with the following objective for a sequence of length  $N$ :

$$\arg \max_{y_1, \dots, y_N} \sum_{i=1}^N W_{wt}(x_i, y_i) \phi_{wt}(x_i, y_i) + \sum_{i=2}^N W_{tt}(y_{i-1}, y_i) \phi_{tt}(y_{i-1}, y_i) \quad (18)$$

Our word vocabulary is the set  $\mathcal{W} = \{\text{Dogs}, \text{fly}\}$ . Consider the following scoring functions and weights:

	$N$	$V$
Dogs	2	1
fly	1	1

(a)  $\phi_{wt}$ 

	$N$	$V$
$N$	0	1
$V$	1	-1

(b)  $\phi_{tt}$ 

	$N$	$V$
Dogs	1	1
fly	1	1

(a)  $W_{wt}$ 

	$N$	$V$
$N$	0	1
$V$	1	0

(b)  $W_{tt}$ 

Read entries in the  $\phi_{tt}$  and  $W_{tt}$  tables as  $(y_{i-1}, y_i)$ . For example, the element at position  $(N, V)$  refers to the score for the assignment  $y_{i-1} = N, y_i = V$ .

Finally, suppose that we want to tag the sentence  $\mathbf{x} = [x_1, x_2]$ : “Dogs fly” with ground truth tagging  $\mathbf{y} = [y_1, y_2] = [N, V]$ .

- (a) (2 points) **Numerical Answer:** Compute the scores of all possible output tag sequences under MAP inference

Sequence	Score
NN	
NV	
VN	
VV	

Sequence	Score
NN	3
NV	4
VN	3
VV	2

- (b) (2 points) **Numerical Answer:** Compute the scores of all possible output tag sequences under Loss-Augmented inference. Use unnormalized Hamming loss.

Sequence	Score
NN	
NV	
VN	
VV	

Sequence	Score
NN	4
NV	4
VN	5
VV	3

2. Recall that the update rule for structured perceptron and structured SVM can be written as:

$$\mathbf{w} \leftarrow \mathbf{w} + \phi(\mathbf{x}, \mathbf{y}^*) - \phi(\mathbf{x}, \hat{\mathbf{y}}) - \lambda \mathbf{w} \quad (19)$$

where  $\mathbf{w} = (W_{wt}, W_{tt})$  is the set of weights,  $\mathbf{y}^*$  is the correct sequence of tags,  $\hat{\mathbf{y}}$  is the sequence of tags returned by the appropriate inference method and  $\phi$  is our feature function. For the structured perceptron case,  $\lambda = 0$ .

- (a) (1 point) **Select One:** Which tag sequence will be chosen as  $\hat{\mathbf{y}}$  during structured perceptron training?

☐ NN

☐ NV

☐ VN

☐ VV

$(N, V)$  using the best scoring structure under MAP inference

(b) (1 point) **Select One:** Which tag sequence will be chosen as  $\hat{\mathbf{y}}$  during structured SVM training?

☐ NN

☐ NV

☐ VN

☐ VV

$(V, N)$  using the best scoring structure under LAI

3. (1 point) **Select all that apply:** Consider a POS-tagging model with a tag set of size 2, i.e.  $\{0, 1\}$ . which of the following loss functions can be used as a replacement for hamming loss in loss-augmented inference?

- ☐ Elementwise Cross-Entropy Loss
- ☐ Elementwise Squared Error
- ☐ Edit Distance
- ☐ None of the above

Elementwise MSE, Edit Distance since both can compute distance between optimal assignments in label space

## 13 Bayesian Nonparametrics

We did not cover the stick-breaking construction or the Indian Buffet Process (IBP); they are out of scope.

1. (1 point) **True or False:** Bayesian nonparametric models *do* have parameters, but the *number* of parameters in use can grow or shrink with the dataset.

☐ True

☐ False

True

2. (1 point) **True or False:** If a distribution is independent and identically distributed, then it is also exchangeable.

☐ True

☐ False

True. (The opposite is not true: exchangeable does not imply i.i.d.)

3. (2 points) **Select all that apply:** Which of the following are constructions of the Dirichlet Process?

☐ Chinese Restaurant Franchise

☐ Stick breaking construction

☐ Chinese Restaurant Process

☐ Dirichlet process mixture model

☐ Polya Urn Scheme

☐ None of the above

Stick breaking construction, Chinese Restaurant Process, Polya Urn Scheme

4. (1 point) **Short answer:** Describe one modeling advantage that a (Gaussian) Dirichlet Process Mixture Model (DPMM) has over a Gaussian Mixture Model (GMM).

---

---

---

The DPMM can adaptively learn the appropriate number of clusters, whereas the GMM requires the number of clusters to be set as a hyperparameter.

5. (1 point) **True or False:** Suppose we define a distribution over binary matrices  $p(\mathbf{Z} \mid \alpha)$  that follows the exchangeable Indian Buffet Process with Poisson strength parameter  $\alpha$ . To draw a sample left-ordered-form matrix  $\mathbf{Z}$  from this distribution, you must resort to approximate inference techniques such as Gibbs sampling or variational inference.

☐ True

☐ False

False. Approximate inference would only be necessary if the model incorporated some observed variables.

6. (1 point) **Select all that apply:** Now suppose we define a distribution over binary matrices  $p(\mathbf{Z} \mid \alpha)$  that follows the exchangeable Indian Buffet Process with Poisson strength parameter  $\alpha$  and where  $\mathbf{Z}$  has  $N$  rows. Given that  $\alpha + N = 6$ , which of the following would lead to the greatest number of non-zero entries in expectation? (You may select multiple if there is a tie.)

☐  $\alpha = 1, N = 5$

☐  $\alpha = 2, N = 4$

☐  $\alpha = 3, N = 3$

☐  $\alpha = 4, N = 2$

☐  $\alpha = 5, N = 1$

☐ None of the above

$\alpha = 3, N = 3$  since the number of expected ones is  $\alpha N$

7. **Numerical answer:** Suppose you have a sample from a Chinese Restaurant Process with strength parameter  $\alpha = 4$ , as shown in Figure 6. The tables are numbered circles. There are three customers at table 1, two at table 2, and one at table 3.

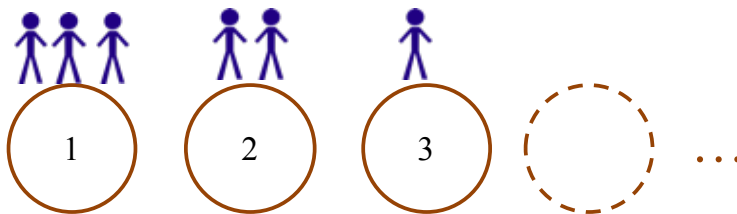


Figure 6

- (a) (1 point) What is the probability of the next (i.e. 7th) customer sitting at table 1?

3/10

- (b) (1 point) What is the probability of the next (i.e. 7th) customer sitting at table 3?

1/10

- (c) (1 point) What is the probability of the next (i.e. 7th) customer sitting at a new table?

4/10