Problem 1: Statistics Review (20 points)

1. **(10 points)** Recall the Dirichlet distribution, a family of continuous multivariate probability distribution used as prior to the multinomial distribution, with the PDF specified as:

$$f(\mathbf{x};\alpha) = \frac{\Gamma(\sum_{i=1}^{K}\alpha_i)}{\prod_{i=1}^{K}\Gamma(\alpha_i)}\prod_{i=1}^{K}x_i^{\alpha_i-1},$$

where $\Gamma(x)$ is the Gamma function with derivative $\Gamma(x)' = \phi(x)$, the digamma function. Derive the maximum likelihood estimator for parameter α for the Dirichlet distribution.

2. **(10 points)** Use the *pdf kernel method* to derive the posterior distribution the following conjugate pair: prior $p(\theta) = \mathcal{N}(\theta; 0, I)$ and likelihood function $p(x;\theta) = \mathcal{N}(x;\mu,\Sigma)$. Show all steps.

Answer: Q1

To derive the maximum likelihood estimator (MLE) for the parameter vector "α" in the Dirichlet distribution, need to maximize the log-likelihood function. The log-likelihood function for the Dirichlet distribution is given by:

$$\mathrm{L}(\alpha) = \log(\mathrm{x};\alpha) = \log\Gamma\sum_{i=1}^{K}a_i - \sum_{i=1}^{K}log\Gamma(a_i) + \sum_{i=1}^{K}(a_i-1)\log(x_i)$$

To find the MLE, need to differentiate the log-likelihood function with respect to each parameter a_i and set the derivatives equal to zero:

$$\frac{\partial L(\alpha)}{\partial\alpha_i} = \frac{\partial(\log\left(\Gamma\sum_{i=1}^{K}\alpha_i\right) - \sum_{i=1}^{K}log\Gamma(\alpha_i) + \sum_{i=1}^{K}(\alpha_i-1)\log(x_i))}{\partial\alpha_i}$$

$$= \phi\left(\sum_{i=1}^{K}\alpha_i\right) - \phi(\alpha_i) + \log(\alpha_i) = 0$$

where $\phi(\mathrm{x})$ is the digamma function, the derivative of the log of the gamma function. Rearranging the equation, obtain:

$$\phi(\alpha_{\mathrm{i}}) - \phi\left(\sum_{i=1}^{K}\alpha_i\right) = \log(\alpha_i)$$

This equation does not have a closed-form solution, so typically solve it iteratively using numerical methods such as Newton's method or gradient descent. The iterative procedure involves starting with an initial guess for α and updating it until convergence is achieved.

During each iteration, compute the digamma function values and the log of the data points. Then, update the parameter vector α using the equation:

$$\alpha_i^{t+1} = \alpha_i^t + \phi_i^t - \phi\left(\sum_{i=1}^{K}\alpha_i^t\right) + \log(x_i)$$

where t is the iteration number.

repeatting this iterative process until the parameter estimates converge. The resulting parameter vector α will be the maximum likelihood estimator for the Dirichlet distribution.

Answer: Q2

To derive the posterior distribution using the PDF kernel method for the conjugate pair consisting of a Gaussian prior and Gaussian likelihood, follow these steps:

1. Prior Distribution: $\mathrm{p}(\theta) = \mathrm{N}(\theta; 0, \mathrm{I})$

   This is a multivariate Gaussian distribution with mean 0 and identity covariance matrix I.

2. Likelihood Function: $\mathrm{p(x|\theta)} = N(x; \mu, \sigma)$
   This is a multivariate Gaussian distribution with mean μ and covariance matrix σ.

3. Compute the joint distribution: : $\mathrm{p(x, \theta)} = \mathrm{p(x|\theta)} * \mathrm{p(\theta)}$
   Since the prior and likelihood are independent, can multiply them to obtain the joint distribution.

$$\mathrm{p(x, \theta)} = (x; \mu, \sigma) * N(\theta; 0, \mathrm{I})$$

4. Apply the pdf kernel method:
   use the properties of Gaussian distributions to simplify the joint distribution.

$$\mathrm{p(x, \theta)} = \frac{1}{(2\pi)^{d/2} \times |\sigma|^{0.5}} \exp(-0.5 \times (x - \mu)^T \times \sigma^{-1} \times (x - \mu)) \times \frac{1}{(2\pi)^{\frac{k}{2}} \times |I|^{0.5}} \exp(-0.5 \times \theta^T * \theta)$$

   where *d* is the dimensionality of x, and *k* is the dimensionality of θ.

5. Simplify the joint distribution: Combining the terms, obtain:

$$\mathrm{p(x, \theta)} = \frac{1}{(2\pi)^{d/2} \times |\sigma|^{0.5} \times (2\pi)^{\frac{k}{2}} \times |I|^{0.5}} \exp(-0.5 \times (x - \mu)^T \times \sigma^{-1} \times (x - \mu) - 0.5 \times \theta^T * \theta)$$

$$= \frac{1}{(2\pi)^{\frac{d}{2}+\frac{k}{2}} \times |\sigma|^{0.5} \times |I|^{0.5}} \exp(-0.5 \times (x - \mu)^T \times \sigma^{-1} \times (x - \mu) - 0.5 \times \theta^T * \theta)$$

6. Compute the posterior distribution:
   To obtain the posterior distribution, need to normalize the joint distribution by dividing it by the marginal likelihood, which acts as a normalization constant. The marginal likelihood can be obtained by integrating the joint distribution over all possible values of θ.

$$\mathrm{p(\theta|x)} = \frac{p(x, \theta)}{\int p(x, \theta) d\theta}$$

   The denominator is the normalization constant that ensures the posterior distribution integrates to 1.

   By simplifying and normalizing the joint distribution, obtain the posterior distribution.

Problem 2: Topic Models (30 points)

1. **(5 points)** Discuss the similarities and differences between the multinomial mixture model and the mixture of unigrams model.
2. **(15 points)** Derive the following quantities in the pLSA model and show all your steps.
   (a) The joint density of all random variables $p(w, d, z; \theta, \beta)$.
   (b) The conditional density $p(z \mid w, d; \theta, \beta)$.
   (c) The conditional density of $p(w \mid z, d; \theta, \beta)$.

   Note: in this notation, θ, β are given as model hyperparameters, hence they are separated from the random variable using ';'.
3. **(10 points)** Write the log-likelihood function of the pLSA model as sum of two terms: the Evidence Lower Bound (ELBO) and the KL divergence between a variational distribution and the latent posterior. Point out each term in your answer.

Answer: Q1

The multinomial mixture model and the mixture of unigrams model are two different approaches used in text modeling, specifically in the context of document modeling.

Similarities:

Probability-based modeling: Both models are probabilistic models that aim to capture the distribution of words in documents.

Mixture modeling: Both models assume that each document is a mixture of different components (topics or word distributions).

Differences:

Representation of documents: Multinomial Mixture Model: In this model, documents are represented using a bag-of-words approach, where the document is represented as a vector of word counts or frequencies. Mixture of Unigrams Model: This model represents documents as a sequence of words, with each word being generated independently from a mixture of word distributions.

Modeling approach: Multinomial Mixture Model: It assumes that each document is generated by choosing a single mixture component (topic or word distribution) and then generating words from that component according to a multinomial distribution.

Mixture of Unigrams Model: It assumes that each word in a document is generated independently by sampling a mixture component (word distribution) and then generating the word from that component's distribution.

Complexity and flexibility: Multinomial Mixture Model: This model is more complex and flexible as it allows for a more nuanced representation of documents by modeling the joint probability of words and topics. It can capture dependencies between words and topics.

Mixture of Unigrams Model: This model is simpler and less flexible compared to the multinomial mixture model. It assumes independence between words in a document and does not capture word co-occurrence patterns or dependencies between words and topics.

Inference and training: Multinomial Mixture Model: Inference in this model typically involves techniques like Expectation-Maximization (EM) or variational inference to estimate the latent variables (topics or mixture components) and their distributions. Training involves optimizing the model parameters to maximize the likelihood of the observed data.

Mixture of Unigrams Model: Inference in this model can be performed using techniques like the EM algorithm or Gibbs sampling. Training involves estimating the mixture weights and

word distributions.

Answer: Q2(a)

In the probabilistic Latent Semantic Analysis (pLSA) model, have three random variables: w (word), d (document), and z (topic). The model assumes a generative process where each word in a document is associated with a latent topic. To derive the joint density of all random variables $p(w, d, z; \theta, \beta)$, have follow steps:

1. Decompose the joint density using the chain rule of probability:

$$p(w, d, z; \theta, \beta) = p(w|d, z; \theta, \beta) * p(d, z; \theta, \beta)$$

2. Apply the conditional independence assumption:

assume that the word w is conditionally independent of the document d given the topic z. Therefore,

$$p(w, d, z; \theta, \beta) = p(w|z; \beta) * p(d, z; \theta)$$

3. Express the conditional probabilities using the model parameters:

$p(w|z; \beta)$ is the probability of word w given topic z, which is parameterized by the matrix β.

$p(d, z; \theta)$is the probability of document d and topic z, which is parameterized by the matrix θ.

4. Incorporate the topic mixture proportion:

assume that the topic z is generated from a multinomial distribution with topic mixture proportions given by the matrix θ. Therefore, express $p(d, z; \theta)$ as the product of $p(d|z; \theta)$ and $p(z; \theta)$.

5. Express $p(d|z; \theta)$ and $p(z; \theta)$ using the model parameters:

$p(d|z; \theta)$ is the probability of document d given topic z, which is parameterized by the matrix θ.

$p(z; \theta)$ is the probability of topic z, which is parameterized by the matrix θ.

6. Putting it all together,

$$p(w, d, z; \theta, \beta) = p(w|z; \beta) * p(d|z; \theta) * p(z; \theta)$$

This represents the joint density of all random variables in the pLSA model.

note that the pLSA model assumes the parameter matrices β and θ are learned from the training data using methods such as expectation-maximization (EM) algorithm or variational inference.

Answer: Q2(b)

To derive the conditional density $p(z|w, d; \theta, \beta)$ in the pLSA model, use Bayes' theorem and the joint density of all random variables $p(w, d, z; \theta, \beta)$. the steps are:

1. Apply Bayes' theorem:

According to Bayes' theorem,

$$p(z|w, d; \theta, \beta) = \frac{p(w, d, z; \theta, \beta)}{p(w, d; \theta, \beta)}$$

2. Express the joint density:

From the previous derivation, have:

$$p(w, d, z; \theta, \beta) = p(w|z; \beta) * p(d|z; \theta) * p(z; \theta)$$

3. Compute the marginal density $p(w, d; \theta, \beta)$:

To compute the marginal density, need to sum/integrate the joint density over all possible values

of z:

$$p(w, d; \theta, \beta) = \sum_{z} p(w|z; \beta) * p(d|z; \theta) * p(z; \theta)$$

4. Substitute the expressions:

$$p(z|w, d; \theta, \beta) = \frac{[p(w|z; \beta) * p(d|z; \theta) * p(z; \theta)]}{\sum_z p(w|z; \beta) * p(d|z; \theta) * p(z; \theta)}$$

5. Simplify the expression:

Depending on the specific form of the pLSA model and the assumptions made, further simplifications may be possible. However, in general, the conditional density $p(z|w, d; \theta, \beta)$ is a complex expression that involves the model parameters θ and β, as well as the observed word w and document d.

Answer: Q2(c)

To derive the conditional density $p(w|z, d; \theta, \beta)$ in the pLSA model, can use Bayes' theorem and the joint density of all random variables $p(w, d, z; \theta, \beta)$. Here are the steps:

1. Apply Bayes' theorem:

According to Bayes' theorem, have:

$$p(w|z, d; \theta, \beta) = \frac{p(w, d, z; \theta, \beta)}{p(d, z; \theta, \beta)}$$

2. Express the joint density:

From the previous derivation, have:

$$p(w, d, z; \theta, \beta) = p(w|z; \beta) * p(d|z; \theta) * p(z; \theta)$$

3. Compute the marginal density $p(z, d; \theta, \beta)$:

To compute the marginal density, need to sum the joint density over all possible values of w:

$$p(z, d; \theta, \beta) = \sum_{w} p(w|z; \beta) * p(d|z; \theta) * p(z; \theta)$$

4. Substitute the expressions:

$$p(w|z, d; \theta, \beta) = \frac{[p(w|z; \beta) * p(d|z; \theta) * p(z; \theta)]}{\sum_w p(w|z; \beta) * p(d|z; \theta) * p(z; \theta)}$$

5. Simplify the expression:

Depending on the specific form of the pLSA model and the assumptions made, further simplifications may be possible. However, in general, the conditional density $p(w|z, d; \theta, \beta)$ is a complex expression that involves the model parameters θ and β, as well as the observed topic z and document d.

Answer: Q3

In the pLSA model, the log-likelihood function can be expressed as the sum of two terms: the Evidence Lower Bound (ELBO) and the Kullback-Leibler (KL) divergence between a variational distribution and the latent posterior. Here is the breakdown of each term:

Evidence Lower Bound (ELBO):

The ELBO is a lower bound on the log-likelihood and is used in variational inference to approximate the true posterior distribution. In the pLSA model, the ELBO term can be written as:

$$\text{ELBO} = \sum_{d}\sum_{w} n(d,w)\log(p(d,w;\theta,\beta))$$

where $n(d,w)$ represents the count of word w in document d, and $p(d,w;\theta,\beta)$ denotes the joint probability of document d and word w given the model parameters θ and β.

KL Divergence between Variational Distribution and Latent Posterior:

In variational inference, a variational distribution $q(z|d,w)$ is introduced to approximate the true but intractable posterior distribution $p(z|d,w;\theta,\beta)$. The KL divergence measures the difference between these two distributions and is given by:

$$\text{KL}(q(z|d,w)||\, p(z|d,w;\theta,\beta))$$

This term quantifies how well the variational distribution $q(z|d,w)$ approximates the true latent posterior $p(z|d,w;\theta,\beta)$.

Hence, the log-likelihood function of the pLSA model can be written as the sum of the ELBO and the KL divergence term:

$$\log-\text{likelihood} = \text{ELBO} - \text{KL}(q(z|d,w)||\, p(z|d,w;\theta,\beta))$$

note that the specific forms of the ELBO and the KL divergence term depend on the assumptions and parameterizations used in the pLSA model and the variational inference approach employed.