In The Name of God

Sharif University of Technology

Electrical Engineering Department

Deep Generative Models

Assignment 4

Fall 2024

Instructor: Dr. S. Amini

1 Question: Understanding Noise Contrastive Estimation (NCE)

You are familiar with Contrastive Divergence (CD) as an approximate method for training undirected graphical models, such as Restricted Boltzmann Machines (RBMs). Noise Contrastive Estimation (NCE) is another method used to train probabilistic models. In this question, you will gradually explore NCE and its relationship to CD.

Part 1: Recall of Contrastive Divergence (CD)

1.1 Define Contrastive Divergence (CD) and describe how it is used in the training of a probabilistic model, such as a Restricted Boltzmann Machine (RBM). What is the objective of the CD algorithm?

Part 2: Introducing Noise Contrastive Estimation (NCE)

2.1 Noise Contrastive Estimation (NCE) is another approach for estimating the parameters of a probabilistic model. In the case of an RBM, explain what you think would be the challenge if you tried to use CD to estimate the model's parameters without having access to the full partition function Z.

2.2 NCE tries to overcome the difficulty of estimating the partition function. How does NCE propose to estimate the model's parameters without explicitly calculating the partition function? Describe the basic idea behind NCE in your own words.

Part 3: Formulating NCE

3.1 In NCE, the objective is to distinguish between samples drawn from the true data distribution $p_{\text{data}}(x)$ and samples drawn from a noise distribution $q(x)$. What is the role of the noise distribution in NCE? Why is it important to choose a suitable noise distribution?

3.2 Suppose we have a set of data samples $x_1, x_2, \ldots, x_n$ and a noise distribution $q(x)$. How would you formulate the objective function for NCE?

Part 4: NCE vs. CD

4.1 Compare the differences between Contrastive Divergence (CD) and Noise Contrastive Estimation (NCE) in terms of their goals, learning objectives, and computational requirements.

4.2 In which situations might Noise Contrastive Estimation (NCE) be preferred over Contrastive Divergence (CD)? Discuss any advantages NCE might offer over CD.

Part 5: Practical Use of NCE

5.1 Describe an example application where Noise Contrastive Estimation (NCE) might be used in practice. Why would NCE be useful in this scenario?
Hint: One example could be training large-scale language models, where the goal is to estimate the probability distribution over words. Here, the partition function is difficult to compute, and NCE can be used to estimate the model's parameters efficiently by distinguishing between real words and noise.

2 Question: Understanding Score Functions and Their Applications

In this question, you will explore the concept of score functions, their role in statistical inference, and how they relate to specific probability distributions.

Part 1: Theoretical Foundations of Score Functions

1.1 Define the score function. How is the score function related to the likelihood function? Write the mathematical expression for the score function.

1.2 Discuss the importance of the score function in the context of Maximum Likelihood Estimation (MLE). How does the score function help in finding the maximum likelihood estimate?

Part 2: Score Function for Specific Distributions

2.1 Consider the probability density function (PDF) of a Normal distribution $\mathcal{N}(\mu, \sigma^2)$, given by:

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

Find the score function for the mean μ and the variance σ^2 of a Normal distribution.

2.2 For an Exponential distribution with parameter λ, the PDF is given by:

$$f(x; \lambda) = \lambda \exp(-\lambda x), \quad x \geq 0$$

Find the score function for λ.

Part 3: Properties of Score Functions

3.1 Prove that the expected value of the score function is zero, i.e.,

$$\mathbb{E}[s(\theta; X)] = 0$$

3.2 Prove or explain why the Fisher Information, defined as the variance of the score function, is always non-negative:

$$I(\theta) = \mathbb{E}[s(\theta; X)^2]$$

Part 4: Application of Score Functions in Statistical Inference

4.1 Discuss how score functions are used in hypothesis testing. In particular, explain the score test (or Lagrange Multiplier test) and its relationship to the likelihood ratio test.

4.2 Score functions also play a key role in asymptotic analysis. Describe how the score function is used to derive the asymptotic normality of Maximum Likelihood Estimators (MLEs).

$$\hat{\theta}_{\text{MLE}} \xrightarrow{d} \mathcal{N}(\theta, I(\theta)^{-1})$$

4.3 In the case of the Normal distribution, use the score function to derive an asymptotic confidence interval for μ.

3 Question: Implicit and Explicit Score Matching in Generative Models

In this question, we explore the relationship between implicit and explicit score matching and demonstrate why these two loss functions are equivalent objectives when training generative model parameters.

Part 1: Theoretical Proof of Equivalence between Implicit and Explicit Score Matching

1.1 Define the **explicit score matching** loss and the **implicit score matching** loss. Suppose you have a generative model with a parameterized density $p_\theta(x)$, where θ are the model parameters.

Hint: The explicit score matching objective involves the gradient of the log-likelihood of the model's distribution with respect to the data. The loss for explicit score matching is given by:

$$\mathcal{L}_{\text{explicit}}(\theta) = \mathbb{E}_{p_{\text{data}}(x)} \left[\|\nabla_x \log p_\theta(x) - \nabla_x \log p_{\text{data}}(x)\|^2 \right]$$

The implicit score matching objective involves the gradient of the model's score function. The loss for implicit score matching is:

$$\mathcal{L}_{\text{implicit}}(\theta) = \mathbb{E}_{p_{\text{data}}(x)} \left[\|\nabla_x \log p_\theta(x)\|^2 \right]$$

1.2 Show that the explicit and implicit score matching objectives are equivalent in terms of the generative model parameters. Derive this equivalence step by step.

Hint: The key to the proof lies in understanding that the gradient of the log-likelihood with respect to the model's parameters is closely related to the score function. Start by writing the objective for explicit score matching and use the fact that $\nabla_x \log p_\theta(x)$ is the score of the model. By expanding both losses and leveraging the properties of gradients, you can show that the two objectives are equivalent.

Part 2: Example of Equivalence in a Simple Case

2.1 Consider a simple generative model where $p_\theta(x)$ is a Gaussian distribution with mean μ and variance σ^2, i.e., $p_\theta(x) = \mathcal{N}(x; \mu, \sigma^2)$. Write out the explicit and implicit score matching objectives for this case.

Hint: The Gaussian distribution is parameterized by μ and σ^2. For simplicity, assume you are optimizing over μ, and use the gradient of the log-likelihood for both score matching losses.

2.2 Show that for this Gaussian case, the explicit and implicit score matching objectives lead to the same optimal value of μ.

Hint: By setting the gradient of the explicit score matching loss with respect to μ equal to zero, you can show that both methods ultimately result in the same estimator for μ, which is the sample mean of the data.

4 Question: Proof of Gaussian Parameters in the Encoder of a Generative Model

In a generative model with a forward diffusion process, the conditional distribution $q(x_{t-1} \mid x_t, x_0)$ is given as a Gaussian:

$$q(x_{t-1} \mid x_t, x_0) = \mathcal{N}(x_{t-1} \mid \mu_q(x_t, x_0, t), \Sigma_q(t)),$$

where the parameters are defined as:

1. The mean:
$$\mu_q(x_t, x_0, t) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})x_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)x_0}{1-\bar{\alpha}_t}.$$

2. The covariance:
$$\Sigma_q(t) = \sigma_q^2(t)\mathbf{I},$$

where:
$$\sigma_q^2(t) = \frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}.$$

Proof

1. Prove that $q(x_{t-1})$ is a Gaussian distribution. Specifically:
   (a) Start with the forward diffusion process and its associated Gaussian properties.
   (b) Derive the conditional distribution $q(x_{t-1} \mid x_t, x_0)$ using the laws of Gaussian distributions.
   (c) Show that the resulting distribution is indeed Gaussian with the given mean $\mu_q(x_t, x_0, t)$ and covariance $\Sigma_q(t)$.
2. Provide a detailed mathematical proof of the expressions for $\mu_q(x_t, x_0, t)$ and $\Sigma_q(t)$.
3. Explain intuitively why the combination of the forward process conditionals leads to this specific parameterization.

Hint

To prove Gaussianity:

- Use the property that the conditional distribution of a joint Gaussian is also Gaussian. Apply this to the variables x_t, x_{t-1}, and x_0.