

CS294-158 Deep Unsupervised Learning

April 2024

This document contains all the possible exam questions and the answers for CS294-158 Spring 2024. The exam will be closed-book. There will be no clarification questions during the exam, because this handout already clarifies everything and is yours to study.

Happy studying!

1 Autoregressive Models

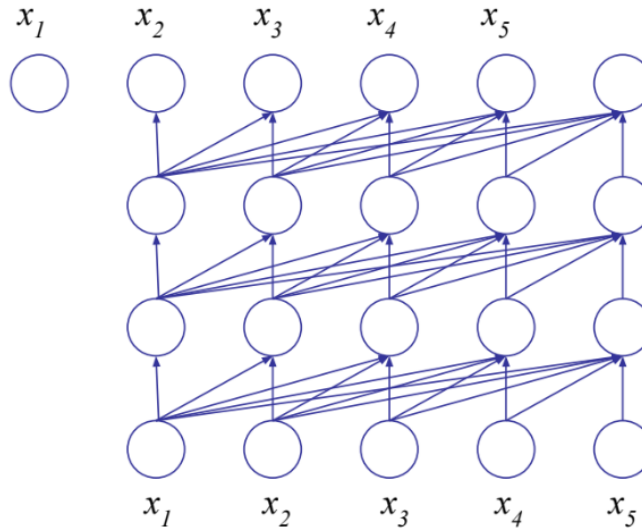
Q: (a) Write out the chain rule for a distribution over $x_1, x_2, \dots, x_d$.
 (b) Draw an example of a Masked Autoregressive Distribution Estimation (MADE) model + describe what makes this a proper probability model.
 (c) Describe 4 key enhancements modern Causal Masked Neural Models have beyond MADE.

A:

(a)

$$p(x_1, x_2, \dots, x_d) = p(x_1)p(x_2 | x_1)p(x_3 | x_1, x_2) \dots p(x_d | x_1, x_2, \dots, x_{d-1})$$

(b)



The connections are such that for output $\hat{x}_i$ only inputs with index $1, \dots, i-1$ can feed into it.

(c) (1) every node is vector valued; (2) parameter sharing; (3) coordinate coding;
 (4) transformer architecture

2 Flows

- Q:** (a) Describe the training objective for flows.
 (b) Describe sampling for flows.
 (c) What properties does the function f_θ have to satisfy for practical flows?

A:

(a)

$$x = f_\theta^{-1}(z)$$

$$z = f_\theta(x)$$

$$\max_{\theta} \sum_i \log p_\theta(x^{(i)}) = \max_{\theta} \sum_i \log p_Z(f_\theta(x^{(i)})) + \log \left| \frac{\partial f_\theta}{\partial x}(x^{(i)}) \right|$$

(b)



Step 1: sample $z \sim p_Z(z)$

Step 2: $x = f_\theta^{-1}(z)$

- (c) f_θ needs to (i) be differentiable, (ii) be invertible, (iii) have a tractable Jacobian determinant

3 VAE

Q: (a) Draw the basic discrete latent variable model and write out the log-likelihood objective.
 (b) Derive the Variational Lower Bound (VLB) using Jensen's Inequality.
 (c) To (approximately) optimize the VLB, which distribution do we sample z from?

A:

(a)



$$\max_{\theta} \sum_i \log p_{\theta}(x^{(i)}) = \sum_i \log \sum_z p_Z(z) p_{\theta}(x^{(i)} | z)$$

(b)

$$\begin{aligned} \max_{\theta} \sum_i \log p_{\theta}(x^{(i)}) &= \max_{\theta} \sum_i \log \left(\sum_z p_Z(z) p_{\theta}(x^{(i)} | z) \right) \\ &= \max_{\theta} \sum_i \log \left(\sum_z \frac{q(z | x^{(i)})}{q(z | x^{(i)})} p_Z(z) p_{\theta}(x^{(i)} | z) \right) \\ &\geq \max_{\theta} \sum_i \mathbb{E}_{q(z | x^{(i)})} \left[\log p_Z(z) - \log q(z | x^{(i)}) + \log p_{\theta}(x^{(i)} | z) \right] \end{aligned}$$

(c) $q(z | x^{(i)})$

4 GANs

- Q:** (a) Write out the original GAN min-max formulation.
 (b) Write out the non-saturating formulation.
 (c) Write out the Inception Score.
 (d) What's the Bayes optimal discriminator?

A:

(a)

$$\min_G \max_D \mathbb{E}_{x \sim p_{\text{data}}} [\log D(x)] + \mathbb{E}_{z \sim p(z)} [\log(1 - D(G(z)))]$$

(b)

$$\max_D L^{(D)} = \max_D \mathbb{E}_{x \sim p_{\text{data}}} [\log D(x)] + \mathbb{E}_{z \sim p(z)} [\log(1 - D(G(z)))]$$

$$\min_G L^{(G)} = \min_G -\mathbb{E}_{z \sim p(z)} \log(D(G(z)))$$

(d)

$$IS(x) = \exp(\mathbb{E}_{x \sim p_g} [D_{KL}(p(y | x) || p(y))])$$

(d)

$$D^*(x) = \frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_g(x)}$$

5 Diffusion Models

- Q:** (a) Describe training for Denoising Diffusion Probabilistic Model (DDPM)
 (b) What additional term is used when using classifier-guidance during inference?
 (c) Derive what additional terms are used when using classifier-free guidance?

A:

(a)

Algorithm 1 Training

```

1: repeat
2:    $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ 
3:    $t \sim \text{Uniform}(\{1, \dots, T\})$ 
4:    $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 
5:   Take gradient descent step on
        $\nabla_{\theta} \|\epsilon - \epsilon_{\theta}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, t)\|^2$ 
6: until converged
  
```

(b)

$$+w\nabla_x \log p(y | x)$$

(c)

$$\begin{aligned}
 w\nabla_x \log p(y | x) &= w\nabla_x \log \frac{p(x|y)p(y)}{p(x)} \\
 &= w\nabla_x \log p(x | y) + w\nabla_x p(y) - w\nabla_x \log p(x) \\
 &= w\nabla_x \log p(x | y) + 0 - w\nabla_x \log p(x)
 \end{aligned}$$

6 Contrastive Learning

Q: (a) What is the InfoNCE loss, as presented in the CPC paper?
(b) What are the positive and negative samples in the loss and what role do they play?

A:

(a)

$$\mathbb{L}_N = -\mathbb{E}[\log \frac{f(x, c)}{\sum_{x' \in X} f(x', c)}]$$

where f is some positive score.

(b) c is the context vector for the positive pair (x, c) . $(x', c) \forall x' \neq x$ are negative samples. Minimizing the InfoNCE loss maximizes the log probability of correctly classifying the positive pair, out of all possible pairs.

7 Compression Fundamentals

Q: (a) What's the Kraft-McMillan Inequality? (both parts!) And what's an important consequence?

(b) State and prove the Shannon coding theorem (1948).

(c) Prove we can have a prefix code with code lengths $l_i = \lceil \log_2 \frac{1}{p(x_i)} \rceil$?

(d) What's the expected code length for such code?

A:

(a) Part I: For any uniquely decodable code C, we have that:

$$\sum_{(s,w) \in C} 2^{-l(w)} \leq 1$$

where $l(w)$ is the length of codeword w for symbol s .

Part II: For any set of lengths L , if

$$\sum_{l \in L} 2^{-l} \leq 1$$

then there is a prefix code C of the same size such that $l(w_i) = l_i$ for all i

Important Consequence: For any uniquely decodable code C, there is always a prefix code with the same codeword lengths

(b) For any message distribution $P(X)$ and associated uniquely decodable code C :

$$H(X) \leq l_a(C)$$

Proof:

$$\begin{aligned} H(X) - l_a(C) &= \sum_i p(x_i) \log_2 \frac{1}{p(x_i)} - \sum_i p(x_i) l_i \\ &= \sum_i p(x_i) \left(\log_2 \frac{1}{p(x_i)} - l_i \right) \\ &= \sum_i p(x_i) \left(\log_2 \frac{1}{p(x_i)} - \log_2 2^{l_i} \right) \\ &= \sum_i p(x_i) \log_2 \frac{2^{-l_i}}{p(x_i)} \\ &\leq \log_2 \left(\sum_i 2^{-l_i} \right) \\ &\leq \log_2 (1) \\ &= 0 \end{aligned}$$

(c) Yes, because such code satisfies the Kraft-McMillan inequality, which can

be proven as follows:

$$\begin{aligned}
\sum_i 2^{-l_i} &= \sum_i 2^{-\lceil \log_2 \frac{1}{p(x_i)} \rceil} \\
&\leq \sum_i 2^{-\log_2 \frac{1}{p(x_i)}} \\
&= \sum_i p(x_i) \\
&= 1
\end{aligned}$$

(d)

$$\begin{aligned}
l_a &= \sum_i p(x_i) l_i = \sum_i p(x_i) \lceil \log \frac{1}{p(x_i)} \rceil \\
&\leq \sum_i p(x_i) (1 + \log \frac{1}{p(x_i)}) \\
&= 1 + \sum_i p(x_i) \log \frac{1}{p(x_i)} \\
&= 1 + H(X)
\end{aligned}$$