Data Analytics & Machine Learning
Informatics
Technical University of Munich

Eexam
Place student sticker here

Note:

- During the attendance check a sticker containing a unique code will be put on this exam.
- This code contains a unique number that associates this exam with your registration number.
- This number is printed both next to the code and to the signature field in the attendance check list.

Advanced Machine Learning: Deep Generative Models

Exam:	CIT4230003 / Endterm	**Date:**	Monday 31st July, 2023
Examiner:	Prof. Dr. Stephan Günnemann	**Time:**	13:30 – 14:30

	P 1	P 2	P 3	P 4
I				

Working instructions

- This exam consists of **12 pages** with a total of **4 problems**.
  Please make sure now that you received a complete copy of the exam.
- The total amount of achievable credits in this exam is 28 credits.
- Detaching pages from the exam is prohibited.
- Allowed resources:
  - one A4 sheet of handwritten notes, two sides.
- **No other material (e.g. books, cell phones, calculators) is allowed!**
- Physically turn off all electronic devices, put them into your bag and close the bag.
- There is scratch paper at the end of the exam.
- Write your answers only in the provided solution boxes or the scratch paper.
- If you solve a task on the scratch paper, clearly reference it in the main solution box.
- All sheets (including scratch paper) have to be returned at the end.
- **Only use a black or a blue pen (no pencils, red or greens pens!)**
- **For problems that say "Justify your answer" you only get points if you provide a valid explanation.**
- **For problems that say "Derive" you only get points if you provide a valid mathematical derivation.**
- **For problems that say "Prove" you only get points if you provide a valid mathematical proof.**
- If a problem does not say "Justify your answer", "Derive" or "Prove", it is sufficient to only provide the correct answer.

Left room from __________ to __________ / Early submission at __________

Problem 1 Normalizing flows (5 credits)

In this task will focus on the reverse parametrization for normalizing flows on $\mathbb{R}^d$.

0
1
2

a) Let $\boldsymbol{x} \in \mathbb{R}^2$ and the transformation is defined as follows:

$$\boldsymbol{A} = \boldsymbol{a}^T \boldsymbol{a}$$
$$\boldsymbol{z} = \sigma(\boldsymbol{A}\boldsymbol{x}),$$

where $\boldsymbol{a} \in \mathbb{R}^{1\times 2}_{>0}$ and σ is the element-wise sigmoid activation.

Please state whether this transformation leads to a valid normalizing flow. Justify your answer accordingly.

0
1
2

b) Now, let $\boldsymbol{x} \in \mathbb{R}^3$ and the transformation is defined as follows:

$$z_1 = (x_3 + x_2)^3$$
$$z_2 = x_1^4 x_2 + x_3$$
$$z_3 = e^{x_3}.$$

Please state whether this transformation leads to a valid normalizing flow. Justify your answer accordingly.

c) Lastly, let's assume you are given a transformation $f : \mathbb{R} \rightarrow \mathbb{R}$, where we know that the Jacobian determinant of its inverse is equal to 1. How does this affect the normalizing flow?

0

1

Please use the change of variable formula and a possible parametrization of f^{-1} to explain.

Problem 2 **Variational Inference & Variational Autoencoder (9 credits)**

We want to draw samples from a log-normal distribution $\log \mathcal{N}(\mu, \sigma^2)$, where $\mu, \sigma \in \mathbb{R}$, with reparametrization. The probability density function of the log-normal distribution is defined as:

$$q_{\mu,\sigma^2}(z) = \begin{cases} \frac{1}{z\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln z - \mu)^2}{2\sigma^2}\right) & \text{if } z > 0 \\ 0 & \text{otherwise} \end{cases}$$

Its cumulative density function is given as:

$$Q_{\mu,\sigma^2}(a) = \Pr(z \leq a) = \int_{-\infty}^{a} q_{\mu,\sigma}(z)dz = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{\ln a - \mu}{\sigma\sqrt{2}}\right)\right]$$

Recall that the error function $\operatorname{erf}(z)$ is an invertible function that is defined as $\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z \exp(-t^2)dt$.

0 1 2 3

a) Suppose you have access to an algorithm that produces samples ϵ from a standard normal distribution $\mathcal{N}(0, 1)$. Find a deterministic transformation $T : \mathbb{R} \to \mathbb{R}_{>0}$ that transforms a sample $\epsilon \sim \mathcal{N}(0, 1)$ into a sample from the log-normal distribution $\log \mathcal{N}(0, 1)$.
Hint: The cumulative density function of a normal distribution $\mathcal{N}(\mu, \sigma^2)$ is given as:

$$F_{\mu,\sigma^2}(a) = \Pr(z \leq a) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{a - \mu}{\sigma\sqrt{2}}\right)\right]$$

0 1 2 3

b) Now suppose you have access to an algorithm that produces samples z from a log-normal distribution $\sim \log \mathcal{N}(0, 1)$. Find a deterministic transformation $M_{\mu,\sigma^2} : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ that transforms a sample $z \sim \log \mathcal{N}(0, 1)$ into a sample from the log-normal distribution $\log \mathcal{N}(\mu, \sigma^2)$.

c) Now suppose you have access to an algorithm that produces samples ϵ from a standard normal distribution $\mathcal{N}(0, 1)$. Find a deterministic transformation $C_{\mu,\sigma^2} : \mathbb{R} \to \mathbb{R}_{>0}$ that transforms a sample $\epsilon \sim \mathcal{N}(0, 1)$ into a sample from the log-normal distribution $\log\mathcal{N}(\mu, \sigma^2)$. 0 1

Hint: Use the results from the previous subproblems.

0
1
2

d) We want to model the distribution of data samples $p(x)$ using a Variational Autoencoder. Recall that this assumes a latent variable structure $p(x, z) = p(x|z)p(z)$ and we need to model the distribution $p_\theta(x|z)$ and the variational distribution $q_\phi(z)$ respectively. We learn the parameters of our model by optimizing the ELBO:

$$\mathcal{L}(\theta, \phi) = \underset{z \sim q_\phi(z)}{\mathbb{E}} \left[\log p_\theta(x|z)\right] - \mathbb{KL}\left[q_\phi(z) \mid p(z)\right]$$

Here, $\mathbb{KL}$ is the Kullback-Leibler divergence $\mathbb{KL}\left[p(z) \| q(z)\right] = \int p(z) \log \frac{p(z)}{q(z)} dz$. For simplicity, assume that the latent variable z is scalar.

Instead of assuming a standard normal prior $p(z) = \mathcal{N}(0, 1)$ on the latent variable z, we want to employ a log-normal prior $p(z) = \log \mathcal{N}(0, 1)$. Justify why parametrizing $q_\phi(z)$ as a normal distribution $\mathcal{N}(\mu, \sigma^2)$ is not a practical idea. Furthermore, propose an alternative suitable parametrization and briefly outline how we can backpropagate through sampling from $q_\phi(z)$.

Hint: You may refer to the procedure of c), even if you could not derive it.

Problem 3 Generative Adversarial Networks (8 credits)

For $\pi = \frac{1}{2}$, GANs are trained by optimizing the model parameters θ according to

$$\min_{\theta} \max_{\phi} \frac{1}{2} \underbrace{\mathop{\mathbb{E}}_{p^*(\boldsymbol{x})}[\log D_\phi(\boldsymbol{x})]}_{E_1} + \frac{1}{2} \underbrace{\mathop{\mathbb{E}}_{p(\boldsymbol{z})}[\log(1 - D_\phi(f_\theta(\boldsymbol{z})))]}_{E_2}.$$

a) Based on this training objective, explain in one sentence each

- the meaning of the first expected value E_1,
- the meaning of the second expected value E_2,
- and what is adversarial about this formulation.

0 1 2 3

0
1
2
3
4
5

b) Show that the loss

$$\mathcal{L} = \max_{\phi} \frac{1}{2} \mathop{\mathbb{E}}_{p^*(\boldsymbol{x})}[\log D_\phi(\boldsymbol{x})] + \frac{1}{2} \mathop{\mathbb{E}}_{p(\boldsymbol{z})}[\log(1 - D_\phi(f_\theta(\boldsymbol{z})))]$$

from the GAN objective is equivalent to the Jensen-Shannon divergence (JSD) between the data distribution p^* and the learned, generated distribution p_θ, i.e.

$$\mathcal{L} = \mathrm{JSD}(p^*, p_\theta).$$

The JSD between two probability densities p and q is defined as

$$\mathrm{JSD}(p, q) = \frac{1}{2}\left[\mathbb{KL}\left(p \| \tfrac{1}{2}(p+q)\right) + \mathbb{KL}\left(q \| \tfrac{1}{2}(p+q)\right)\right]$$

where $\mathbb{KL}$ is the Kullback-Leibler (KL) divergence $\mathbb{KL}(p\|q) = \mathbb{E}_p\left[\log \frac{p}{q}\right]$.

Hint: Remember the general form of the optimal discriminator.

Hint: For GANs, it holds for functions h that

$$\mathop{\mathbb{E}}_{p(\boldsymbol{z})}[h(f_\theta(\boldsymbol{z}))] = \mathop{\mathbb{E}}_{p_\theta(\boldsymbol{x})}[h(\boldsymbol{x})].$$

Problem 4 Denoising Diffusion (6 credits)

Consider a denoising diffusion model with N diffusion steps and the usual forward parametrization $q_{\varphi_{\mathbf{x}_0}}$ and reverse process p_θ.

$$\alpha_n = 1 - \beta_n \quad \bar{\alpha}_n = \prod_{i=1}^{n} \alpha_i \quad \tilde{\beta}_n = \frac{1-\bar{\alpha}_{n-1}}{1-\bar{\alpha}_n}\beta_n$$

$$q_{\varphi(\mathbf{x}_0)}(\mathbf{z}_n) = \mathcal{N}\left(\sqrt{\bar{\alpha}_n}\mathbf{x}_0, (1-\bar{\alpha}_n\mathbf{I})\right)$$

$$q_{\varphi(\mathbf{x}_0)}(\mathbf{z}_{n-1} \mid \mathbf{z}_n) = \mathcal{N}\left(\frac{\sqrt{\alpha_n}(1-\bar{\alpha}_{n-1})}{1-\bar{\alpha}_n}\mathbf{z}_n + \frac{\sqrt{\bar{\alpha}_{n-1}}\beta_n}{1-\bar{\alpha}_n}\mathbf{x}_0, \tilde{\beta}_n\mathbf{I}\right)$$

$$\mathbf{x}_0 = \frac{\mathbf{z}_n - \sqrt{1-\bar{\alpha}_n}\epsilon_\theta(\mathbf{z}_n, n)}{\sqrt{\bar{\alpha}_n}}$$

a) Why do we optimize the ELBO instead of the data log-likelihood?

0
1

b) Why does model training fail if $\beta_n > 1$?

0
1

c) Why does model training fail if $\beta_n = 1$?

0
1

0
1
2

d) Which of these beta schedules are *invalid*? Justify your answer.

1. $\beta_n = \sin\left(\frac{n}{N}\right)$
2. $\beta_n = 1 - \frac{1}{n}$
3. $\beta_n = \log_e\left(1 + \frac{n}{N}\right)$
4. $\beta_n = -\cos\left(\frac{\pi n}{N}\right)$

Additional space for solutions–clearly mark the (sub)problem your answers are related to and strike out invalid solutions.