Data Analytics & Machine Learning
Informatics
Technical University of Munich

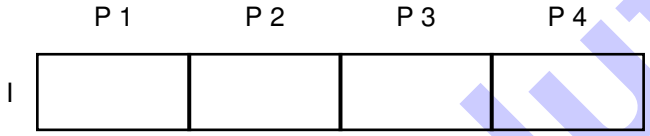
Ecorrection
Place student sticker here

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



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 (after problem 10).
- 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:

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

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.

No, this transformation is not invertible. Trivially $\boldsymbol{A}$ does not have full rank. Therefore, the determinant of $\boldsymbol{A}$ is zero and the transformation is non-invertible.

✓ for no.
✓ for pointing out that it is not invertible.
✓ for providing a reason why it is not invertible.

0 1 2

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

$$\begin{aligned} z_1 &= (x_3 + x_2)^3 \\ z_2 &= x_1^4 x_2 + x_3 \\ z_3 &= e^{x_3}. \end{aligned}$$

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

No, this is not a valid transformation. To disprove the bijectivity of the transformation, we can find a counter example. For any assignment of $\boldsymbol{x}$, we can choose a different assignment that maps to the same $\boldsymbol{z}$ by replacing x_1 with $-x_1$.

✓ for no.
✓ for pointing out that it is not invertible.
✓ for providing a reason (counter example) why it is not invertible.

c) Lastly, let's assume you are given a transformation $f : \mathbb{R} \to \mathbb{R}$, 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.

A flow which Jacobian determinant is equal to 1 is volume preserving. This means that the transformed distribution has the same volume, i.e., the change of variable formula is given by $p_2(x) = p_1(f^{-1}(x)) * 1$. Thus, a normalizing flow with a Jacobian determinant of one is not expressive and can not model any other distribution than $p_1(z)$.
A trivial example is $f^{-1}(x) = x + b$, where $b \in \mathbb{R}$, which includes the identity map and any translation.

✓ for pointing out it is volume preserving (but does not have to name **volume preserving**) by using the change of variable formula.
✓ for giving an example.

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:

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

We want to find T such that $\Pr(T(\epsilon) \leq a) = Q_{0,1}(a)$

$$\begin{aligned} \Pr(T(\epsilon) \leq a) &= \Pr(\epsilon \leq T^{-1}(a)) \\ &= F_{0,1}(T^{-1}(a)) \\ &= \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{T^{-1}(a)}{\sqrt{2}}\right)\right] \\ &\overset{!}{=} Q_{0,1}(a) \\ &= \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{\ln a}{\sqrt{2}}\right)\right] \end{aligned}$$

Since $\operatorname{erf}(z)$ is invertible, we just have to match the arguments of the error function.

$$T^{-1}(a) = ln(a) \Rightarrow T(\epsilon) = \exp(\epsilon)$$

✓ for stating that the cdfs have to match. ✓ for writing out both cdfs and ✓ for matching the arguments of the error functions. ✓ for the correct transformation $T(\epsilon)$.

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 $\log\mathcal{N}(\mu, \sigma^2)$.

We want to find M_{μ,σ^2} such that $\Pr(M_{\mu,\sigma^2}(z) \leq a) = Q_{\mu,\sigma^2}(a)$

$$\begin{aligned} \Pr(M_{\mu,\sigma^2}(z) \leq a) &= \Pr(z \leq \Pr(M^{-1}_{\mu,\sigma^2}(a)) \\ &= Q_{0,1}(M^{-1}_{\mu,\sigma^2}(a)) \\ &= \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{\ln M^{-1}_{\mu,\sigma^2}(a)}{\sqrt{2}}\right)\right] \\ &\overset{!}{=} Q_{\mu,\sigma^2}(a) \\ &= \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{\ln a - \mu}{\sigma\sqrt{2}}\right)\right] \end{aligned}$$

Similar to before, we match the arguments of the error function.

$$\begin{aligned} \ln M^{-1}_{\mu,\sigma^2}(a) &= \frac{\ln a - \mu}{\sigma} \\ \sigma \ln M^{-1}_{\mu,\sigma^2}(a) + \mu &= \ln a \\ M^{-1}_{\mu,\sigma^2}(a)^\sigma \exp(\mu) &= a \\ \Rightarrow M_{\mu,\sigma^2}(z) &= z^\sigma \exp(\mu) \end{aligned}$$

✓ for stating that the cdfs have to match. ✓ for writing out both cdfs and ✓ for matching the arguments of the error functions. ✓ for solving the equation for a. ✓ for the correct transformation $M_{\mu,\sigma^2}(z)$.

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.

We simply compose the transformations T which provides samples from $\log \mathcal{N}(0,1)$ and M_{μ,σ^2} which transforms them into samples from $\log \mathcal{N}(\mu, \sigma^2)$.

$$C_{\mu,\sigma^2}(\epsilon) = \left(M_{\mu,\sigma^2} \circ T\right)(\epsilon) = M_{\mu,\sigma^2}\left((T(\epsilon)\right) = \exp(\epsilon)^\sigma \exp(\mu) = \exp(\sigma\epsilon + \mu)$$

✓ for composing the two transformations ✓ for the correct transformation C_{μ,σ^2}. Alternatively, if this is solved like a) and b), award ✓ for the correct derivation and ✓ for the correct transformation $C(\mu, \sigma^2)$.

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:

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

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 $p(z) = \log \mathcal{N}(0, 1)$. Argue 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.

The KL-divergence $\mathbb{KL}(q_\phi(z) \mid p(z))$ is defined as:

$$\mathbb{KL}\left[q_\phi(z) \mid p(z)\right] = \int_{\mathbb{R}} q_\phi(z) \log\left(\frac{q_\phi(z)}{p(z)}\right) dz \tag{2.1}$$

Since $p(z)$ will be zero for $z \leq 0$ while $q_\phi(z) > 0$, the KL-divergence term diverges to ∞, which prevents gradient-based optimization.
If we instead parametrize $q_\phi(z) = \log \mathcal{N}(\mu, \sigma^2)$, both arguments of the KL-divergence have the same support, ensuring finite values. By employing the reparametrization scheme of c), we can backpropagate through sampling from $q_\phi(z)$.
✓ for recognizing that $p(z)$ and $q_\phi(z)$ have different support. ✓ for concluding that the KL divergence is infinite. ✓ for parametrizing $q_\phi(z)$ as log-normal instead. ✓ for stating that reparametrization is needed and for proposing (or referring to) the procedure of c) (it does not have to be explicited to get marks, just mentioning it is fine).

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

- E_1: Rewards the discriminator for recognizing samples from the data distribution
- E_2: Rewards the discriminator for rejecting samples from the generated distribution and the generator for fooling the discriminator
- Discriminator and generator are adversaries because they optimize the same objective in opposite directions

✓ for explaining the meaning of E_1
✓ for explaining the meaning of E_2
✓ for explaining what is adversarial about a GAN (-✓ if the explanation is non-technical or not related to the training objective, e.g. “the generator tries to fool the discriminator”)

0 1 2 3 4 5

b) Show that the loss

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

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.

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

for some constant $c \in \mathbb{R}$ that does not depend on p^* or θ. The JSD between two probability densities p and q is defined as

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

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

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

The optimal discriminator is given by

$$D_{\phi^*}(\boldsymbol{x}) = \frac{p^*(\boldsymbol{x})}{p^*(\boldsymbol{x}) + p_\theta(\boldsymbol{x})}.$$

$$\mathcal{L} = \max_{\phi} \tfrac{1}{2} \mathop{\mathbb{E}}_{p^*(\boldsymbol{x})} [\log D_\phi(\boldsymbol{x})] + \tfrac{1}{2} \mathop{\mathbb{E}}_{p(\boldsymbol{z})} [\log(1 - D_\phi(f_\theta(\boldsymbol{z})))] \tag{3.1}$$

$$= \max_{\phi} \tfrac{1}{2} \mathop{\mathbb{E}}_{p^*(\boldsymbol{x})} [\log D_\phi(\boldsymbol{x})] + \tfrac{1}{2} \mathop{\mathbb{E}}_{p_\theta(\boldsymbol{x})} [\log(1 - D_\phi(\boldsymbol{x}))] \tag{3.2}$$

Now plug in the optimal discriminator.

$$= \tfrac{1}{2} \mathop{\mathbb{E}}_{p^*(\boldsymbol{x})} [\log \frac{p^*(\boldsymbol{x})}{p^*(\boldsymbol{x}) + p_\theta(\boldsymbol{x})}] + \tfrac{1}{2} \mathop{\mathbb{E}}_{p_\theta(\boldsymbol{x})} \left[\log \left(1 - \frac{p^*(\boldsymbol{x})}{p^*(\boldsymbol{x}) + p_\theta(\boldsymbol{x})}\right)\right] \tag{3.3}$$

$$= \tfrac{1}{2} \mathop{\mathbb{E}}_{p^*(\boldsymbol{x})} [\log \frac{p^*(\boldsymbol{x})}{p^*(\boldsymbol{x}) + p_\theta(\boldsymbol{x})}] + \tfrac{1}{2} \mathop{\mathbb{E}}_{p_\theta(\boldsymbol{x})} \left[\log \left(\frac{p_\theta(\boldsymbol{x})}{p^*(\boldsymbol{x}) + p_\theta(\boldsymbol{x})}\right)\right] \tag{3.4}$$

$$= \tfrac{1}{2} \mathop{\mathbb{E}}_{p^*(\boldsymbol{x})} [\log \frac{p^*(\boldsymbol{x})}{\frac{p^*(\boldsymbol{x}) + p_\theta(\boldsymbol{x})}{2}}] + \tfrac{1}{2} \mathop{\mathbb{E}}_{p_\theta(\boldsymbol{x})} \left[\log \left(\frac{p_\theta(\boldsymbol{x})}{\frac{p^*(\boldsymbol{x}) + p_\theta(\boldsymbol{x})}{2}}\right)\right] - \log(2) \tag{3.5}$$

$$= \tfrac{1}{2}\mathbb{KL}\left(p^* \| \frac{p^*(\boldsymbol{x}) + p_\theta(\boldsymbol{x})}{2}\right) + \tfrac{1}{2}\mathbb{KL}\left(p_\theta \| \frac{p^*(\boldsymbol{x}) + p_\theta(\boldsymbol{x})}{2}\right) - \log(2) \tag{3.6}$$

$$= \mathrm{JSD}(p^*, p_\theta) - \log(2) \tag{3.7}$$

✓ for replacing $\mathbb{E}_{p(\boldsymbol{z})}$ with $\mathbb{E}_{p(\boldsymbol{x})}$ according to the second hint
✓ for plugging in the optimal discriminator
✓✓✓ for transforming (3.3) into (3.7)

Problem 4 Denoising Diffusion (6 credits)

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

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

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

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

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

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

0 1

The data log-likelihood $\log p(\boldsymbol{x})$ requires us to marginalize out the latent variables $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N$ which is intractable.

✓ for saying that the data log-likelihood is intractable
✓ for stating that the reason is marginalizing the $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N$

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

0 1

If $\beta_n > 1$, $\alpha_n = 1 - \beta_n < 0$ which means that $\bar{\alpha}_n$ is going to oscillate in sign with at least one $\alpha_n < 0$. Then we would have to take the square root of a negative number during training when sampling $\boldsymbol{z}_n \sim q_{\phi(\boldsymbol{x}_0)}(\boldsymbol{z}_n)$.

✓ for providing a reason why the model fails
✓ for pointing out where exactly it fails

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

0 1

If $\beta_n = 1$, we get $\alpha_n = 1 - \beta_n = 0$ and therefore $\bar{\alpha}_{n'} = 0$ for $n' \geq n$. In training this would mean that all information from $\boldsymbol{x}_0$ would be lost from the n-th step on. During sampling, we would also have to divide by $\sqrt{0} = 0$ when estimating $\boldsymbol{x}_0$ from $\boldsymbol{z}_n$.

✓ for providing a reason why the model fails
✓ for pointing out where exactly it fails

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. $\beta_n = -\cos\left(\frac{\pi n}{N}\right)$

Schedule 2 is invalid because $\beta_1 = 1 - \frac{1}{1} = 0$. Schedule 4 is invalid because $\beta_n \leq 0$.

✓ for identifying schedule 2 as invalid. ✓ for identifying the reason.
✓ for identifying schedule 4 as invalid. ✓ for identifying the reason.
-✓ for identifying schedule 1 falsely as invalid.
-✓ for identifying schedule 3 falsely as invalid.

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