ATML (deep generative modelling)

Sample exam

(This sample exam totals 45 marks; this track's part of the real exam will be 25 marks.)

(1) *[12 marks]* **Comparing generative model classes.** This question considers models that generate samples by transforming latent noise using a (learnt) function $f : \mathbb{R}^{d_{\text{latent}}} \to \mathbb{R}^{d_{\text{data}}}$.

Let $p_{\text{latent}} = \mathcal{N}(0, I_{d_{\text{latent}}})$ be the latent distribution and p be the induced approximation to the data distribution, *i.e.*, if $z \sim p_{\text{latent}}$ and $x = f(z)$, then $x \sim p$.

(a) *[4 marks]* **Density estimation.** We would like to estimate the density $p(x)$ for $x \in \mathbb{R}^{d_{\text{data}}}$. For each of the following model classes, briefly explain how to estimate $p(x)$ or why this cannot be done in general:

(i) f is the generator of a typical generative adversarial network (GAN);

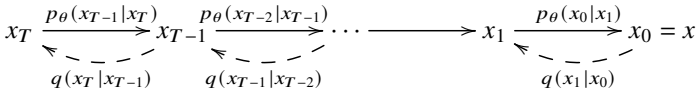
(ii) f is a normalising flow.

(b) *[8 marks]* **GAN with normalising flow as generator.** Suppose that f is a normalising flow (*i.e.*, a smooth function with smooth inverse whose gradient can be tractably computed) and that f is trained using a GAN objective, *i.e.*, jointly with a discriminator $D : \mathbb{R}^{d_{\text{data}}} \to [0, 1]$ using an adversarial loss.

(i) What would be the advantages of such a model, supposing that training were successful, compared to a normalising flow trained using maximum likelihood and to a typical GAN?

(ii) What training difficulties would you expect when training such a model, compared to a normalising flow trained using maximum likelihood and to a typical GAN?

(2) *[11 marks]* **Diffusion models.** This question considers a (discrete-time) diffusion model:



with the variance-exploding noising process $q(x_t \mid x_{t-1}) = \mathcal{N}(x_t; x_{t-1}, \sigma^2 I)$.

(a) *[2 marks]* Explain how to efficiently sample from $q(x_t \mid x_0)$.

(b) *[3 marks]* After training a model on a dataset of samples of x to approximate the reverse of the noising process, we obtain a model $p_\theta(x_{t-1} \mid x_t)$. Explain how to use this model to approximately sample from $p_\theta(x_0)$, including the choice of an initial distribution for x_T.

(c) *[6 marks]* How does the quality of distribution approximation provided by the procedure in (b) depend on:

(i) The value of σ, assuming T is varied to hold the total variance added after T steps constant?

(ii) The number of steps T, assuming σ is held constant?

(3) *[12 marks]* **Latent generative models.** A common way of training generative models of very large images x is to first train a variational autoencoder with a low-dimensional latent space (with encoder $q(z \mid x)$ and decoder $p(x \mid z)$), then to train another model (often a neural ODE) to generate samples z from $q(z \mid x)$.

The resulting model can be used to generate images by first sampling z from the second-stage model, then sampling x from $p(x \mid z)$.

(a) *[4 marks]* Why may such a two-stage procedure be preferred to training a single neural ODE on images x directly?

(b) *[4 marks]* Why may such a two-stage procedure be preferred to simply sampling z from the prior of the VAE?

(c) *[4 marks]* Suppose that the VAE is trained on unlabelled images, but that we have access to class labels for each image in the dataset. How would you modify *only* the second-stage training procedure to obtain a generative model of images conditioned on class labels?

(4) *[10 marks]* **Deployment risks and failure modes.** Suppose that a generative model of videos, conditioned on text descriptions, is trained on videos of house pets (*e.g.*, cats and dogs) collected from a video sharing site. The model is deployed as a service that creates videos based on user-provided text prompts, with the intended use of entertaining users with videos of animals doing amusing things.

(a) *[5 marks]* To prevent the model from leaking private information, the model developers filter the training data to remove any videos that contain humans using an off-the-shelf face detection model. Discuss the potential risks and failure modes of this approach.

(b) *[5 marks]* To reduce the data requirements for training, the developers initialise the model using a pretrained text-to-video model trained on a large, diverse, and unfiltered dataset of videos and text captions. How could this choice impact the risks and failure modes in part (a)?

ATML (deep generative modelling)
Sample exam solutions

(1) (a) (i) The density cannot be estimated in general. Firstly, the target distribution may not have a density in $\mathbb{R}^{d_{\text{data}}}$ (for example, if $d_{\text{latent}} < d_{\text{data}}$, so the distribution is supported on a lower-dimensional manifold). Secondly, even if the density exists, the mapping f may not be invertible, *i.e.*, there may be multiple z such that $f(z) = x$, and these z may not be tractably found.

(ii) The density can be estimated using the change of variables formula:

$$p(x) = p_{\text{latent}}(f^{-1}(x)) \left|\det\left(\frac{\partial f^{-1}}{\partial x}\right)\right| = p_{\text{latent}}(z) \left|\det\left(\frac{\partial f}{\partial z}\right)\right|^{-1},$$

where $x = f(z)$. Since f is a normalising flow, z is unique and can be tractably computed, as can the determinant of the Jacobian.

(b) (i) Advantages compared to a normalising flow trained using maximum likelihood include higher sample quality, because the GAN loss encourages perceptually better samples due to the discriminator's inductive bias towards learning natural measures of realism and the implicit minimisation of divergences (JSD) that optimise in regions of high density.
Advantages compared to a typical GAN include tractable density estimation (cf. (a)(ii)).
(Other answers to both are possible if well justified.)

(ii) Training difficulties compared to a normalising flow trained using maximum likelihood include instability due to adversarial training, mode collapse, difficulty of reducing mass of regions where the discriminator has high confidence due to normalising flows' inductive bias towards mode connectivity (whereas a maximum-likelihood-trained normalising flow is not penalised for placing mass in such regions).
Training difficulties compared to a typical GAN include increased cost due to the input dimension being higher for a NF generator ($d_{\text{latent}} = d_{\text{data}}$), difficulty of reducing mass of regions where the discriminator has high confidence due to normalising flows' inductive bias towards mode connectivity (whereas a GAN can more flexibly reduce the mass of such regions).
(Other answers to both are possible if well justified.)

(2) (a) We have $q(x_t \mid x_0) = \mathcal{N}(x_t; x_0, t\sigma^2 I)$, *i.e.*, x_t can be sampled by adding Gaussian noise with variance $t\sigma^2$ to x_0.

(b) We first sample $x_T \sim \mathcal{N}(0, T\sigma^2 I)$, then for $t = T, T-1, \ldots, 1$ we sample $x_{t-1} \sim p_\theta(x_{t-1} \mid x_t)$. The final sample x_0 is approximately from $p_\theta(x_0)$. (Other choices for the initial distribution of x_T are possible, such as taking its mean to be the empirical data mean, increasing it by the empirical data variance, or learning it as a model parameter.)

(c) (i) Smaller σ generally leads to better approximations, because each step of the reverse process is better approximated by a Gaussian when the noise added in the noising process is small.

(ii) Larger T generally leads to better approximations, because the distribution of the final noised sample x_T is closer to a Gaussian when more noise is added (*i.e.*, the final signal-to-noise ratio is lower).

(3) (a) The neural ODE would need to model dynamics in the high-dimensional image space and would be more costly to train and sample from, while the latent neural ODE models dynamics in the lower-dimensional latent space. The distribution of latent representations learnt by the VAE can also be more amenable to generalisation than the distribution of images. (Other answers are possible if well justified.)

(b) The VAE prior may be a poor fit to the aggregate posterior, *i.e.*, the distribution of $z \sim q(z \mid x)$ when x is drawn from the data distribution. Thus, samples from the VAE prior may be decoded to low-quality images, whereas samples from the second-stage model are trained to match the aggregate posterior and so decode to higher-quality images. (Other answers are possible if well justified.)

(c) We train a *conditional* neural ODE $p_{\text{latent}}(z \mid y)$ on samples $z \sim q(z \mid x)$ for x drawn from the dataset, where y is the class label associated with x. At generation time, we first sample $z \sim p_{\text{latent}}(z \mid y)$ for the desired class label y, then decode z to $x \sim p(x \mid z)$.

(4) (a) Possible risks and failure modes include:

- Biased performance of the face detection model, leading to failure to remove some videos containing humans (or some groups of humans in different proportions), or other human body parts appearing in videos, leading to violations of humans' privacy.
- Adversarial attacks on the model, leading to apparance of humans in generated videos, lack of appearance of animals, generation of other inappropriate content.

- Loss of diversity in the training data due to filtering, leading to memorisation of training data.
- High computation cost of training and inference; associated social and environmental cost.

(Other answers are possible if well justified.)

(b) Possible effects include:

- Decreased sensitivity to reduced diversity in the training data, if the pretraining dataset is more diverse than the fine-tuning dataset.
- Inherited biases and sensitivity to adversarial attacks from the pretrained model, leading to increased undesired bias and vulnerability to exploits.
- Decreased cost of training.

(Other answers are possible if well justified.)