Machine Learning and Physics
Winter Semester 2025/26

Exercise Sheet 10

Due: 19.01.26 14h00

Christof Gehrig, Gerrit Gerhartz

General Regulations.

- Please hand in your solutions in groups of two (preferably from the same tutorial group).
- Your solutions to theoretical exercises can be either handwritten notes (scanned), or typeset using LaTeX. For scanned handwritten notes, please ensure they are legible and not blurry.
- For the practical exercises, always provide the (commented) code as well as the output, and don't forget to explain/interpret the latter.
- Please hand in a **single PDF** that includes both the exported notebook and your solutions to the theoretical exercises. Submit the PDF to the Übungsgruppenverwaltung once per group, making sure to include the names of both group members in the submission.
- You can find all the data in the GitHub Repository.

1 Restricted Boltzmann Machine

Consider an RBM with a layer of n visible units $\mathbf{v} \in \{0,1\}^n$ and m hidden units $\mathbf{h} \in \{0,1\}^m$. The energy function is defined as:

$$E(\mathbf{v},\mathbf{h}) = -\sum_{i=1}^{n} a_i v_i - \sum_{j=1}^{m} b_j h_j - \sum_{i,j} v_i W_{ij} h_j$$

Where a_i, b_j are biases and W_{ij} are weights. The joint probability follows the Boltzmann distribution:

$$P(\mathbf{v},\mathbf{h}) = \frac{1}{Z} e^{-E(\mathbf{v},\mathbf{h})}$$

(a) What physical system does this model represent if we set $W_{ij} = 0$? What role does W_{ij} play physically? (2 pts)

(b) Prove that the hidden units are conditionally independent given the visible units, i.e.:

$$P(\mathbf{h} \mid \mathbf{v}) = \prod_j P(h_j \mid \mathbf{v})$$

(2 pts)

(c) To train the RBM, the goal is to maximize the log-likelihood of the visible data $\mathcal{L} = \ln P(\mathbf{v})$. Show that the gradient with respect to a weight W_{ij} is

$$\frac{\partial \mathcal{L}}{\partial W_{ij}} = \langle v_i h_j \rangle_{\text{data}} - \langle v_i h_j \rangle_{\text{model}}.$$

Here, $\langle \cdot \rangle_{\text{data}}$ is the expectation when $\mathbf{v}$ is clamped to the training data and $\langle \cdot \rangle_{\text{model}}$ is the expectation over the free-running "thermal" distribution of the model. (3 pts)

2 Instantaneous Change of Variables

In this exercise, we try to give an informal derivation of the Instantaneous Change of Variables formula

$$\frac{\mathrm{d}\ln p(\mathbf{z}(t))}{\mathrm{d}t} = -\mathrm{Tr}\left(\frac{\partial f}{\partial \mathbf{z}(t)}\right)$$

for one-dimensional distributions.

Consider a distribution $q(z)$ at time t that is transformed to a new distribution $p(z)$ at time $t+\delta t$ as a result of a transformation from z to x. Also consider nearby values z and $z+\Delta z$ along with corresponding values x and $x+\Delta x$ as shown in the following figure.

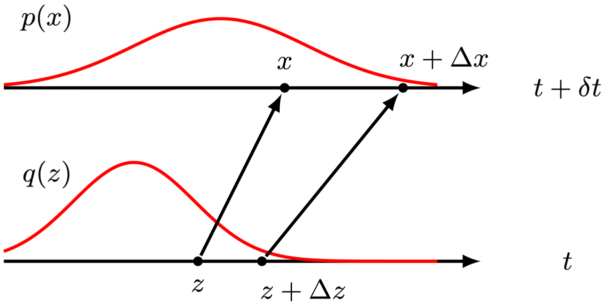

Figure 1: Schematic version of the transformation from z to x.[1]

(a) Write down an equation that expresses that the probability mass in the interval Δz is the same as that in the interval Δx.

(b) Write down an equation that shows how the probability density changes in going from t to $t+\delta t$, expressed in terms of the derivative $\frac{\mathrm{d}q(t)}{\mathrm{d}t}$.

(c) Third, write down an equation for Δx in terms of Δz by introducing the function $f(z) = \frac{\mathrm{d}z}{\mathrm{d}t}$.

(d) Finally, by combining these three equations and taking the limit $\delta t \to 0$, show that

$$\frac{\mathrm{d}}{\mathrm{d}t}\ln q(z) = -f'(z),$$

which is the on-dimensional version of the equation above.

(4 pts)

3 Flow Matching

In this exercise, you will study flow matching (https://arxiv.org/abs/2210.02747), a very popular generative model related to diffusion models.

Given samples from a base distribution p and target distribution q (on the same input domain), the aim is to learn a velocity field v_t that pushes samples from p forward to the distribution of q (by integrating along the velocity field). The flow field v_t is optimized for every time step $t \in [0,1]$ by minimizing the following loss function (cf. Eq. (23) in the paper):

$$\mathcal{L}_{\mathrm{CFM}}(\theta) = \mathbb{E}_{t\sim U[0,1],x_1\sim q,x_0\sim p}\left\|v_t\big(\psi_t(x_0)\big) - (x_1 - x_0)\right\|^2, \quad \psi_t(x_0) = (1-t)x_0 + tx_1,$$

meaning that at points, which interpolate between two samples $x_0 \sim p$ and $x_1 \sim q$, the flow field is trained to point from x_0 to x_1.

[1] Schematic from Bishop, C. M., & Bishop, H. (2024). Deep Learning: Foundations and Concepts. Springer Nature.

(a) Implement this training procedure in the Jupyter notebook and use it to reproduce the setup of Fig. 4 in the paper. In Fig. 4, the authors convert samples from a 2D normal distribution into a 2D checkerboard distribution. (4 pts)

(b) Assuming successful training, based on the loss function above, argue theoretically how the optimal velocity field looks at $t = 0$. Confirm your hypothesis experimentally by evaluating your model on a grid for $[-0.5, 0.5]^2$ for $t = 0$. (3 pts)

(c) Given that flow matching aims to produce integration trajectories which are as straight as possible, why can such a velocity field at $t = 0$ be suboptimal, e.g. in our example? How can minibatch optimal transport help with this problem? (2 pts)