Machine Learning and Physics
Winter Semester 2024/2025

Exercise Sheet #10

Due: 20.01.24 14h00

Peter Lippmann

General Regulations.

- Please hand in your solutions in groups of two (preferably from the same tutorial group). Submissions by a single person alone will not be corrected.
- Your solutions to theoretical exercises can be either handwritten notes (scanned), or typeset using LaTeX. For scanned handwritten notes please make sure that they are legible and not too blurry.
- For the practical exercises, the data and a skeleton for your jupyter notebook are available at `https://github.com/sciai-lab/mlph_w24`. Always provide the (commented) code as well as the output, and don't forget to explain/interpret the latter. Please hand in your notebook (`.ipynb`), as well as an exported pdf-version of it.
- Submit all your files in the Übungsgruppenverwaltung, only once for your group of two. Specify all names of your group in the submission.

On **Wednesday, January 15, 2025**, your student council MathPhysInfo and the Department of Physics and Astronomy will host the **A**rbeitsgruppen**I**nspirations**M**esse (research group inspiration fair) – short **AIM**. Starting at **3:00 p.m.**, in the KIP and in the PI (INF 226 & 227), you can meet professors, PhD students, master students and other members from different research groups and talk about thesis projects, a student job or the current research in the groups in general. Your student council will provide you with snacks and drinks.

1 Optimal transport

In Optimal Transport (OT) the objective is to move mass from one distribution to the other in an optimal way, i.e. to find a transport plan that minimizes the total transportation cost.

(a) Let us consider two discrete distributions. Suppose we have a sets of points at locations $x_i \in R^d$ with supply of mass $a_i \in R^+$ ("sources") and a set of points located at $y_j \in R^d$ with demand $b_j \in R^+$ ("sinks"), s.t. $\sum_i a_i = \sum_j b_j$. Assuming that the transportation costs are linear in the transported mass and transportation distance, write down the optimal transport objective and the constraints. Argue why the mass should be transported along straight lines between sources and sinks in order to minimize the cost. (2 pts)

(b) A linear programming (LP) problem aims to minimize a linear objective function subject to linear constraints. In matrix and vector form, the problem can be written as:

$$\text{minimize} \quad \mathbf{c}^T\mathbf{x}$$

subject to:

$$\mathbf{A}\mathbf{x} \leq \mathbf{b}, \quad \mathbf{x} \geq \mathbf{0}$$

where $\mathbf{x} \in \mathbb{R}^n$ is the vector of decision variables, $\mathbf{c} \in \mathbb{R}^n$ is the vector of coefficients in the objective function, $\mathbf{A} \in \mathbb{R}^{m\times n}$ is a matrix and $\mathbf{b} \in \mathbb{R}^m$ a vector, which together define the constraints.

Rewrite the optimization problem from a) in the form of a linear program. *Hint: Show how you can express equality constraints using the inequality* $\mathbf{Ax} \leq \mathbf{b}$. *Then use equality constraints for simplicity.* (2 pts)

(c) Use a standard solver for linear programming to solve the 5d optimal transport problem provided in the jupyter notebook. What is the final transportation cost? (2 pts)

(d) Consider a transportation plan in which two transportation routes intersect. Prove that such transportation plans are never optimal, assuming that the transportation cost is linear in the transportation distance. (2 pts)

2 Flow matching for generative modeling

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 push 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):

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

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.

(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)

3 Adversarial attacks and AI safety

In the lecture, you have seen how probes are used to inspect the internal activations of LLMs to detect malicious behaviour of the LLMS, e.g. whether they are lying (recall ex. 2 from sheet 5). However, you also discussed that these probes can be tricked in the following way: the input to the model is modified in a systematic attack such that the probe will fail to detect lies (labelling them incorrectly as truths).

(a) Explain the connection of tricking a probe to adversarial attacks. (2 pts)

(b) Following up on the LLM lie detector exercise on sheet 5, train again an LLM lie detector on the cities dataset. Then, take a single sample from the cities dataset (with label 0, i.e. a lie) and optimize a perturbation vector via gradient descent, s.t. the lie detector incorrectly classifies the sample (plus the perturbations) as a true statement. Which loss function do you use? Add this perturbation vector to the rest of the samples with label 0. How do the lie detector predictions change? (3 pts)

(c) Add an additional side loss which should ensure that the perturbation is small. Why can the lie detector no longer be tricked (even for a single sample)? Try to explain your finding using the linear representation hypothesis. (2 pts)