Introduction to Data Science **Assignment 1**

Instructors: **Dr. Bahrak, Dr. Yaghoobzadeh** TAs: جواد کاویان، محمدرضا علوی، فاطمه محمدی

Deadline: 15th Farvardin

Introduction

In the first task of this assignment, you are going to gain hands-on experience of a well known sampling algorithm called Langevin Dynamics.
In the second task, you will explore Tableau's capabilities, create interactive dashboards with filters, parameters, and KPIs to build a structured and insightful data story.

Task 1: Sampling F1

As mentioned in the introduction, in this part of the assignment, we will focus on learning Langevin dynamic sampling. However, before proceeding, it is essential to familiarize ourselves with a concept known as the score function.

Score Function

For the probability distribution with probability density function $P(x)$, score function is:

$$\nabla_x \log p(x)$$

A real-world use case arises when we have a distribution for which the normalization constant ($Z(\theta)$) is unknown. For instance, consider a distribution given by:

$$p(x) = \frac{e^{f_\theta(x)}}{Z(\theta)}$$

given a set of samples to estimate the density, $p(x)$ should be marginalized over x so that they sum up to one and $Z(\theta)$ is calculated. The point is, in many scenarios, x is a high-dimensional data which makes this estimation impossible. Thus we need to be able to get samples from a distribution whose normalization constant is unknown. In the above density function, score function is:

$$\nabla_x \log \frac{e^{f_\theta(x)}}{Z(\theta)} = \nabla_x [f_\theta(x) - \log Z(\theta)] = \nabla_x f_\theta(x)$$

As you can see, we omitted $Z(\theta)$. Langevin Dynamics is the algorithm which takes samples from a distribution, when provided with its score function.

Langevin Dynamics

To take sample using langevin dynamics algorithm, take the following steps:

1. Initialize a point randomly, given an arbitrary distribution $\pi : X^0 \sim \pi(x)$
2. Repeat until convergence(a fixed number of steps):

a. $X^{t+1} \leftarrow X^{t} + \epsilon S_{\theta}(X^{t}) + \sqrt{2\epsilon}Z^{t}$; where ϵ is a hyperparameter specifying step size, S_{θ} is score function and Z^{t} is a gaussian noise, sampled from standard normal distribution

What to do

First, make a 2D gaussian distribution with $mean = [-5, 5]$ and $cov = 5I$ where I is the identity matrix; then plot it. The output should look like the figure 1:

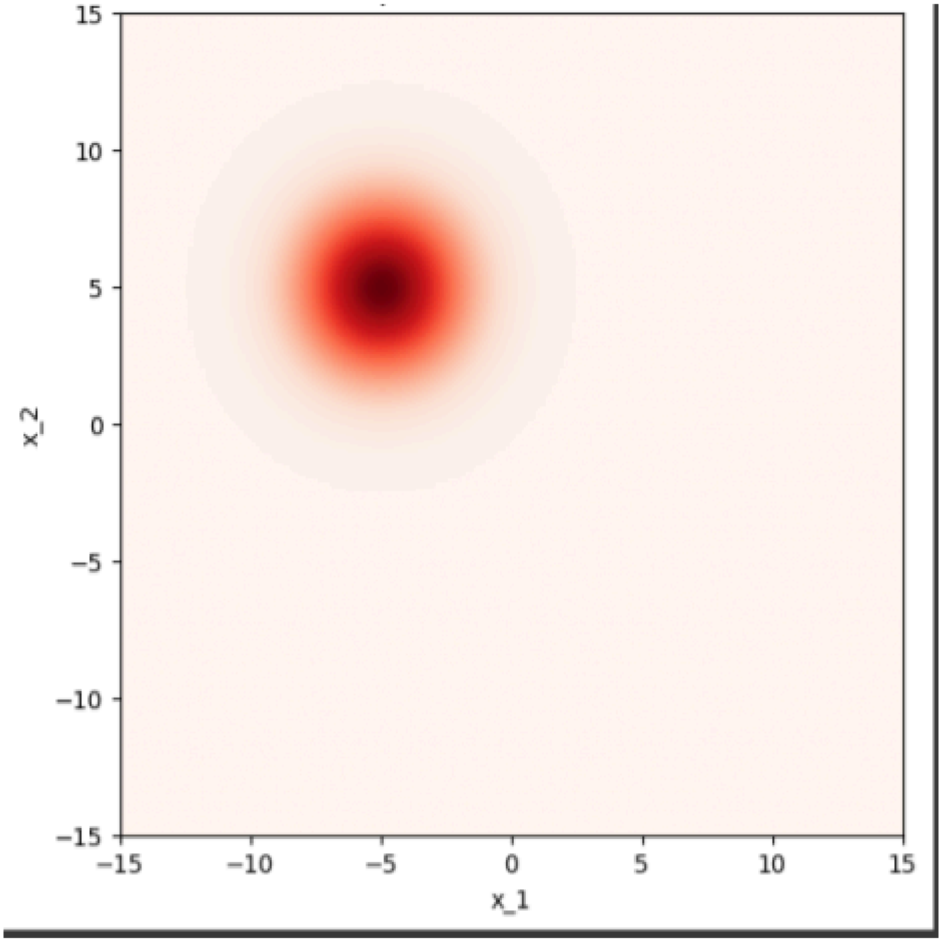

figure1:Heatmap of 2D gaussian

Then you need to implement the score function theoretically with the provided formula and pdf of the distribution. To make sure that you have calculated the score function correctly, you can plot a quiver from points of the grid to their corresponding score. The final result should look something like this:

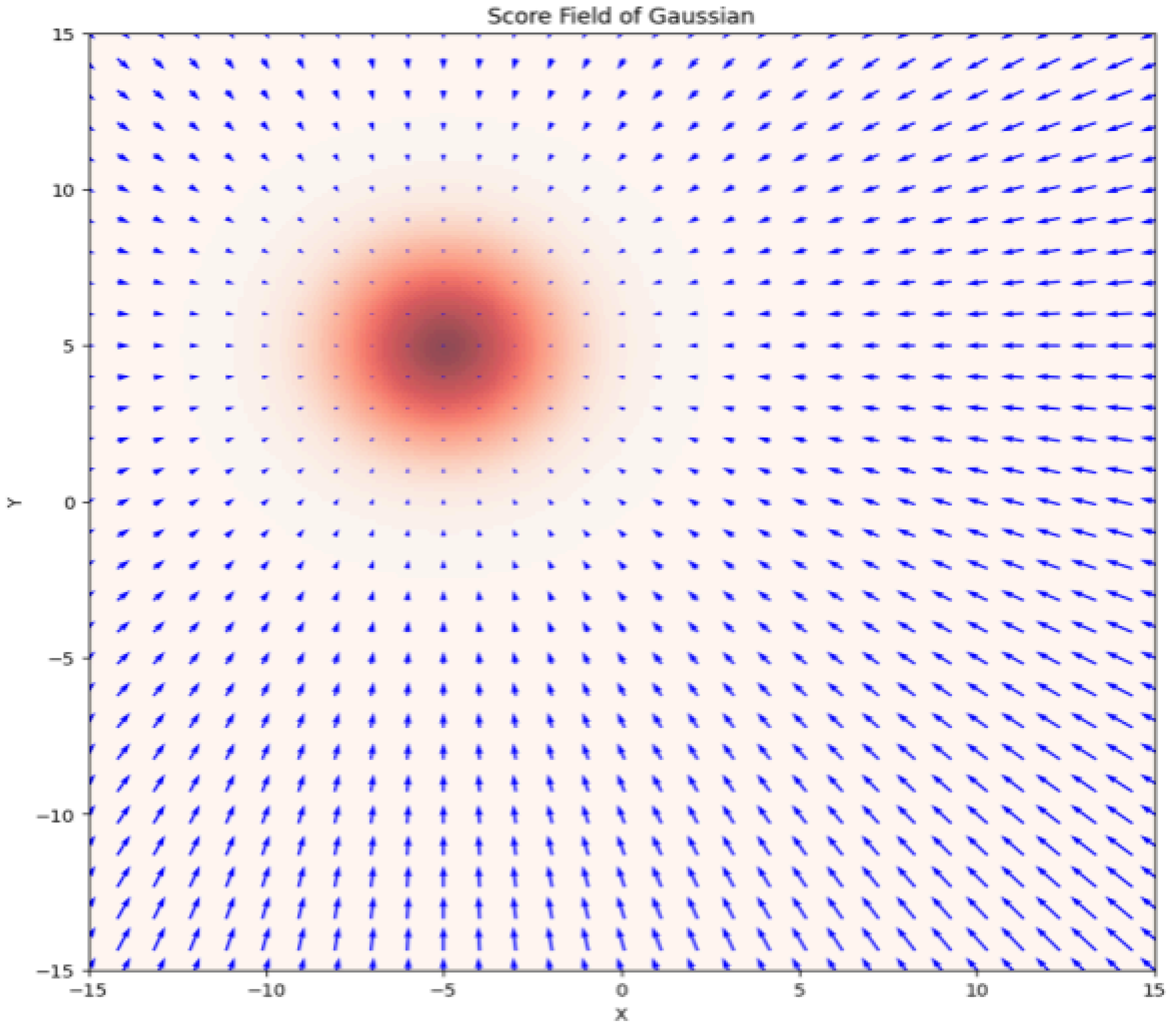

Figure 2: score field of distribution

Now you have all the required utility for langevin sampling. Implement a function which takes an array of initial points and runs langevin dynamics on them and returns the converged samples. To see the progress of the algorithm, you should keep a trajectory of points and plot them:

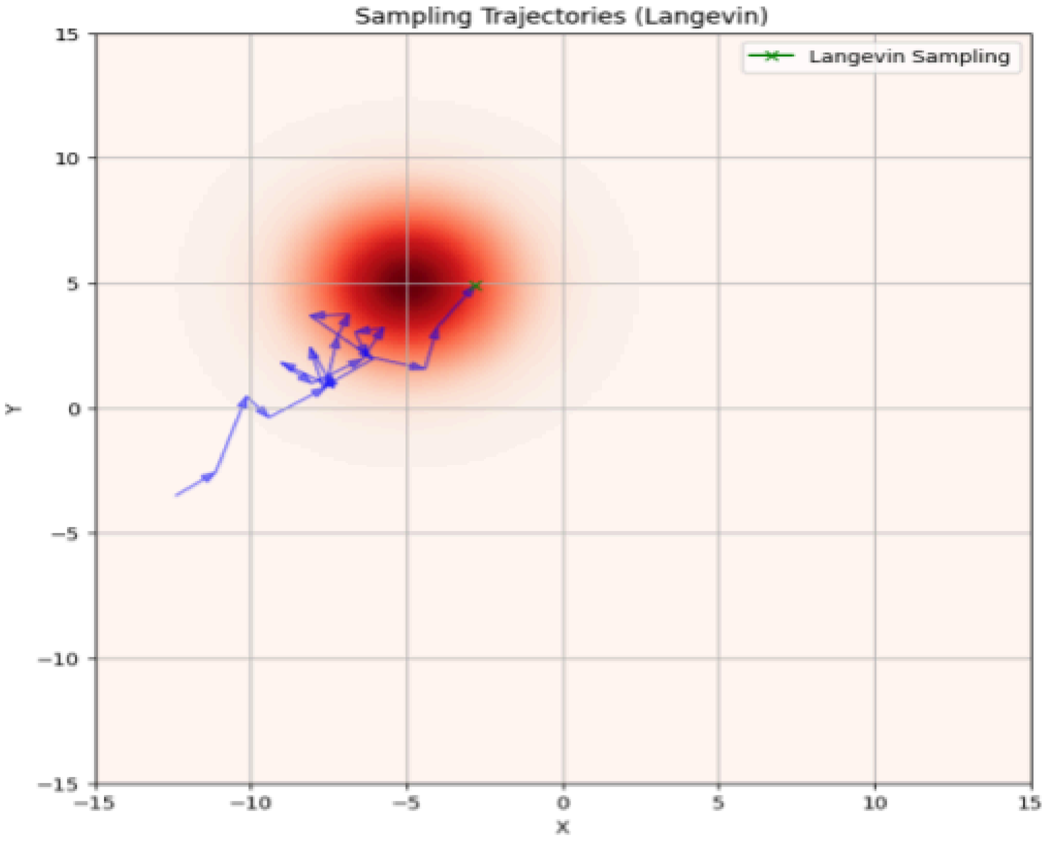

Figure 3:sampling trajectories

Now we want you to take 1000 samples using langevin dynamics and 1000 samples using numpy.random.multivariate_normal method. Compare the visualizations. How else can you compare these two methods?

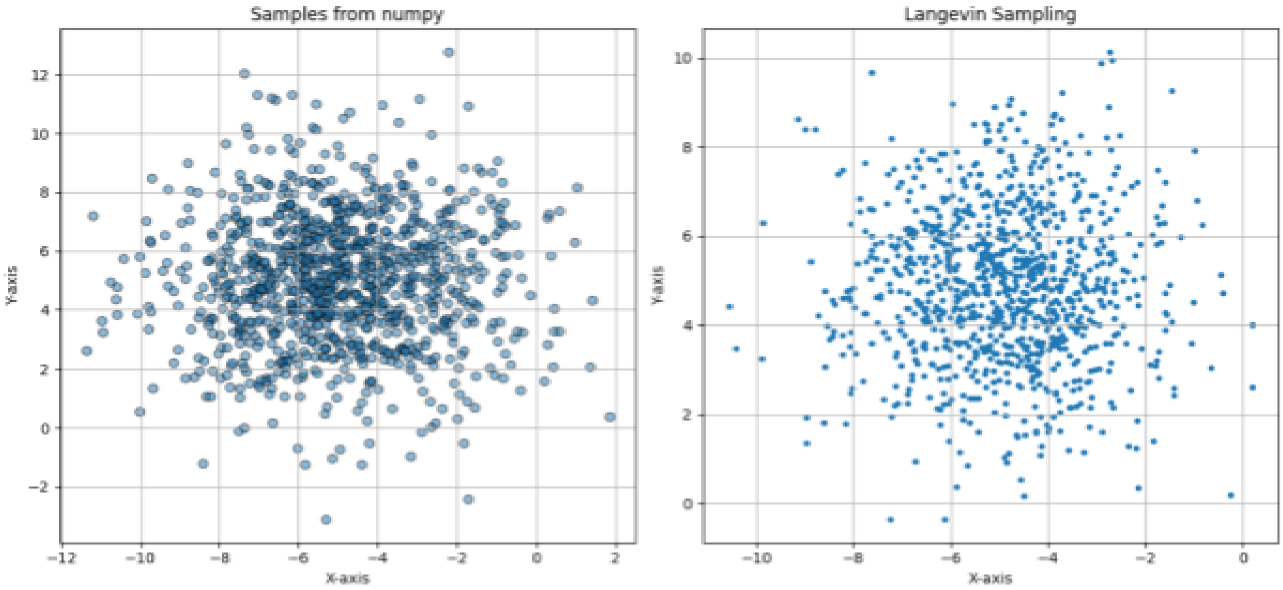

Figure 4: comparing samples

Questions(5% Bonus)

1- In the sampling part, suppose that instead of a gaussian distribution, we have a mixture of gaussians whose density function is:

$$p(x) = \alpha N(x;\, \mu_1, \Sigma_1) + (1 - \alpha)N(x;\, \mu_2, \Sigma_2)$$

for some $\alpha;\ 0 < \alpha < 1$ and normal distributions $N(x;\, \mu_1, \Sigma_1)$ and $N(x;\, \mu_2, \Sigma_2)$, will we be able to take proper samples using langevin dynamics algorithm? Justify your answer.

Task 2: Tableau

In this task, you will utilize Tableau to analyze Airbnb data and create a data-driven story through multiple interactive dashboards. The objective is to build a well-structured narrative using several dashboards, each focusing on different aspects of the data.

Dataset Details

1. **Airbnb_Listings.xls** - contains information about Airbnb listings, including:
   - Neighbourhood – The specific area where the listing is located.
   - Price – Cost per night for each listing.
   - Availability 365 – Number of days a listing is available for booking per year.
   - Calculated Host Listings Count – The total number of listings a host has.
   - Host Id – Unique identifier for the host.
   - Host Name – Name of the host.
   - Id – Unique identifier for the listing.
   - Last Review – Date of the most recent review for the listing.
   - Minimum Nights – Minimum number of nights required per booking.
   - Name – Name of the Airbnb listing.
   - Number Of Reviews – Total number of reviews received for the listing.
   - Reviews Per Month – Average number of reviews per month.
   - Room Type – Type of room (e.g., Entire home/apt, Private room, Shared room).
2. **`Neighborhood_Locations.xlsx`** – Provides latitude, longitude, and neighborhood group classifications for mapping.
   - Neighbourhood: The specific area where the Airbnb listing is located.
   - Neighbourhood Group: A broader classification of neighbourhoods.
   - Latitude: The geographical latitude of the listing.
   - Longitude: The geographical longitude of the listing.

Assignment Requirements

1. Data Preparation F2

- Import both datasets into Tableau.
- Join the datasets properly using neighborhood information to enable location-based analysis. After joining, each location must have longitude and latitude.Handle null values if needed.

2. Multiple Dashboards & Storytelling A

Instead of creating just one dashboard, create multiple dashboards that work together to tell a clear and structured story about Airbnb trends.

- At least three dashboards, each focusing on different aspects of the data.
- Include interactive elements such as filters or parameters to enable dynamic exploration.
- Use KPIs effectively to highlight key performance metrics in the analysis.

3. Insights & Storytelling Structure A

Your story should be structured as:

- Introduction: Define the key focus of the analysis.
- Findings across multiple dashboards:
  - Each dashboard should highlight different aspects of the story.
  - Example insights: seasonal trends, Geographic Price Distribution.
- Conclusion: Summarize the main findings and patterns seen in all the dashboards.

If meaningful insights are found, highlight them. If not, summarize the main trends in a structured way. You do not need to prove the only one relationship, in addition to the lack of an event that is logical.

4. Design Best Practices

- Follow **Gestalt Principles** to keep the dashboards structured and easy to read.
- Use **preattentive attributes** like color, size, and position to highlight key insights.

Presentation & Submission

- Upload your dashboards to Tableau Public.
- Upload a PDF document that includes:
  - The Tableau Public link to your dashboards.
  - A short written summary explaining the story behind your dashboards.

    - Screenshots of all worksheets and dashboards.
    - A structured narrative highlighting key insights and trends.
    - A list of parameters, filters and KPIs used in the dashboards.
- On presentation day, all group members must be prepared to answer questions about their dashboards and how they contribute to the overall story.

Evaluation Criteria

- Data Integration – Properly joining and preparing datasets.
- Dashboard Quality – Well-structured, interactive, and connected dashboards.
- Visualization Variety – At least three dashboards with multiple types of visualizations.
- Storytelling – A single, structured narrative that connects all dashboards.
- Use of Interactivity – Meaningful filters, parameters, KPIs and dashboard actions.
- Design Principles – Applying Gestalt principles and preattentive attributes.
- Group Participation – Everyone should be able to explain their work.

Notes

- Upload your work as a zip file in this format on the website: DS_CA1_[Std number].zip. If the project is done in a group, include all of the group members' student numbers in the name.
- If the project is done in a group, only one member must upload the work.
- We will run your code during the project delivery, so make sure your results are reproducible.
- We will also use your public Tableau dashboard, so ensure it remains available and is not deleted.