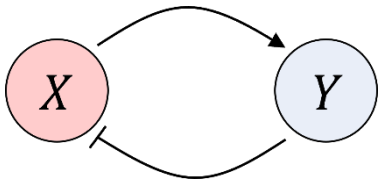
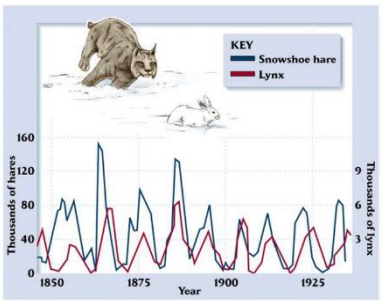
BCS410. Practical Class 4 & Homework 4 [Due: May 05]

Caution: You must submit the code you wrote and used to solve each problem to KLMS. When submitting, name the folder with your student ID and submit each problem's code file as 'ProblemX_sol' (e.g., 'Problem1_sol'). You are free to use Python libraries. However, you are not allowed to ask AI to generate the entire code for you. If detected, you will receive a score of zero. You must provide your answers in a descriptive format for each problem.

Background:

1. Overview of the Lotka-Volterra model (Predator-Prey model): The Lotka-Volterra model captures the dynamics of a biological system involving two interacting species—one acting as the predator and the other as the prey. Independently introduced by Alfred J. Lotka in 1925 and Vito Volterra in 1926, the model has since been widely applied to describe activation-inhibition systems. These arise when the two variables exert opposing effects on each other's growth rates, creating a form of negative feedback. Such activation-inhibition configurations are essential for producing periodic behaviors across domains ranging from molecular biology to ecology.





2. Ordinary differential equations for the Lotka-Volterra model: Let $x(t)$ and $y(t)$ denote the populations of prey and predators at time t, respectively. The model dynamics are governed by the following system of ordinary differential equations (ODEs):

$$\frac{dx}{dt} = \alpha x - \beta xy$$

$$\frac{dy}{dt} = \delta xy - \gamma y$$

where α is the prey's birth rate, β is the predation rate coefficient, γ is the predator's death rate, and δ is the rate at which predators reproduce per prey consumed.

Problem 1 (10 pts): Simulate the model using the following two sets of parameter values:

1) $\alpha = 1,\ \beta = 1,\ \gamma = 1,\ \delta = 1$ 2) $\alpha = 1,\ \beta = 0.05,\ \gamma = 0.1,\ \delta = 1,$

For each case, plot $x(t)$ and $y(t)$ over time and compare them with the data provided in 'LV_data.csv'. Include a brief interpretation in your report describing how the choice of parameter values affects the model's behavior and its ability to capture the dynamics observed in the data.

Problem 2 (10 pts): Estimate the four parameters of the model by fitting it to the provided dataset, 'LV_data.csv', using the scipy.optimize.minimize function (import it with "from scipy.optimize import minimize"). Use the Nelder-Mead simplex algorithm (the default in both MATLAB and Python—you don't need to understand its

internal workings), as specified in the skeleton code, and set the initial guess for all parameters to 1. Additionally, consider: What changes if you use the 1-norm error instead of the 2-norm? You should report the estimated parameter values and graphs of $x(t)$ and $y(t)$ over time with the data.

Problem 3 (20 pts): Modify the code used to solve problem 2 and search minimum error among various initial guesses (Initial guess for $\alpha = (0, 0.2, 0.4, \dots, 1)$, $\beta = (0, 0.03, 0.06, \dots, 0.21)$, $\delta = (0, 0.03, 0.06, \dots, 0.21)$, and $\gamma = (0, 0.2, 0.4, \dots, 1)$. That is, among total $6 \cdot 8 \cdot 8 \cdot 6$ guesses, specify minimum error and corresponding parameter set. Hint: You will need to set $0 < \alpha, \gamma < 2$ and $0 < \beta, \delta < 0.22$ boundary for the range of search so that it does not blow up.

Problem 4 (30 pts): Use the simulated annealing method to find the parameter values that minimize the error, subject to the constraints: $0 < \alpha, \gamma < 2$ and $0 < \beta, \delta < 0.22$. You may need to tune the hyperparameters of the algorithm (e.g., the initial temperature, cooling schedule, step size) to identify the optimal set of parameters that allows the model to accurately capture the trends in the data.

Bonus problem (5 bonus points):

1. Background

1-1. Clustering: Clustering is the task of dividing data into groups (clusters) such that data points in the same group are more similar to each other than to those in other groups. In this problem, our goal is to assign cluster labels to each data point so that within-cluster distances are minimized.

1-2. A single-cell RNA-seq dataset: A single-cell RNA-seq dataset typically captures gene expression profiles at the resolution of individual cells. The data is often structured as a matrix with genes as rows and cells as columns, where each entry represents the number of mRNA transcripts (counts) observed for a particular gene in a particular cell.

	Cell1	Cell2	...	CellN
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
...	.	.	.	.
...	.	.	.	.
...	.	.	.	.
GeneM	25	0	.	0

As described above, single-cell RNA sequencing (scRNA-seq) allows the measurement of gene expression at the level of individual cells. One of the main tasks in analyzing such data is to cluster cells based on their gene expression profiles, enabling the identification of distinct cell types or states. Unlike conventional clustering methods such as k-means, Simulated Annealing (SA) algorithm is a probabilistic metaheuristic that can escape local minima, making it suitable for optimizing clustering objectives in high-dimensional and noisy biological data like scRNA-seq.

2. Problem: Please analyze the scRNA-seq dataset 'single_cell_RNA_seq_data.csv' using the SA algorithm, i.e., cluster the cells into 2, 3, and 4 groups using the SA algorithm, respectively.

Hint: Define the objective function as follows:

$$\text{Total Within Cluster Sum of Squares (WCSS)} = \sum_{i=1}^{k} \sum_{x_j \in C_i} \|x_j - \mu_i\|^2$$

where C_i is cluster i, x_j is a cell, and μ_i is the centroid of cluster i. Then, randomly assign each data point to one of k clusters and begin your SA algorithm. You need to check the robustness of your SA algorithm; that is, it should produce similar results across independent trials with different initial guesses. Furthermore, you should visualize the results, for example, using a scatter plot.