Freie Universität Berlin
Liu-Wei Wang, Max von Kleist

7. Homework
Foundations of Mathematics and Statistics

Deadline: January 08, 10:00 (**before** the lecture)

The homework should be worked out in groups. Pen & paper exercises will be discussed on the board. Python programs must be submitted via Whiteboard, plots printed and handed in (write the names of the group members and their student numbers on the sheet).

Homework 1 (MLE and statistical testing, (programming counts 2/6))
Consider the Markov Model from the last assignment, task 2. You want to statistically assess whether the data actually supports a first-order Markov model (as depicted in the last assignment, task 2), vs. a simple coin flip. You thought of the following test

$$\mathcal{H}_0 : p^*_{00} \quad \geq \quad \tilde{p}^*_{00} \tag{1}$$
$$vs. \tag{2}$$
$$\mathcal{H}_1 : p^*_{00} \quad < \quad \tilde{p}^*_{00} \tag{3}$$

where p^*_{00} denotes the optimal parameter estimate for the input data and $\tilde{p}^*_{00}$ is an optimal parameter estimate if the original data is randomly (bootstrap) resampled; i.e. where the order of appearance of '0' and '1' is ignored, while their frequencies are somewhat *statistically* conserved. The same test can be performed with regards to parameter p_{11}.
a) [programming; submit via whiteboard] Write a program that computes the MLE estimate θ^* for the sequence of observations (y_i) provided in the file “Input.txt” via Whiteboard using 'scipy.optimize.minimize'. Set the seed of your random number generator to 'Seed.txt'; generate one (re-)sampled dataset (with replacement) using the function random.choices and compute the respective MLEs. For this, set your initial parameter guess to $p_{00} = \theta^0_1 = 0.2 = \theta^0_2 = p_{11}$. Write the MLE estimate into file "Exc7Task1a.txt" using 2 digits after the comma (same format as in reference file). Name your program Exc7Task1a.py and upload via Whiteboard.
b) [programming; print and hand in] Now, generate 1000 (re-)sampling datasets (with replacement) using the function random.choices and compute their respective MLEs using initial parameter guesses $p_{00} = \theta^0_1 = 0.2 = \theta^0_2 = p_{11}$. Plot a histogram an empirical cumulative density function (ecdf) of the MLE estimates from the 1000 bootstrap samples. Mark the values of p^*_{00} and p^*_{11} in those distribution plots. Compute the probability of the null hypothesis (as stated above), for p_{00} and p_{11} respectively and interpret this finding (max. 2 sentences).
Tipp: Define parameter bounds such that $0 < \theta_i < 1$ as previous.

Homework 2 (Newton-Algorithm, (programming counts 2/6 + 1/6 +1/6))
You are given the model:

$$x_i = \sin(\theta \cdot t_i) + x_0$$

with $x_0 = 2$. Your measurement data (t_i, y_i) is provided in the file "Data.txt", which you can load with np.loadtxt('Data.txt',ndmin=2,converters = float,delimiter=","). Assume that your data has an additive *gaussian* measurement error $\eta_i \sim \mathcal{N}(0, \sigma^2)$. Derive your objective function (the simplest form/have a look at the slides).
Implement the Newton method to estimate the parameter θ of the model. Stop the algorithm, if a maximum of 30 updates have been exceeded, or the parameters change by less than $\epsilon = 10^{-8}$, i.e. $||\theta^{s+1} - \theta^s||_1 = \sum_{j=1}^m |\theta^{s+1}_j - \theta^s_j| < \epsilon$. Set your initial parameter guess to $\theta^0 = 3$.
a) (**to be uploaded via Whiteboard**) Upload your Newton method as “Exc7Task2a.py”. The

program should generate an Output-file “Exc7Task2a.txt” that saves the parameter estimates θ^s obtained from the method (including the initial value).

$$2.315 \tag{4}$$

$$5.345 \tag{5}$$

b) (**to be printed and discussed**) Plot the value of your objective function, and its first two derivatives for the problem in task a) into one plot as a smooth function of $\theta \in [0, 20]$. Mark the y-value 0 as a reference.

- Which optima in your objective function $g(\theta)$ exist?

c) (**to be printed and discussed**) Plot model prediction vs. data for the parameters corresponding to the optima identified in b).

- Explain if/or if not the Newton algorithm is guaranteed to provide the global optimal parameter derived in a)?

Homework 3 (Least-Squares, pen & paper)
Assume a gaussian error model, i.e.

$$y_i = x_{i|\theta} + \eta_i$$

with $\eta_i \sim \mathcal{N}(0, \sigma_i^2)$.
Write down the likelihood function and then derive a least-squares criterion $g(\theta)$ that is proportional to your negative log-likelihood function.

Homework 4 (Optimization, pen & paper)
Given the least squares problem with $N = 2$ data points $t = (1, 3)$, $y = (6, 7)$.

$$g(\theta) = \sum_{i}^{N=2} (x_{i|\theta} - y_i)^2$$

with model

$$x_i = \theta_1 \cdot e^{\theta_2 \cdot t_i}$$

The estimable parameter $\theta^s = (\theta_1, \theta_2)^T = (1, 1)^T$. Compute one step of Newton's method $\theta^{s+1} = \theta^s - H^{-1}(\theta^s) \cdot \nabla g(\theta^s)$.

good luck! . . .