Final exam

STAT 517 - Winter 2023

1. Gerstein & Mandelbrot (1964) propose several point process models for neural spike data that are motivated by the similarities between the biological processes in the neuron and a Brownian motion. This question refers to the simplest of those models, which can be described in the following way:

   - Let the electrical state of polarization of the somatic and dendritic membrane of the neuron be specified by a single number. As time passes and the electrical state of the membrane varies, the state point will move back and forth along a straight line.
   - Choose a particular point on the line (say, 0) as the resting potential of the membrane.
   - Choose another point on the line, $z > 0$, as the "threshold". If at any time the state point reaches the threshold the neuron will produce an action potential.
   - The electric potential of the membrane varies over time as a Brownian motion without drift and volatility σ.
   - Immediately after the state point has attained the threshold and caused the production of an action potential, it returns to the resting potential, only to begin again on its random walk.

   Under this model, answer the following questions:

   (a) What is the distribution of the interspike times? Are the interspike times independent from each other? Make sure you justify your answers.

   (b) What is the distribution of $N(t)$, the number of spikes in the interval $[0, t]$ under this model? You can assume that the process starts at the resting potential.

   (c) How does your answer change if you drop the assumption that the process starts at the resting potential.

   (d) Are the parameters of this model identifiable? Recall that a the parameters of a model described by the probability distribution $p(x \mid \theta)$ are identifiable if and only if $p(x \mid \theta_1) = p(x \mid \theta_2)$ implies that $\theta_1 = \theta_2$ for all values of θ_1 and θ_2. If your answer is negative, what constraints could be introduced to the parameter space to make the parameters identifiable.

(e) Consider the data contained in the file `spikes.R`. Fit the point process described above to this data and provide point and interval estimators for the (identifiable) model parameter(s).

(f) Fit a homogeneous Poisson process to this data and provide a point and interval estimate for the parameters.

(g) Which of these two models seems to provide a better fit for the data?

2. **Sequential experimental design in an optimization setting.**. Consider the problem of finding the maximum of an unknown function $f : [a, b] \subset \mathbb{R} \to \mathbb{R}$. For example, f might represent the response of a physical system to a change in temperature. We assume that, while the function f is unknown, you can run some experiments in which a given input x is put into the system and a measurement $y = f(x)$ is collected. (We assume that these measurements are free of observational noise). This process can be repeated sequentially a finite number of times n. The value of n can be thought of as a *budget*.

   One way to determine the set of points on which to collect data is to decide in advance the values of $x_1, \ldots, x_n$ and then use the data to estimate f. For example, a common choice in one-dimensional settings like this one is to simply pick an equally spaced grid on $[a, b]$. An alternative is to pick the points sequentially. For example, you might start with a regular grid with $m \ll n$ points and, starting with the $m + 1$ point, apply some rule that tells you where to locate the next point in an "optimal" way one at a time.

   When the goal is to find the maximum of f, a widely used criteria for allocating new points is the expected improvement of the function f at a new point x^*. The expected improvement is defined as:

$$EI(x^*) = \mathsf{E}\left\{\max\{f(x^*) - y_{max,m}, 0\} \mid x_1, \ldots, x_m\right\}$$

   where $y_{max,m} = \max_{1 \le i \le m} \{f(x_i)\}$. The next point x^* is chosen to maximize $EI(x^*)$.

   (a) Assume that f is modeled using a Gaussian process with constant mean function and an exponential covariance matrix $\gamma(x, x') = \sigma^2 \exp\left\{-\frac{|x-x'|}{\lambda}\right\}$. Find a closed form expression for the expected improvement.

   (b) Can you find a closed form expression for $x_{m+1} = \arg\max_{x^* \in [a,b]} EI(x^*)$? If not, how can you determine x_{m+1} in practice?

   Assume now that $[a, b] = [0, 1.2]$ and you are interested in finding the maximum of the function $f(x) = (6x - 2)\sin(12x - 4)$. (While in this case you know the function and you could find its maximum exactly, we will treat it, for the purpose of items (c) and (d) below, as a "black-box" that you can evaluate at any point but do not completely know.)

(c) Assume you use the naive strategy of evaluating f on a regular grid with with n points as your "design". Provide a point and interval estimates for f, as well as your best guess to the location and value of the maximum of f.

(d) Implement the sequential strategy starting with $m = 4$ equally spaced points to obtain a total of $n = 15$ "optimal" points. Again, provide a point and interval estimates for f, as well as your best guess to the location and value of the maximum of f.

(e) Which method seems to perform best in finding the maximum?

Please note that both μ and λ are unknown and need to estimated from the data each time.

References

Gerstein, G. L., & Mandelbrot, B. (1964). Random walk models for the spike activity of a single neuron. Biophysical journal, 4(1), 41-68.