**Exercises 3 – Advanced statistics – Tuesday, 13th January 2015**

*Please return the results by next week, Tuesday 22th Jan 2015 1pm, with name and student number on each page. Exercises should be done individually. They can be send to* `m.r.feyereisen@uva.nl` in a single mail. *The whole sheet is worth 40 points.*

1. Signals and backgrounds in low-number statistics.

   (a) Consider a Poisson process with a known background $b$ and a unknown signal contribution $\mu \geq 0$, such that the average number of expected counts are $\langle c \rangle = \mu + b$. Suppose we measure a value of $c = 3$, whereas the expected background is $b = 1.5$. What is the $p - value$ of this 'excess'? (1pt)

   (b) Derive, based on the previous numbers, a proper 95% CL *upper* limit and a proper 95% CL *lower* limit on $\mu$. Derive also the 68% CL central confidence interval on $\mu$. (4pt)

2. Minimizers – Gradient decent

   In this exercise, we will concentrate on the two-dimensional function

   $$f(x, y) = (x - 2)^2 + (y - 3)^2 + xy \ ,$$

   (a) As a warm up, show analytically that the minimum of the function is given by $x = \frac{2}{3}$ and $y = \frac{8}{3}$. (1pt)

   (b) Write a gradient decent algorithm in order to minimize the function. Select three or more random points in the range $-10 \leq x, y \leq 10$, and set $\gamma$ to a reasonable value. Plot the first twenty steps that the minimizer takes when approaching the minimum for each of the three initial conditions. For what values of $\gamma$ does this provide good convergence (visual inspection of the plot is enough)? (5pt)

   *Hint:* Remember that first-order derivatives can be estimated numerically by e.g. $\frac{\partial}{\partial x} f(x, y) = (f(x + dx, y) - f(x, y))/dx$, with a suitably chose $dx$. You can instead also use the analytical derivatives of $f(x, y)$ in the algorithm, $\partial f(x, y)/\partial x$ and $\partial f(x, y)/\partial y$.

3. Minimizers – Simulated Annealing

   We will consider the one dimensional function

   $$f(x) = \frac{x^2}{100} - \cos(x) \ ,$$

   in the range $-10 \leq x \leq 10$.

(a) Set up the simulated annealing algorithm as discussed in the lecture. As proposal distribution we simply take a flat distribution, $g(x \to x') = \text{const}$ (this means that proposed points are uniformly randomly distributed in the range $-10 \leq x \leq 10$). For the acceptance propability

$$A(x \to x') = \min(1, \exp(-f(x')/T + f(x)/T)) \tag{1}$$

select the temperature $T = 1.0$. What fraction of proposed points is accepted at this temperature? (5pt)

*Hints:* The algorithm has the following steps

   i. Start with a random point in the range $|x| < 10$ and save it in a list.

   ii. Select a proposal point $x'$ in the range $|x'| \leq 10$.

   iii. Calculate the acceptance rate $A(x \to x')$, and accept the new point with the propability that is given by $A(x \to x')$. A simple way to do that is to draw a random number in the range $r \in [0, 1]$, and accept the point $x'$ if $r \leq A(x \to x')$ (acceptance rejection method).

   iv. If the point is accepted, set $x = x'$, and save it in the list. Otherwise, do nothing. In any case, go back to step ii (or break the loop if the enough iterations were performed).

(b) Plot histograms of the accepted points for different temperatures. What temperature is sufficiently low to singlet out the central global minimm of the function $f(x)$? (4pt)

4. Searching a "X-ray" line

This work is based on unbinned mock data that is available online. The file contains a list of events in ASCII format, with only the energy information given for each event. The energy ranges from 1 to 10 keV. The file contains the background flux as well as a known line at around 3.0 keV.

We assume that the detector has an energy resolution of $\Delta E/E = 10\%$, and the energy dispersion can be described by a normal distribution. The energy spectrum is then given by

$$\frac{dN}{dE} = A_s \cdot N(E|\bar{E}, \Delta E) + \frac{dN_B}{dE} \ , \tag{2}$$

where $\bar{E} = 3.0\,\text{keV}$, and $\Delta E = 0.1\bar{E}$. Here, $N(x|\mu, \sigma)$ is the normal distribution, $A_s$ the signal normalization, and we assume that the background rate $dN_B/dE = 100\,\text{keV}^{-1}$ is known.

(a) Read the event lists into python (`numpy.loadtxt`) and bin them into 50 linear bins (`numpy.histogram`). Plot the histogram, together with a simple estimate of the flux variance as $\pm 1\sigma$ errors (`pylab.errorbar`). (3pt)

(b) Perform a fit to the data set (`scipy.optimize.fmin_bfgs`), using as model the above full model with a line signal at 3.0 keV (one free parameter, the normalization). As likelihood function, take the Poisson likelihood

$$\mathcal{L} = \prod_{i=1}^{n_{\rm bins}} P(c_i|\mu_i) \,, \tag{3}$$

where $\mu_i$ are the events expected in energy bin $i$, and $c_i$ are the actually observed events. You can obtain $\mu_i$ by numerically integrating $dN/dE$ over the energy bins (`scipy.integrate.quad`). What is the maximum-log-likelihood-ratio of the background-only ($A_s = 0$) and the background-plus-line model? Can you claim a detection of a 3.0 keV line with a significance of $5\sigma$? (7pt)

(c) Perform a fit to the data set, and calculate an 95.4% CL upper limit on $A_s$ for a line at $\bar{E} = 8\,\text{keV}$. To this end, increase $A_s$ from its best-fit value until $-2\ln\mathcal{L}$ changes by 2.71. (5pt)

(d) What constraint (68% CL) do you obtain on the *energy* $\bar{E}$ of the line around 3.0 keV? Why is the constraint so much smaller than the $\Delta E/E = 10\%$ energy resolution of the detector? (5pt)

*Hint:* Repeat the fit at different line energies $\bar{E}$ around 3.0 keV and plot the $-2\ln\mathcal{L}$ as function of $\bar{E}$. The 68% CL confidence band corresponds to an increase of $-2\ln\mathcal{L}$ of one.

(e) *Optional (for the enthusiasts):*

i. *Optional:* Instead of fixing the background to a constant rate, include its normalization as an additional free parameter in the fit. How does this affect the results?

ii. *Optional:* Instead of the binned analysis, perform a unbinned analysis. How does this affect the results?