MOL 410/510: Introduction to Biological Dynamics Fall 2011
Problem Set #8 (due 12/9/2011)
3 Questions, all are **MUST DO**

1. **Time Series, smoothing, filtering, and fourier transforms.** Download and unzip the Matlab data from the homework and solutions section of the course website. You will find a file called seriesA.mat. Load the file into Matlab, it contains a vector called A which contains 100 elements from a computer random number generator.

   (a) Calculate the mean and standard deviation of A.

   (b) Is the mean of A statistically different from 0?

   (c) Now find if the first ten elements are statistically different from 0. Do the same for the last ten elements.

   (d) Smoothing time series. Often, time series data is too noisy to show for publication. That is, we want to convey the overall trend in the time series, but the measurements are so noisy that they make trends difficult to observe. We would like to "smooth" over the noise without fundamentally changing the trends in the data. This can be accomplished through convolution. In the next steps, you will smooth the time series in A.

      i. To smooth the data, convolve the time series with a Gaussian kernel. To generate a gaussian kernel, you can use the Matlab function called normpdf (use Matlab's help function to learn more). To perform the convolution, use the Matlab function conv. That is, if the values of your Gaussian kernel is stored in a vector named kern, call A1=conv(A,kern,'same') to perform the convolution (the string switch 'same' tells Matlab that A1 should be the same size as either A or kern, whichever is bigger). Perform the convolution for four different sized kernels, i.e. with standard deviation 1, 3, 10, and 100. Use -10:1:10 as the x-input into normpdf (note: we are assuming here that the time intervals in A are incremented by 1).

      ii. Plot the original time series and the four smoothed time series on the same axes. Use t=1:100 as your x-data. What is the difference between the time series as you increase the standard deviation?

      iii. Use Matlab's fft function to calculate the Fourier transform of the original time series and all of the four convolved time series.

      iv. Plot the power spectra for all time series on the same axes (be sure to label each line clearly). What happens to the power spectrum as you increase the standard deviation of the convolution kernel? Be careful to label your frequency axes properly; if it helps, use fftshift to order the vector that fft outputs more sensibly.

2. **Approximation by frequency components.** Take the following function of time:

$$f(t) = 1 \text{ if } |t| < 10, \;\; 0 \text{ otherwise.} \tag{1}$$

   (a) Choose a sampling rate and range for the function and use Matlab's `fft.m` to compute its Fourier transform.

   (b) Now approximate the function using the real part of successive frequency components, that is, approximate the function $f(t)$ using

$$f(t) \approx \frac{1}{N}\left(\mathrm{Re}(z_0) + 2\sum_{k=1}^{N_c}\mathrm{Re}(z_k)\cos(2\pi w_k t)\right) \tag{2}$$

   where z_k is the the k^{th} frequency component of the Fourier transform of $f(t)$, N_c is the number of (non-DC) frequency components used in the approximation, and N is the total number of samples (points) in $f(t)$. Plot the approximate value of $f(t)$ using successive values for N_c (i.e. $N_c = 0, 1, 2...$). How many components do you need to obtain a good approximation of $f(t)$?

   (c) Now calculate the analytical Fourier transform of the continuous function $f(t)$.

3. **Diffusion Meets Maximum Likelihood.** A group of 10 fluorescent proteins are released at time $t = 0$ at position $x = 0$ in a one-dimensional microfluidic channel. At time $t = 60$ seconds, the proteins are imaged as fluorescent spots at $x =$ -110, -80, -40, -30, -10, 10, 20, 40, 70, and 130 microns.

   (a) What are the mean μ and variance σ^2 of these positions (to estimate the variance use $\sigma^2 = (1/N)\sum_{i=1}^{N}(x_i - \mu)^2$)?

   (b) Assuming the proteins move by diffusion, what is the maximum likelihood estimate from the data of the diffusion constant D?

   (c) What is the range of possible D values, $D_{\min}$ to $D_{\max}$, estimated from ln[Likelihood($D_{\min}$)] = ln[Likelihood($D_{\max}$)] = ln(Maximum Likelihood) - 2.

   (d) For your maximum likelihood value of D, what is the probability distribution of protein positions at time $t = 60$ seconds? What is the variance of this distribution?

   (e) If a set of N data values are drawn from a Gaussian probability distribution of unknown mean and variance, show that the maximum likelihood estimates of the mean and variance of the Gaussian are just the sample mean and the sample variance.