Coding in advance of the Apr 29, 2019 class

For the final three coding exercises, we will analyze the same data set: the set is given as data13_1.txt on the website, and is taken from Table 1 of Ambrosi et al. 2017, Nature, 552, 63. The lure here is the possibility that the bump at around 1.4 TeV in that data might represent something meaningful, e.g., dark matter. Odds are it doesn't, but we'll analyze it to find out!

The file description13.txt describes the data; essentially, we have columns that give the average energy in an energy bin ($E_{\rm avg}$), the flux in that bin (F), the reported statistical uncertainty for that flux ($\sigma_{\rm stat}$), and the reported systematic error in that bin ($\sigma_{\rm sys}$). We have been handed already-processed data; we thus are forced to make assumptions regarding the nature of those uncertainties. In particular, we will assume that $\sigma_{\rm stat}$ and $\sigma_{\rm sys}$ both represent Gaussian uncertainties, to an arbitrary number of standard deviations, and that these uncertainties add in quadrature so that the total uncertainty is

$$\sigma_{\rm tot} = \sqrt{\sigma_{\rm stat}^2 + \sigma_{\rm sys}^2}\ . \tag{1}$$

In a real analysis, if we came up with apparently strong evidence that the bump is a real feature, we would want to proceed via the actual, raw, data, so that we didn't have to make these simplifying assumptions.

Your first task will be to use the affine-invariant MCMC code that I sent you to fit a simple power law model to the data, for which the flux is

$$\Phi = \Phi_0 (E_{\rm avg}/100\ {\rm GeV})^{-\gamma} \tag{2}$$

and thus the two parameters in your model are the normalization Φ_0 and the power-law index γ. Yes, for a two-parameter model you could do a grid search, but as we proceed we'll need to explore more complex models with more parameters, so please use the affine-invariant code even for this simple model.

Note that the output of your code would allow us to do full explorations of the posterior; for example, we could get the full two-dimensional posterior probability density, we could find ways to represent the 68% (or other) credible region in two dimensions or for each parameter separately, and so on. If you feel like performing such analyses, more power to you! But for this particular purpose, let's just aim to get the maximum likelihood for this model; what values of Φ_0 and γ maximize the log likelihood, and what is the value of that maximum log likelihood? In the next two weeks we'll compare that value with what we get with more complicated models, and will use Wilks' Theorem to judge whether the additional model parameters are needed.

As a check on your fit, please plot your best fit against the data; is it reasonable?

Good luck!