Computational Physics / PHYS-GA 2000 / Problem Set #6
Due October 31, 2023

You *must* label all axes of all plots, including giving the *units*!!

1. Calculate derivative of the function:

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-x^2/2\right) \tag{1}$$

three ways: (a) analytically; (b) with second-order finite difference; and (c) with `autodiff` using the `jax` implementation. Compare the results, and also compare the speeds.

2. Here we will perform a simple Principal Components Analysis on a real data set. Download this file which contains the central optical spectra of 9,713 nearby galaxies from the Sloan Digital Sky Survey (note that although this file is 150 Mb, this is a small sample! larger data sets exist of millions, though they are lower quality). I have done some the work for you by interpolating all of the spectra onto the same restframe wavelength grid. Now do the following:

   (a) Read the data in using the `astropy` package, in particular using `astropy.io.fits`. This data set is a special format called the Flexible Image Transport System (FITS) format, common in astronomy. Its only virtue is that it is a standard in astronomy. You should be able to `pip install astropy`, and then:

```
hdu_list = astropy.io.fits.open('specgrid.fits')
logwave = hdu_list['LOGWAVE'].data
flux = hdu_list['LOGWAVE'].flux
```

   `logwave` will be $\log_{10}\lambda$ for λ in Angstroms. `flux` will be in 10^{-17} erg s^{-1} cm^{-2} Å^{-1}, and is the spectrum. Plot a handful of the galaxies.

   (b) There are two processing steps that will make the PCA more meaningful. First, all of these galaxies are at different distances, so their fluxes span a large dynamic range; so first normalize all the fluxes so their integrals over wavelength are the same.

   (c) Second, the mean flux at every wavelength is positive; this will mean that a PCA will spend an eigenvector to explain the mean offset from zero. Instead, we will first subtract off the mean $\vec{f}_m$. This will leave residuals $\vec{r}_i = \vec{f}_i - \vec{f}_m$ of all the galaxies i varying around zero.

   (d) Now perform the PCA. The idea of the PCA is to find the eigenvectors of the covariance matrix of the distribution. This covariance matrix *can* be calculated as follows:

$$\mathbf{C} = \frac{1}{N_{\mathrm{gal}}} \sum_{ij} \vec{r}_i \vec{r}_j \tag{2}$$

where i and j index the galaxies. If I recast the residuals as a matrix R_{ij} this is $\mathbf{R} \cdot \mathbf{R}^T$. So construct this matrix (it should be $N_{\rm wave} \times N_{\rm wave}$, and find its eigenvectors. Make a plot of the first five eigenvectors.

(e) It is also possible to find these eigenvectors directly from $\mathbf{R}$ using SVD. Consider the linear problem, which finds a set of coefficients $\vec{x}$ to multiply the given spectra by, to explain some spectrum $\vec{f}$:

$$\mathbf{R} \cdot \vec{x} = \vec{f} \tag{3}$$

We know the SVD decomposition of $\mathbf{R}$ yields a rotation $\mathbf{V}$ into the space where the covariance matrix of the uncertainties in $\vec{x}$ is diagonal. This covariance matrix is $\mathbf{R}^T \cdot \mathbf{R}$. So now find the eigenvectors using an SVD decomposition of $\mathbf{R}$ and show that the vectors are equivalent to what you found before.