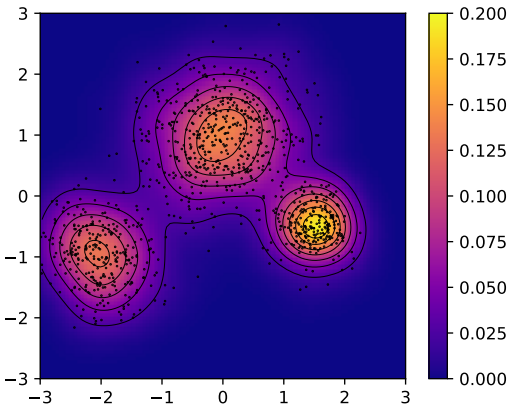
Visualization
Prof. Bernhard Schmitzer, Uni Göttingen, summer term 2022

Problem sheet 3

- *Submission by **Tuesday** 2022-06-07 18:00 via StudIP **as a single PDF/ZIP**. Please combine all results into one PDF or archive. If you work in another format (markdown, jupyter notebooks), add a PDF converted version to your submission.*
- *Use Python 3 for the programming tasks as shown in the lecture. If you cannot install Python on your system, the GWDG jupyter server at https://jupyter-cloud.gwdg.de/ might help. Your submission should contain the final images as well as the code that was used to generate them.*
- *Work in groups of up to three. Clearly indicate names and matrikelnr of all group members at the beginning of the submission.*

Exercise 3.1: density estimation and countour plot.

1. The file `data_points.mat` contains an array `data` of dimensions $N_{\text{sample}} \times 2$ of $N_{\text{sample}} = 10,000$ points sampled from a Gaussian mixture model in 2d, and then truncated to the box $[-3,3]^2$. Import the array into python.
2. As shown in the lecture, perform a Gaussian kernel density estimation (e.g. with `sklearn`) on $[-3,3]^2$. Choose a reasonable width of the kernel and evaluate the density on a fine grid.
   Hint: Kernel width and the resolution of the evaluation grid can be chosen independently.
3. Show the estimated density as color coded image (with a perceptually uniform scale and a colorbar legend), the samples as scattered points and contours of the estimated density. The result may look similar to this:



Exercise 3.2: smoking and life expectancy.

1. The file `data_smokers.mat` contains an array `data` of dimensions $N_{\text{pers}} \times 3$ of type `int` which contains information about $N_{\text{pers}} = 20,000$ persons from $N_{\text{countries}} = 20$ countries. Each row represents one person. The first column encodes the country that they live in, by an integer from 0 to 19. The second column encodes whether that person was a regular smoker (=1) or not (=0). The third column gives the age in full years that this person reached at the time of their death. Import this array into python.

2. Plot histograms over ages for the total population, for smokers and non-smokers (with absolute counts in each bin). In addition, plot the normalized histograms (where entries in all bins sum to one), which represent an approximate probability density function.

3. For each country, determine the average life expectancy of people and the fraction of smokers. Visualize this information.

4. For each country, determine the life expectancy of smokers and non-smokers. Visualize this information.

5. Generate a 2d histogram of people over their country and their age, for smokers and non-smokers. Find a way to visualize this in a single plot as a multi-color image.
   Hints: Think about a good way of normalizing the color channels. Think about a reasonable ordering of the countries.

6. For smokers and non-smokers, the relation 'country that a person lived in' to 'age of that person' is a stochastic functional relation. Visualize this relation for both groups (smokers, non-smokers) to obtain a single chart which conveys the information how long smokers and non-smokers tend to live in various countries and how large the variation of ages is.