Report7

胡琦浩　PB21000235

一、问题

对一个实验谱数值曲线 *p*(*x*) ，自设 *F*(*x*)，分别用直接抽样和舍选法对 *p*(*x*) 抽样。比较原曲线和抽样得到的曲线以验证。讨论抽样效率

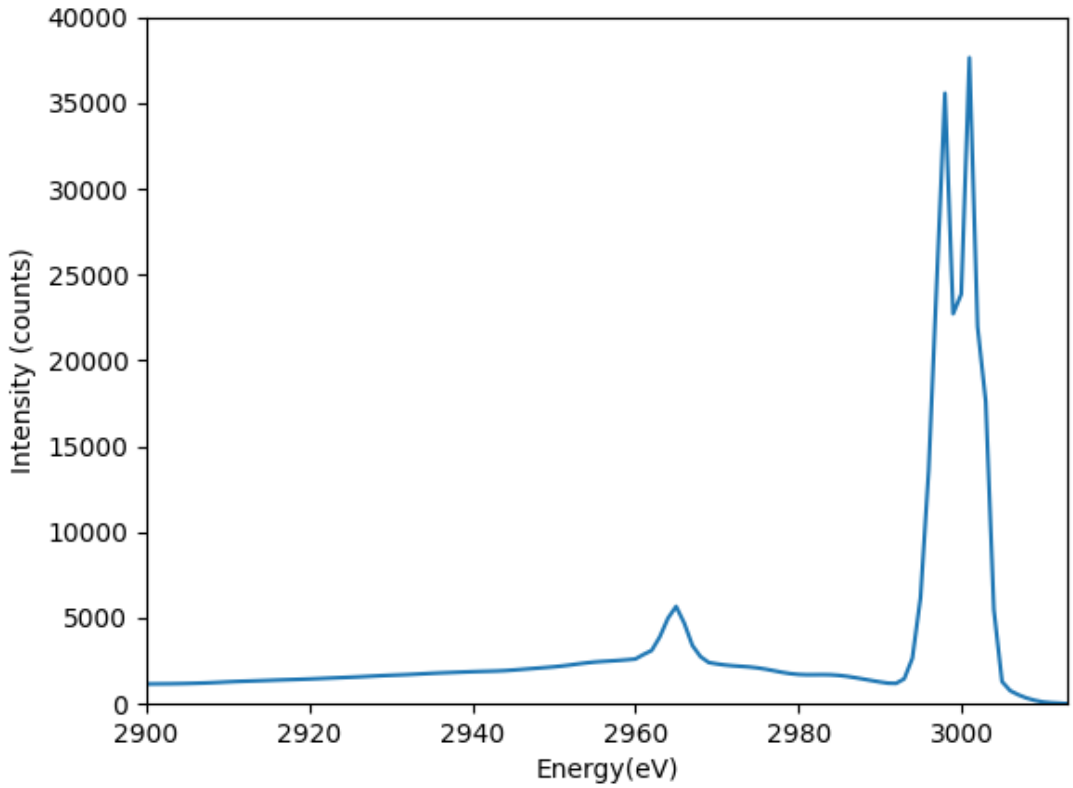


图1: 原实验数据谱线

二、方法

2.1 数学推导

2.1.1 直接抽样法

实验数据为离散型，x范围为[2900, 3013]，$p(x_i) = \frac{n(x_i)}{N}$，得到[0,1]之间的随机数ξ满足:

$$\sum_{i=1}^{n-1} p_i < \xi \leq \sum_{i=1}^{n} p_i$$

，则取$x = x_n$

2.1.2 舍选抽样法

根据$n(x)$的大小,将x分为3个区间:[2900, 2993]，[2993, 3006]，[3006, 3013]

第一个区间[2900, 2993]。$n(x)_{max} = 5672$,不妨利用简单分布舍选法(ξ_1与ξ_2均为[0,1]的随机数)。

取$M_1 = 5672$, $\xi_1 = \frac{\xi_x - 2900}{2993 - 2900}$， $\xi_2 = \frac{\xi_y}{M_1}$

故$\xi_x = 93\xi_1 + 2900, \xi_y = M_1\xi_2$

第二个区间[2993, 3006]。显然这个区间用高斯分布函数来覆盖。经过测试参数,选择:

$$F(x) = 39000e^{-\frac{(x-3000)^2}{100}}$$

由舍选法:

$$\xi_3 = \frac{\int_{2993}^{\xi_x} F(x)\,dx}{\int_{2993}^{3006} F(x)\,dx}$$

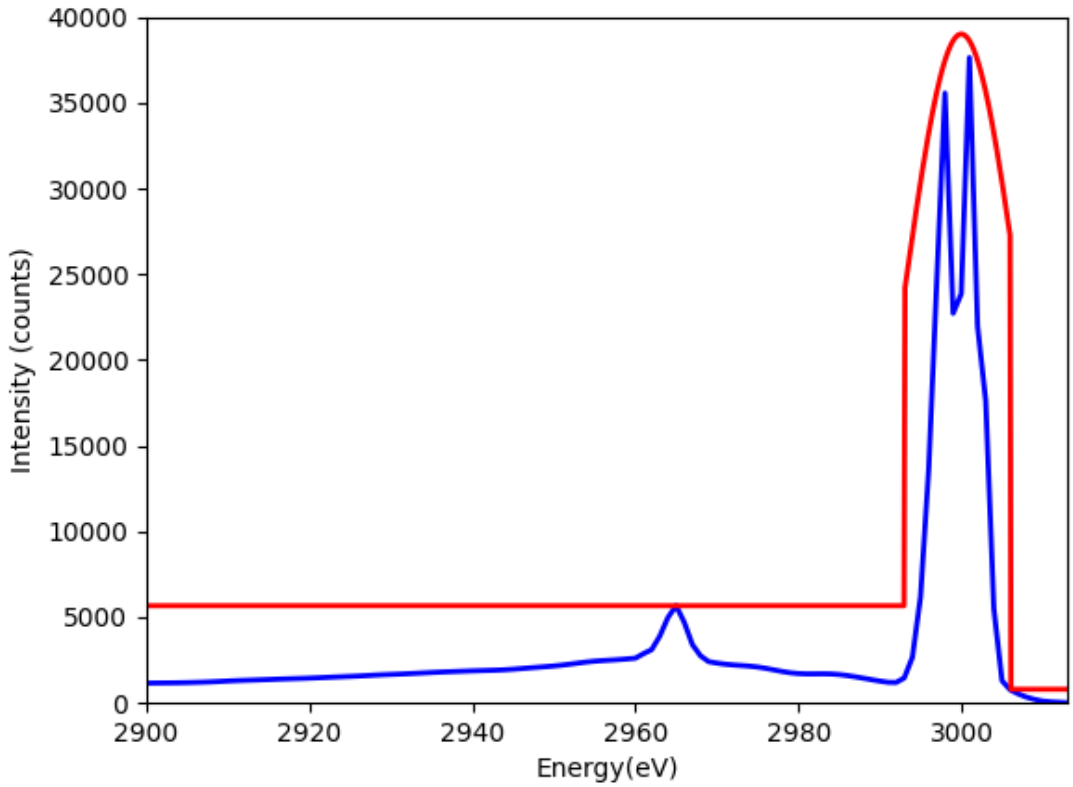
$$\xi_4 = \frac{\xi_y}{F(\xi_x)}$$

由于无法求出$\int F(x)\,dx$的解析表达式,故采用数值求解法得到ξ_x与ξ_3的对应关系。

第三个区间[3006, 3013]。同区间一的做法:

取$M_2 = 800,\ \ \xi_5 = \frac{\xi_x - 3006}{3013-3006},\ \ \xi_6 = \frac{\xi_y}{M_2}$

故$\xi_x = 7\xi_5 + 3006,\ \ \xi_y = M_2\xi_6$



如图所示, 选择的曲线能将原始数据包含

2.2 代码实现

直接抽样法较简单，不做赘述。

舍选抽样法中得到的样本值不一定为整数，则比较大小时，我们无法得知原数据中能量值不为整数时所对应的数目是多少，故采用python中的scipy库来解决

```
from scipy import interpolate

# 样条插值法
energy = read_txt()[0]
num = read_txt()[1]  # 储存每个energy的数量
spline = interpolate.splrep(energy, num, s=0)

num1 = interpolate.splev(elem1, spline)
```

利用这个库，我们就可以估计原数据中能量为elem1时的数量大约为num1。方便舍选法时作比较。

此外求ξ_3与ξ_x的对应关系时

记：

$$G(x) = \frac{\int_{2993}^{x} F(t)\,dt}{\int_{2993}^{3005} F(t)\,dt}$$

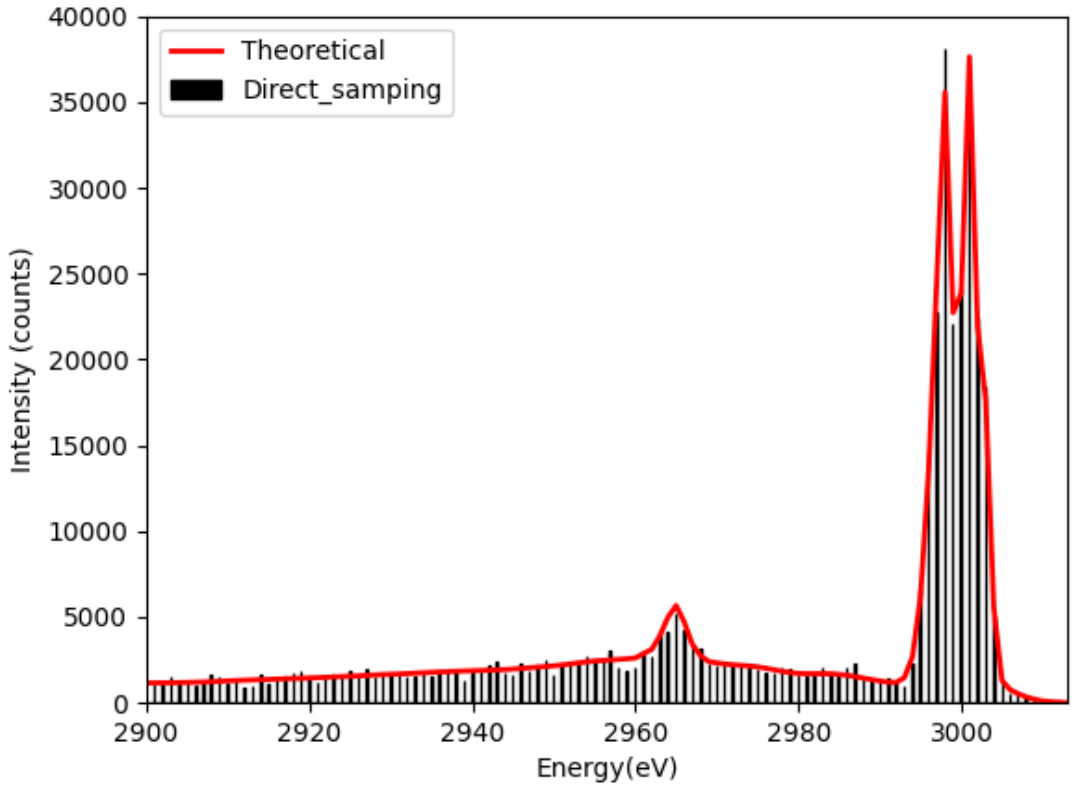
$$G_1(x) = G^{-1}(x)$$

在python中直接用scipy库中的integrate函数求G(x), 显然G(x)为单调递增的函数, 因此用二分法求G(x)的反函数G_1(x), 得到ξ_3与ξ_x的对应关系。

画概率直方图与前几题类似

三、实验结果

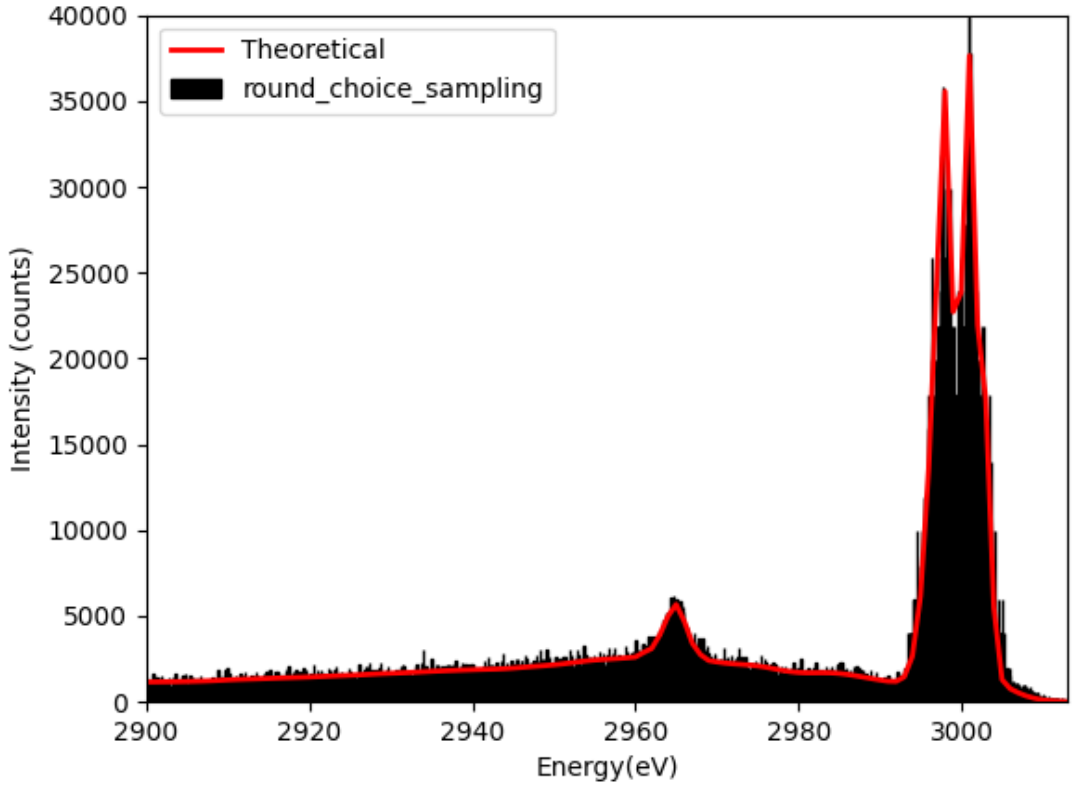
3.1 直接抽样法



由于为直接抽样，故效率为100%

但我们得到的样本值都为整数，离散的，然而实际中的值为连续的。

3.2 舍选抽样法




```
舍选法区间1点数为N1=100000,抽样样本为n1=34075,效率为0.34075
舍选法区间2点数为N2=10000,抽样样本为n2=4873,效率为0.4873
舍选法区间3点数为N3=10000,抽样样本为n3=2961,效率为0.2961
舍选法总点数为N=120000,抽样样本为n=41909,效率为0.3492416666666667
```

可以看出简单的采用区间最大值来进行抽样，效率不会很高。

区间2采用高斯函数来抽样，效率明显高于其他两个区间

四、总结

对于直接抽样，虽然效率高，但得到的是离散的样本数据，不过可以采取插值法来估计中间值

而对于舍选抽样，虽然可以得到连续的样本，但抽样效率不太高，不过可以通过选择更加贴合原始数据曲线的函数来提高抽样效率

因此两种方法各有优劣，在实际问题中，我们需要选择合适的抽样法去解决问题。