计算物理第7题

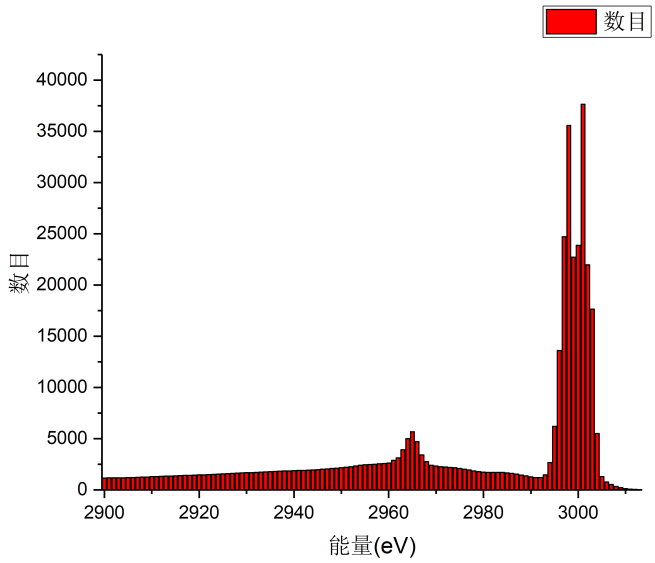
PB18000039 徐祺云

一　作业题目

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。

二　算法及主要公式

在ORIGIN中处理数据文件$data.TXT$得到能量分布图如下：



观察实验谱数值曲线性质：在$[2900, 3013]$电子伏特能量范围内，能谱分布有三个尖峰：在能量为2965eV处第一个小峰值为5672；在能量为2998eV处第二个峰值为35560；在能量为3001处第三个峰值为37630,其中后两个尖峰的能量差较小，比较接近，取$F(x)$时可考虑为一个峰。

这里采取直接抽样和舍选抽样两种抽样方法：

直接抽样法

能量分布是离散型的，取值为$2900, 2901, ..., 3013$，相应值出现的几率为$p_i, i = 1, ..., 114$(自变量取值共有114个)，先把概率归一化，即取截面值σ，从而$p_i = \sigma_i / \sum \sigma_i$，得到概率密度分布函数：

$$f(x) = \sum_{i=1}^{N} p_i \delta (x - x_i)$$

累积分布函数：

$$y_k = F(x) = \int_{-\infty}^{x} f(t)\, dt = \int_{-\infty}^{x} \left(\sum_{i}^{N} p_i \delta (t - x_i) \right) dt = \sum_{i=1}^{k} p_i$$

以上部分是对原始数据$data.TXT$的预处理。

利用16807随机数发生器在[0,1]区间内均匀抽样得到随机数序列$\{\xi_i\}$，对于每一个ξ_i，若找到满足$y_{k-1} < \xi_i \leq y_k$的k值，则随机变量的第k个取值即为欲抽取的值。如此这样操作，最后可以得到抽样曲线。

舍选抽样法

采用直接抽样法时会遇到较大困难，主要是数据的密度分布函数本身是以数值表的形式给出的，对此von Neumann发展了一个简单实用的方法，即舍选法，它不需要计算累计函数。

舍选法的一般思想是，在二维平面随机抽样(x,y)，判断(x,y)点是否在分布曲线下方，若是则选，否则舍去。但是，对于题目给定的数据，分布曲线$p(x)$呈尖峰形状时，抽样效率会很低，这是需要把变换抽样与舍选法结合，引入处处比$p(x)$大的比较函数$F(x)$，在比较函数的面积区内产生随机数点抽样，这样可以提高抽样效率。

为覆盖两个尖峰，考虑取分段函数：

$$F(x)=\begin{cases}2000, x\in[2900,2946)\\ 5672, x\in[2946,2977)\\ 2000, x\in[2977,2994)\\ 37630, x\in[2994,3005)\\ 2000, x\in[3005,3013]\end{cases}$$

近似地，这里认为$data.TXT$中的离散数据可以看成分段阶梯函数，即：对于点x_n，对应数目为y_n，令$\forall x\in[x_n,x_n+1)$有$p(x)=y_n$。

如此定义$p(x)$,$F(x)$后开始抽样：随机抽取$[0,1]$间均匀分布的随机数(ξ_1,ξ_2)，在$F(x)$的面积区内抽取随机数ξ_1，由

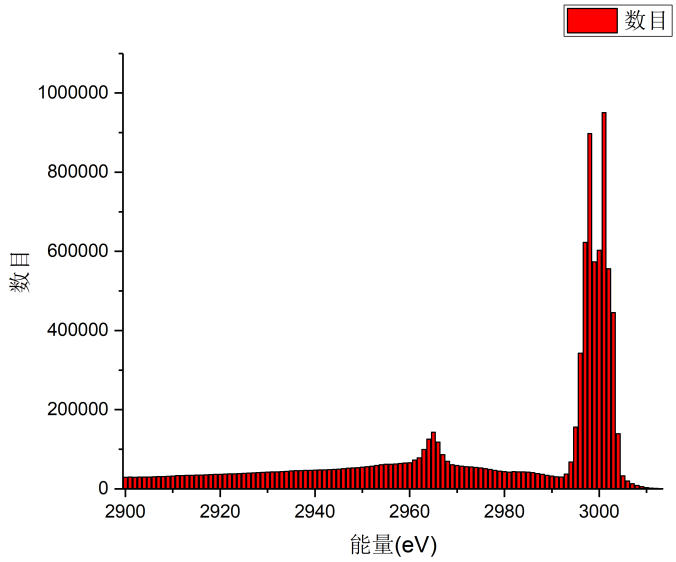
$$\xi_1=\int_a^x F(t)dt/\int_a^b F(t)dt, \xi_2=y/F(x);$$

得到(x,y)，再关于处于$p(x)$面积区上/下进行舍选。

三　计算结果与分析

直接抽样法

取$N=10^7$，初始化处理好概率函数、累积分布函数后抽样，结果导入ORIGIN软件处理如下：



与原数据能量分布图(第一页)对比，可以明显地看到2965附近的一个小峰，以及3000左右的两个尖峰，说明该直接抽样是良好的。但是，直接抽样需要消耗时间空间去初始化各个离散点的概率函数与累积分布函数，使得程序的效率不高。

舍选抽样法

令$H(x)=\int_{2900}^{x}F(t)dt$，得到：

$$H(x)=\begin{cases}2000(x-2900), x\in[2900,2946)\\ 5672(x-2946)+92000, x\in[2946,2977)\\ 2000(x-2977)+267832, x\in[2977,2994)\\ 37630(x-2994)+301832, x\in[2994,3005)\\ 2000(x-3005)+715762, x\in[3005,3013]\end{cases}$$

归一化系数$\int_{a}^{b}F(t)dt=731762$，把$[0,1]$间均匀分布的$\xi_1$ 扩展到$[0,731762]$间均匀分布的$z(z=731762\xi_1)$，求反函数$x=W(z)=H^{-1}(z)$：

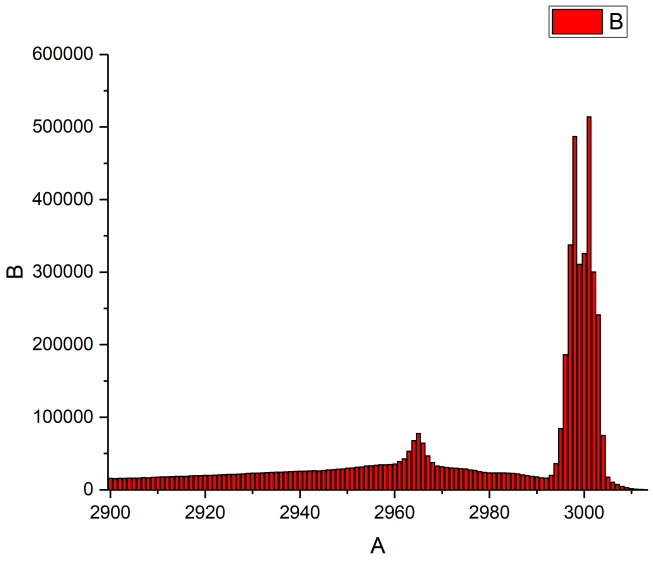
$$x=W(z)=\begin{cases}\dfrac{z}{2000}+2900, z\in[0,92000)\\ \dfrac{z-92000}{5672}+2946, z\in[92000,267832)\\ \dfrac{z-267832}{2000}+2977, z\in[267832,301832)\\ \dfrac{z-301832}{37630}+2994, z\in[301832,715762)\\ \dfrac{z-715762}{2000}+3005, z\in[715762,731762]\end{cases}$$

然后求出$y=F(x)\xi_2$与$p(x)$比较舍选即可。

取$N=10^7$，统计舍选个数，计算与N的比值即可得到实验的抽样效率：

$$\eta=\frac{SelectNum}{TotalNum}=54.1061\%$$

抽样结果导入ORIGIN软件处理如下：



与原数据能量分布图(第一页)对比，趋势大体相同，可以明显地看到2965附近的一个小峰，以及3000左右的两个尖峰，说明该直接抽样是良好的。选取了五段的分段比较函数$F(x)$，抽样效率达到了54.1061%，效率良好；若想继续提高抽样效率，可继续划分更细致的分段函数区间，使得比较函数趋近于$p(x)$，面积区之比$S_{p(x)}/S_{F(x)}$向1逼近。

四　结论

本题练习了离散分布下，$p(x)$有尖峰情况的抽样。先熟悉了直接抽样法，计算了概率密度分布函数和累积分布函数，得到了较为良好的结果；采用舍选抽样时，选取了五段分段阶梯函数作为比较函数$F(x)$，图像和原能谱分布吻合得比较好；抽样效率约为54%，若想继续增大抽样效率，可以继续划分更细致的分段函数区间，使得面积区之比增大，但会带来更繁琐的累积分布与反函数的求解。