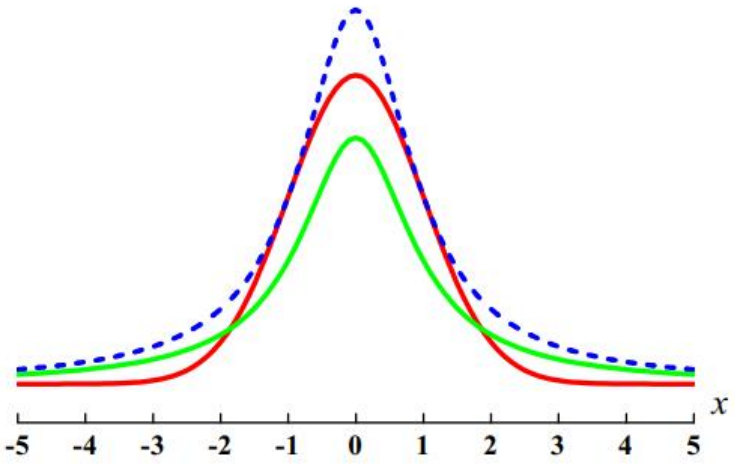
# 09：舍选抽样法

许传奇 PB16021546

## 1 题目

对两个函数线型（Gauss分布和类Lorentz型分布），设其一为 p(x)，另一为 F(x)，用舍选法对 p(x)抽样。将计算得到的归一化频数分布直方图与理论曲线 p(x)进行比较，讨论差异。讨论抽样效率。

$$Gaussian: exp(-\frac{x^2}{2}) \qquad Lorentzian\ like: \frac{1}{1+x^4}$$




## 2 原理与算法

### 2.1 原理

#### 2.1.1 舍选抽样法

在某些情况下，我们无法采用直接抽样法和变换抽样法来进行抽样，这时我们可以采取舍选抽样法。

舍选抽样法的思想是：对于一个难以用常规方法抽样的分布（设其概率密度函数为$f(x)$），我们选取另一个容易抽样的函数（设其概率密度函数为$g(x)$）进行抽样，得到$g(x)$的随机抽样。然后再对这个分布内的点进行舍去和选择，就能得到$f(x)$的抽样。

具体步骤如下：

1. 对$g(x)$进行抽样，得到$\xi_x$（如在[0,1]上抽取$\xi$，再根据累计函数的反函数求得$\xi_x$，就得到了$g(x)$的抽样$\xi_x$）；
2. 抽取在[0, g($\xi_x$)]均匀分布的$\xi_y$，比较$\xi_y < f(\xi_x)$是否成立；
3. 如果不成立，则返回到（1）；如果成立，则$\xi_y$就是f(x)的抽样。

$g(x)$选取的原则是：容易求得其抽样；在抽样的区间内满足：$f(x) \leq g(x)$

### 2.1.2 用Lorentzian like分布函数抽取Gaussian分布函数

首先，因为指数下降更快，为了使$p(x) \leq F(x)$在实轴上成立，只能用Lorentz like分布函数来抽取Gaussian分布函数。设Gaussian分布函数为要抽取的p(x)，Lorentz like分布函数为F(x)。

对Gaussian分布函数进行归一化：

$$\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx = \sqrt{2\pi} \tag{1}$$

因此有：

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{2}$$

1. F(x)系数的选取：

为了使F(x)满足$p(x) \leq F(x)$，设：

$$F(x) = c\frac{1}{1+x^4} \tag{3}$$

由$p(x) \leq F(x)$，有：

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \leq c\frac{1}{1+x^4} \Leftrightarrow c \geq \frac{1}{\sqrt{2\pi}} max\{(1+x^4)e^{-\frac{x^2}{2}}\} \tag{4}$$

设$\phi(x) = (1+x^4)e^{-\frac{x^2}{2}}$，则：

$$\phi'(x) = x(-x^4 + 4x^2 - 1)e^{-\frac{x^2}{2}} \tag{5}$$

$$\phi'(x) = 0 \Leftrightarrow x = 0, \pm\sqrt{2 \pm \sqrt{3}} \tag{6}$$

$\phi'(x)$的图像为：

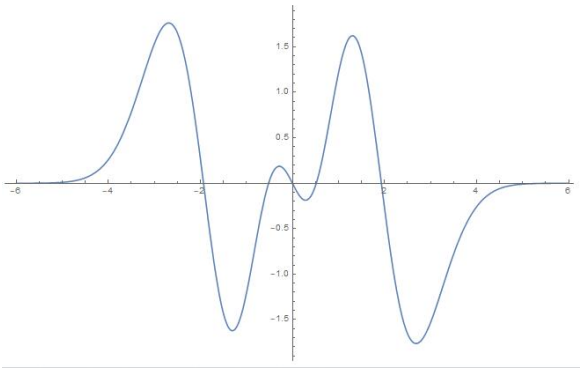


图 1: $\phi'(x)$的图像

可以看出$\phi(x)$的单调递增区间是[-∞, $-\sqrt{2+\sqrt{3}}$]，[$-\sqrt{2-\sqrt{3}}$, 0]和[$\sqrt{2-\sqrt{3}}$, $\sqrt{2+\sqrt{3}}$]；单调递减区间是[$-\sqrt{2+\sqrt{3}}$, $-\sqrt{2-\sqrt{3}}$]，[0, $\sqrt{2-\sqrt{3}}$]和[$\sqrt{2+\sqrt{3}}$, +∞]

对应的函数值为：

$$\phi(0) = 1,\ \phi(\pm\sqrt{2+\sqrt{3}}) \approx 2.30995 \tag{7}$$

所以：

$$c \geq \frac{\phi(\pm\sqrt{2+\sqrt{3}})}{\sqrt{2\pi}} \approx 0.92153762 \tag{8}$$

选取$c = 0.92153762$。

图像如下：

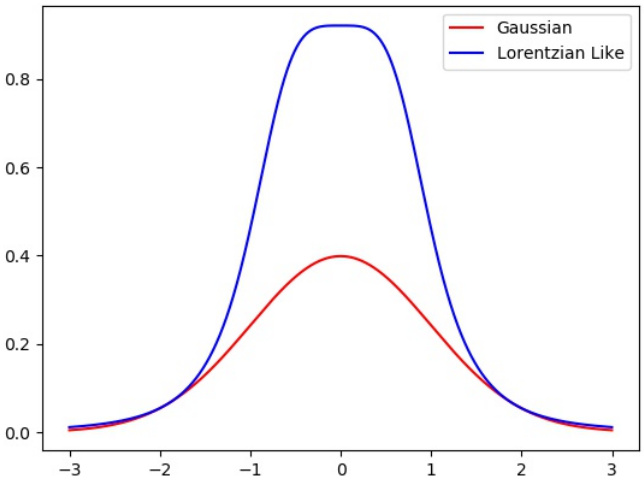



图 2: $p(x)$和$F(x)$的图像

确实满足$p(x) \leq F(x)$。

2. F(x)的累计函数与抽样方法：

   对F(x)积分并归一化得到累计函数：

$$\begin{aligned}\xi &= \frac{\int_{-\infty}^{\xi_x} F(t)dt}{\int_{-\infty}^{\infty} F(t)dt} \\ &= \Psi(\xi_x) \\ &= \frac{1}{4\pi}[2\arctan(1+\sqrt{2}\xi_x) - 2\arctan(1-\sqrt{2}\xi_x) + \ln\frac{\xi_x^2+\sqrt{2}\xi_x+1}{\xi_x^2-\sqrt{2}\xi_x+1}] + \frac{1}{2}\end{aligned} \tag{9}$$

   在[0, 1]上抽取均匀分布的随机数$\xi$，可以得到F(x)的抽样$\xi_x = \Psi^{-1}(\xi)$。

3. 舍选法对Gaussian分布函数进行抽样：

   在[0, F($\xi_x$)]上抽取均匀分布的随机数$\xi_y$，判断$\xi_y \leq p(\xi_x)$是否成立。

   如果成立，则$\xi_x$为p(x)的抽样；否则，重新抽取$\xi_x$。

4. 抽样效率的理论估计：

   由于抽样效率为p(x)曲线下面积和F(x)曲线下面积之比，故抽样效率的理论值为：

$$\eta = \frac{\int_{-\infty}^{\infty} p(x)dx}{\int_{-\infty}^{\infty} F(x)dx} = \frac{\sqrt{2}}{\pi c} \approx 0.488486 \tag{10}$$

## 2.2 算法

按照原理部分的方法进行编程，就可以得到Gaussian分布函数的抽样。

1. 累计函数反函数的计算：

   根据前面的叙述，我们对F(x)进行抽样时，现在[0, 1]上抽取均匀分布的随机数$\xi$，然后通过$\xi_x = \Psi^{-1}(\xi)$计算出的$\xi_x$就是F(x)的抽样。但在本题中，$\Psi^{-1}(\xi)$无法用初等函数表达，甚至都无法求出解析解，因此得用其他的方法进行抽样。

   由于最后我们通过直方图来进行数据的描述，因此我们可以进行离散化的处理，本题的源代码中采取的就是这种方法。

   最后结果中，我们绘制出[-3, 3]上的直方图，且绘制N个矩形。因此把[-3, 3]平均分为N份，每份即为$[-3+\frac{k}{N}\times 6,\ -3+\frac{k+1}{N}\times 6],\ k = 0, 1, 2, ..., N-1$。

   在[0, 1]上抽取均匀分布的随机数$\xi$，从0到N-1循环k，当$\Psi(x_k) \leq \xi \leq \Psi(x_{k+1})$，记$\xi_x = \frac{x_k + x_{k+1}}{2}$，即将此时对应的$\xi_x$由$\Psi^{-1}(\xi)$改成$x_k$和$x_{k+1}$的中点。这样就不用计算$\Psi^{-1}(\xi)$了。

2. 抽样效率的计算：

   每当满足$\xi_y \leq p(\xi_x)$时，记数count加一。循环结束后，返回(count/num)即为抽样效率。

# 3 源文件使用说明

编译并运行“09Acceptance-rejection_Sampling.cpp”，将弹出命令行，要求输入总的抽样个数num。

输入总的抽样个数后，程序运行并将数据输出到文件“num=输入的num.txt”中。同时，命令行上显示本次抽样的抽样效率。

编译并运行“plot.py”即可绘制出直方图。

# 4 计算结果及具体分析

16807的种子值由C语言自带的随机数函数产生，绘制的直方图为[-3, 3]平均分为100份。

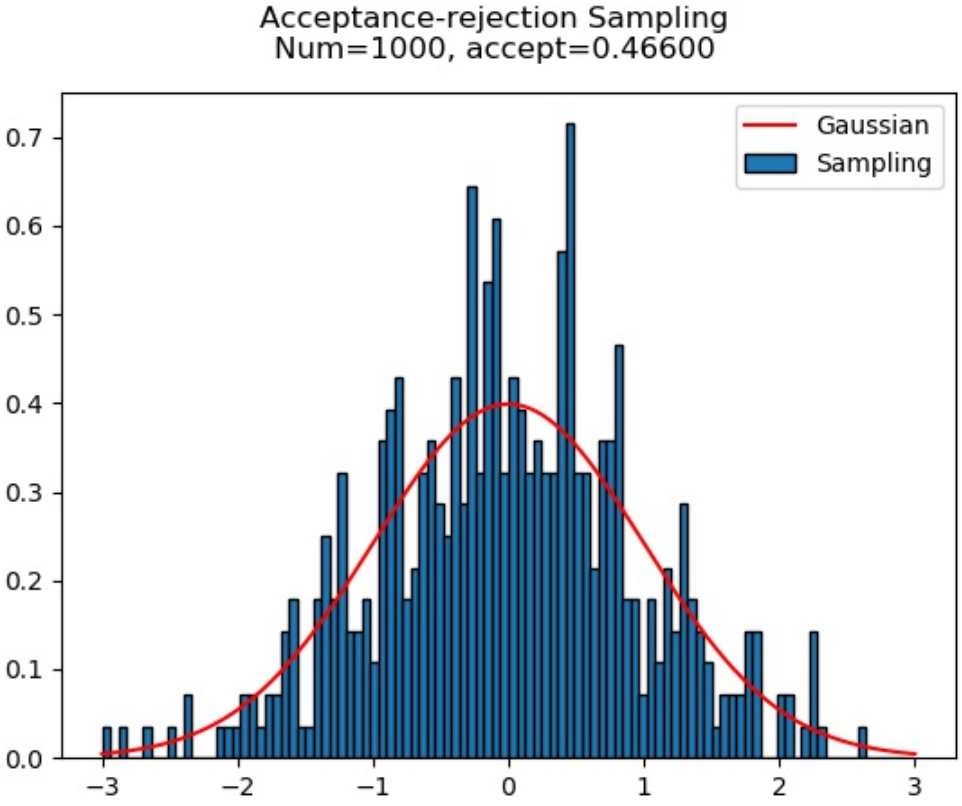



图 3: Num=1000

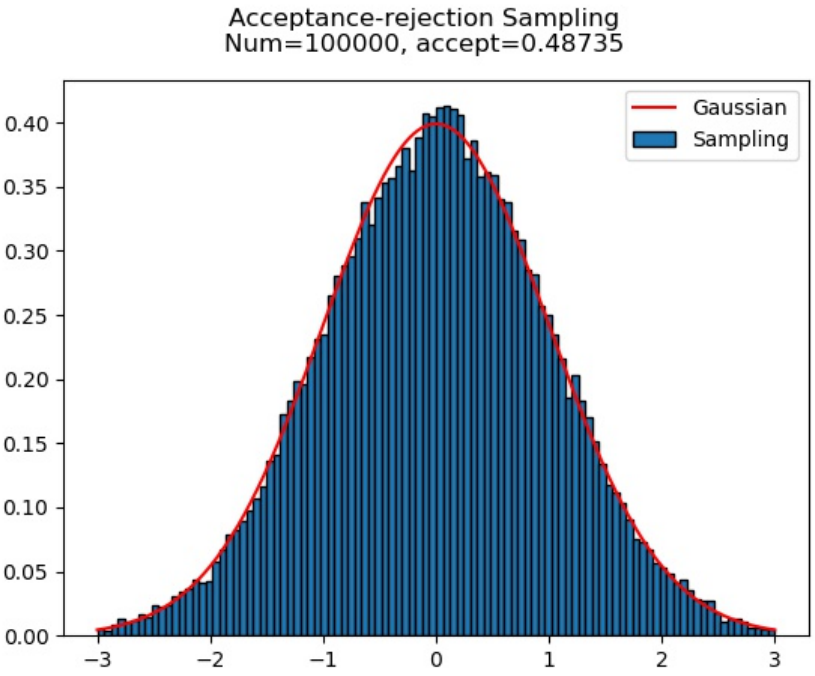



图 4: Num=100000

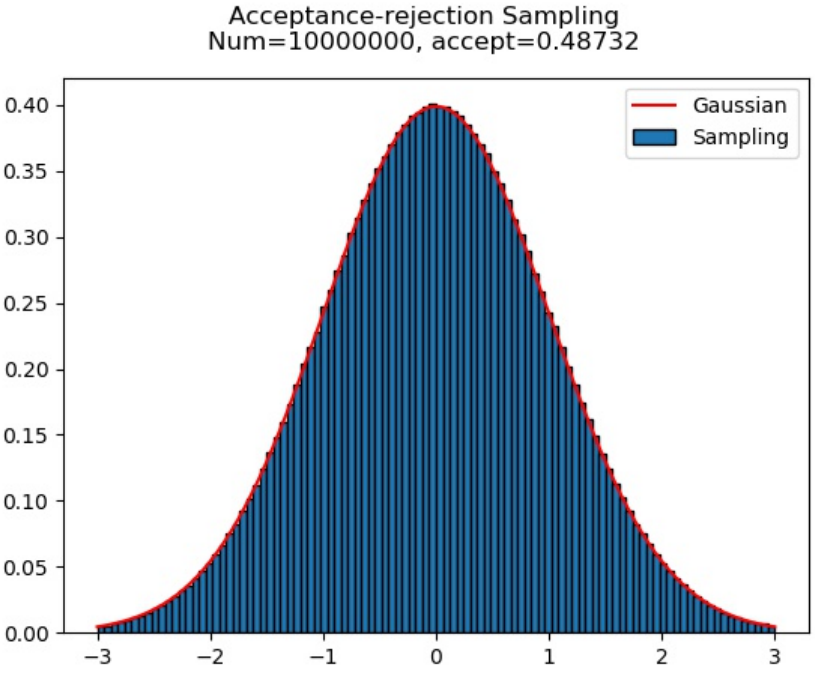



图 5: Num=10000000

# 5 讨论

## 5.1 原理的讨论

本题的原理部分较简单，只需要按照原理部分的舍选法的步骤编程即可。

个人感觉本题出的不好，因为舍选法的核心是将一个难以抽样的函数换成另一个容易抽样的函数进行抽样。但本题中的两个函数都难以进行抽样，因此舍选法在本题相较其他方法也并没有什么优越的地方。对高斯函数抽样不如直接用Box-Muller法。

## 5.2 算法的改进

由于本人的计算机知识并不是很丰富，下面的讨论可能会欠妥，源代码的编写也肯定会有很多需要改进的地方，希望以后学习中能够不断完善。

对于Lorentzian like分布函数的抽样还有其他方法，例如用形式简单的分段函数进行累计函数的拟合，或者进行插值，如下所示：

```
Plot[
 {If[x < -0.8, Exp[2.3 x] + 0.005299373,
   If[x < 0.8, 0.419854 x + 0.5, 0.994700626 - Exp[-2.3 x]]],
  1 / 2 +
   (-2 ArcTan[1 - Sqrt[2] x] + 2 ArcTan[1 + Sqrt[2] x] -
      Log[1 - Sqrt[2] x + x^2] + Log[1 + Sqrt[2] x + x^2]) / (4 π)},
 {x, -3, 3}, PlotLegends → {Interpolation Function, Original Function}]
```

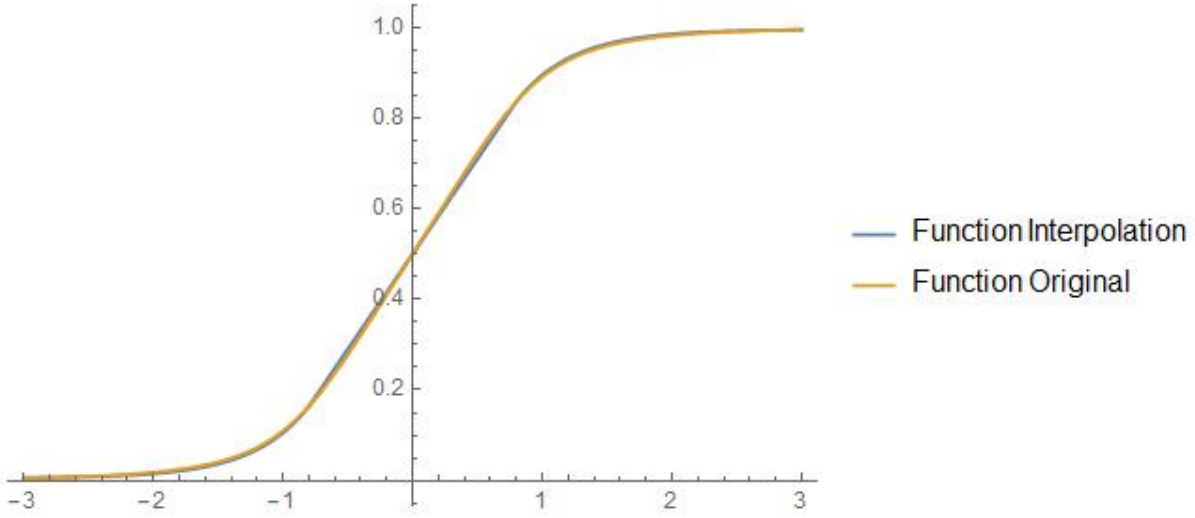



图 6: 三段分段函数

可见在[-3, 3]上用该形式的三段分段函数拟合的十分好。
但得到的抽样并不好，如下所示：

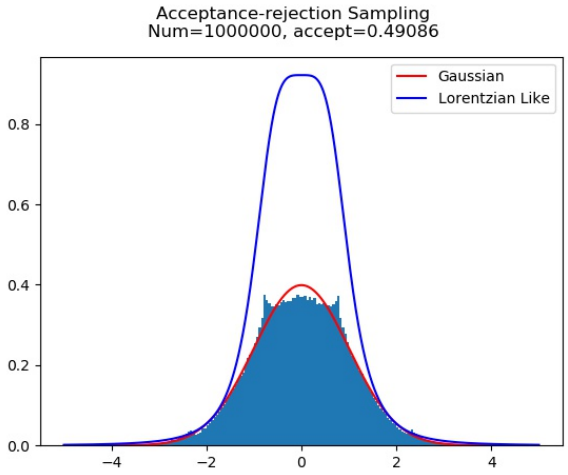



图 7: 三段分段函数

由于0附近是线性拟合，因此理论上直方图是一个平面，因此拟合会有很大差别。
设置五段分段函数，如下所示：
可见在[-3, 3]上用该形式的五段分段函数拟合比三段分段函数拟合更好。

```
Plot[
 {If[x < -0.8, Exp[2.3 x] + 0.005299373,
   If[x < -0.3, 0.402107 x + 0.4858025,
    If[x < 0.3, 0.449432 x + 0.5,
     If[x < 0.8, 0.402107 x + 0.5141975, 0.994700525 - Exp[-2.3 x]]]]],
  1 / 2 +
   (-2 ArcTan[1 - Sqrt[2] x] + 2 ArcTan[1 + Sqrt[2] x] -
     Log[1 - Sqrt[2] x + x^2] + Log[1 + Sqrt[2] x + x^2]) / (4 π)},
 {x, -3, 3}, PlotLegends → {Interpolation Function, Original Function}]
```

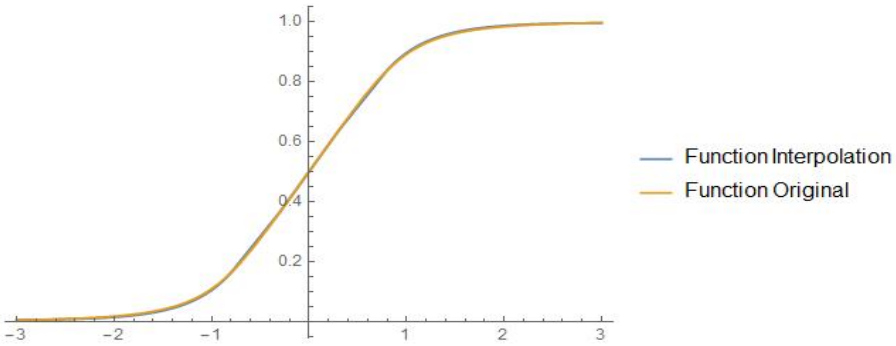



图 8: 五段分段函数

但得到的抽样依然不好，如下所示：

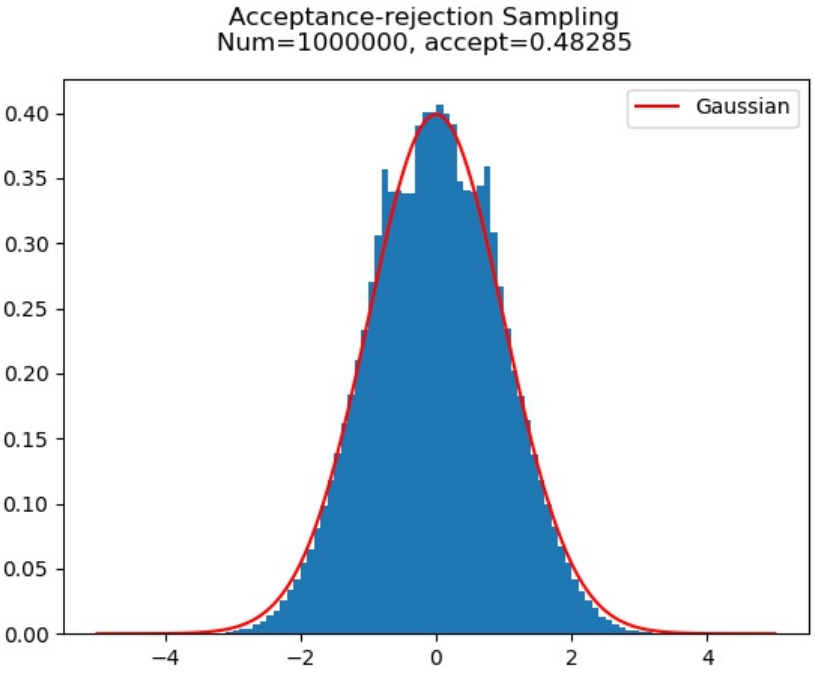



图 9: 五段分段函数

五段分段函数是对三段分段函数0附近的线性拟合函数再进行细分，将刚才的一个线性函数分成三段、三个线性函数。这次拟合的结果比三段分段函数的要好，但仍然有较大差别。

如果能对函数再进行细分，得到结果会更好。

不过显然这个方法十分复杂，不如直接用本题中所用的离散化的方法画直方图。

另外，本题中的离散化方法因为用到了循环，将会使时间复杂度变大，例如10的七次方个点的抽样都需要几分钟了。

## 5.3 结果的讨论

1. 从结果中的图片可以看出，随着抽样点的增加，直方图与曲线拟合的越来越接近，且抽样效率也更加接近理论值。当Num为$10^7$时，模拟的抽样效率为0.48732，而理论值为0.488486，相差非常小。

   这是因为数量少时，统计涨落影响更大，偶然误差更大，将会有较大偏差。

2. 另外，如果不是在整个实轴上进行抽样，而是在某一个给定区间上进行抽样，则可以用Gaussian分布函数抽取Lorentzian like分布函数，因为Guassian分布函数只需要在该区间上在Lorentzian分布函数之上即可。

   同样的，如果是给定区间上，我们的系数c可以再减小。因为从图2中的两个函数图像来看，0附近Lorentzian like分布函数比Gaussian分布函数大很多，抽样效率不高也主要是因为这个原因。如果我们仅在0附近进行抽样，比如[-1, 1]上进行抽样，c不必是之前所选的值，可以取得更小，这样可以显著提高抽样效率。