

随机过程 2021 年秋季学期 Final Project

交叉研 21 班 赵成钢 2021311777

2021 年 12 月

目录

1	一维随机游走	2
1.1	a_1 和 a_2 的条件	2
1.2	停时 T_c 的期望和方差	2
1.2.1	理论分析	2
1.2.2	仿真验证	3
1.3	$p_1 = p_2 = p = 0.5$	6
1.3.1	理论分析	6
1.3.2	仿真验证	7
2	二维随机游走	8
2.1	尝试	8
2.2	Markov 链 (递推)	9
2.3	仿真验证	9
3	Brownian 运动	9
3.1	T_c 和 X_c 的概率分布	9
3.1.1	理论分析	9
3.1.2	仿真验证	11
3.2	最小值和最大值的分布	12
4	代码说明	13

1 一维随机游走

1.1 a_1 和 a_2 的条件

令 Δ_n 为每个时刻移动引起的位置变化, 且 $n \geq 1$, 并有如下 4 种情况:

1. 以 $p_1 p_2$ 的概率同时向右, $\Delta_n = 0$;
2. 以 $(1 - p_1)(1 - p_2)$ 的概率同时向左, $\Delta_n = 0$;
3. 以 $p_1(1 - p_2)$ 的概率, 粒子 1 向右, 粒子 2 向左, $\Delta_n = -2$;
4. 以 $(1 - p_1)p_2$ 的概率, 粒子 1 向左, 粒子 2 向右, $\Delta_n = 2$ 。

则在 n 时刻, 两个粒子因为移动产生的位置差为 $S_n = \sum_{i=1}^n \Delta_i$ 。碰撞时, 有 $S_{T_c} = \sum_{i=1}^{T_c} \Delta_i = -(a_2 - a_1) = 2k$ ($k \in \mathbb{Z}$), 则显然需要满足 $a_2 - a_1$ 为偶数且 $a_1 \leq a_2$ 。

1.2 停时 T_c 的期望和方差

1.2.1 理论分析

首先, 为满足碰撞假设, 应有 $a_2 - a_1$ 为偶数。

根据鞅的构造方式, 构造:

$$\begin{aligned} Z_n &= \sum_{i=1}^n (\Delta_i - E[\Delta_i]) \\ &= S_n - nE[\Delta_1] \\ &= S_n - 2n(1 - 2p) \end{aligned}$$

即 $\{Z_n, n \geq 0\}$ 满足鞅的性质, 且停时满足 $T_c = \inf\{n : Z_n = -(a_2 - a_1) - 2n(1 - 2p)\}$ 。容易验证, 停时定理 1 成立 (同课本例题 3.11), 所以:

$$\begin{aligned} E[Z_{T_c}] &= E[Z_0] = 0 \\ -(a_2 - a_1) - 2E[T_c](1 - 2p) &= 0 \\ E[T_c] &= \frac{a_2 - a_1}{2(2p - 1)} \end{aligned}$$

根据鞅的构造方法，构造（令 $c = 2(1 - 2p)$ ）：

$$\begin{aligned}
M_n &= \sum_{i=1}^n Z_i^2 - E[Z_i^2 | F_{i-1}] \\
&= \sum_{i=1}^n (S_{i-1} + \Delta_i - ic)^2 - E[(S_{i-1} + \Delta_i - ic)^2 | F_{i-1}] \\
&= \sum_{i=1}^n 2S_{i-1}(\Delta_i - ic) + (\Delta_i - ic)^2 - E[2S_{i-1}(\Delta_i - ic) + (\Delta_i - ic)^2 | F_{i-1}] \\
&= S_n^2 + \sum_{i=1}^n -2icS_{i-1} - 2ic\Delta_i + i^2c^2 - E[2S_{i-1}(\Delta_i - ic) + (\Delta_i - ic)^2 | F_{i-1}] \\
&= S_n^2 + \sum_{i=1}^n -2icS_{i-1} - 2ic\Delta_i + i^2c^2 - 2S_{i-1}(1-i)c - E[(\Delta_i - ic)^2] \\
&= S_n^2 + \sum_{i=1}^n -2ic\Delta_i + i^2c^2 - 2cS_{i-1} - E[(\Delta_i - ic)^2] \\
&= S_n^2 + \sum_{i=1}^n -2ic\Delta_i - 2cS_{i-1} - E[\Delta_i^2] + 2ic^2 \\
&= S_n^2 - nE[\Delta_1^2] + (1+n)nc^2 - 2c \sum_{i=1}^n i\Delta_i + S_{i-1} \\
&= S_n^2 - nE[\Delta_1^2] + (1+n)nc^2 - 2ncS_n \\
&= S_n^2 - 4n(p^2 + (1-p)^2) + 4(n^2 + n)(1-2p)^2 - 4n(1-2p)S_n
\end{aligned}$$

同习题 3.21， M_n 满足停时定理 3，故：

$$\begin{aligned}
E[M_{T_c}] &= E[M_0] = 0 \\
(a_1 - a_2)^2 - 4E[T_c](p^2 + (1-p)^2) + 4(E[T_c] + E[T_c^2])(1-2p)^2 - 4E[T_c](1-2p)(a_1 - a_2) &= 0 \\
(a_2 - a_1)^2 - 4\frac{a_2 - a_1}{2(2p-1)}(p^2 + (1-p)^2) + 4(\frac{a_2 - a_1}{2(2p-1)} + E[T_c^2])(1-2p)^2 - 2(a_2 - a_1)^2 &= 0
\end{aligned}$$

解得：

$$\begin{aligned}
E[T_c^2] &= \frac{(a_2 - a_1)(2(a_2 - a_1)p - (a_2 - a_1) - 4p^2 + 4p)}{4(2p-1)^3} \\
\text{Var}[T_c^2] &= E[T_c^2] - (E[T_c])^2 \\
\text{Var}[T_c] &= \frac{(a_2 - a_1)(1-p)p}{(2p-1)^3}
\end{aligned}$$

1.2.2 仿真实验

在实际的模拟中，我们固定 $a_2 - a_1 = 4$ ， p 从 0.51 到 1，每隔 0.01 取一个，进行 10000 次模拟取次数的平均值，同时限定最大游走次数为 4000 步，如果 4000 步后还没有碰到就同样认为

4000 步（如果丢弃这个结果会让模拟的期望变小），虽然这样估计出来的期望也可能变小，但只要最大游走次数越大，那么最后的精度就会越好，所以在实验中我发现只要不丢弃这个结果，返回一个和 4000 在同一个数量级的数就可以得到很好的仿真效果。根据以上设定，如图 1，我们可以画出期望关于概率 p 变化的曲线图（理论和仿真）：

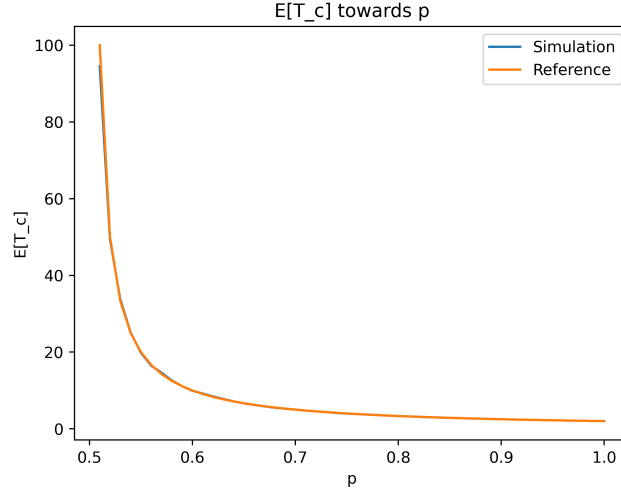


图 1: $E[T_c]$ 关于 p 的曲线（固定 $a_2 - a_1 = 4$ ）

另外，我们也可以固定 $p = 0.8$ ， $a_2 - a_1$ 从 2 变化到 32（不取奇数），进行 10000 次模拟，同时限定最大游走次数为 4000 步。可以得到图 2：

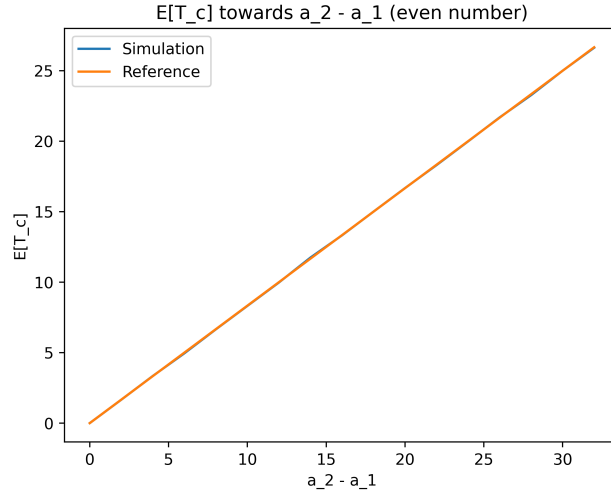


图 2: $E[T_c]$ 关于 $a_2 - a_1$ 的曲线（数据点只有偶数的整数，本图为了直观连接了起来）

对于方差，模拟时需要注意期望直接取样本平均就可以，但方差需要进行无偏估计： $\sigma^2 = \sum_{i=1}^n (X_i - \bar{X})^2$ 。同样的，我们先固定 $a_2 - a_1 = 4$ ， p 从 0.51 到 1，每隔 0.01 取一个，进行 100000

次模拟，限定最大游走次数为 8000。注意到这里的次数取值会比期望更大，是因为方差受误差的影响更大，需要增加模拟次数。根据以上设定，如图 3，我们可以画出方差关于概率 p 变化的曲线图（理论和仿真）：

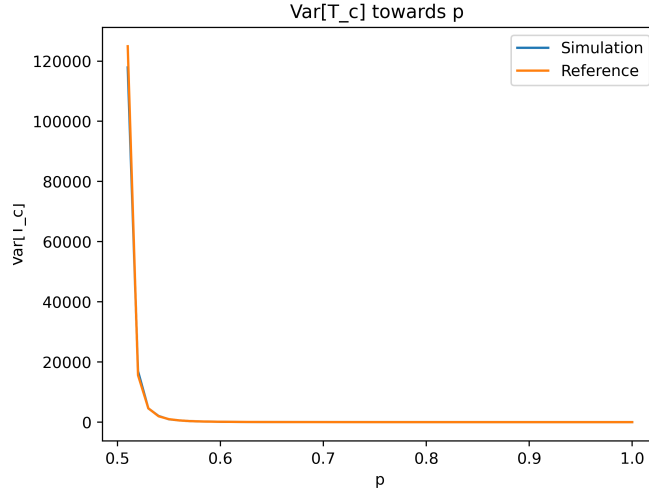


图 3: $Var[T_c]$ 关于 p 的曲线（固定 $a_2 - a_1 = 4$ ）

固定 $p = 0.8$ ， $a_2 - a_1$ 从 2 变化到 32（不取奇数），进行 100000 次模拟，同时限定最大游走次数为 8000 步。可以得到图 4：

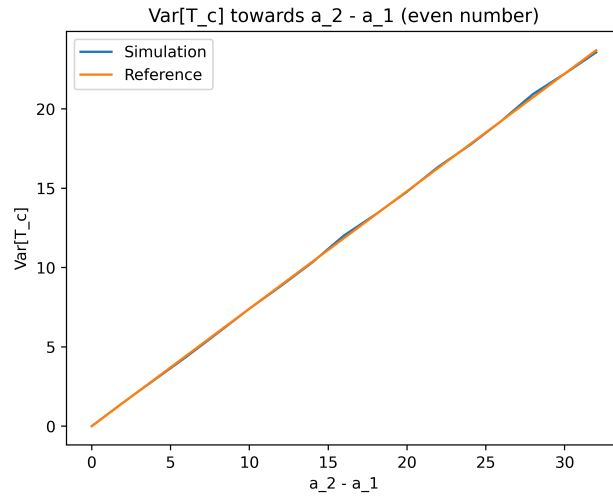


图 4: $E[T_c]$ 关于 $a_2 - a_1$ 的曲线（数据点只有偶数的整数，本图为了直观连接了起来）

可以看到上面 4 条曲线重合得非常好，互相印证了理论和仿真的正确性。

1.3 $p_1 = p_2 = p = 0.5$

1.3.1 理论分析

假设 $E[T_c]$ 存在, 则满足停时定理 1, 根据 1.2 中的结论, 有

$$E[Z_{T_c}] = -(a_2 - a_1) = E[Z_0] = 0$$

这显然不成立, 所以 $E[T_c]$ 不存在。

令 $s = \frac{a_2 - a_1}{2}$, 设 n 时刻已经进行了 x_n 次 $\Delta = -2$ (以 $p = \frac{1}{4}$ 的概率) 的运动, y_n 次 $\Delta = 2$ (以 $p = \frac{1}{4}$ 的概率) 的运动, 则在任意时刻 n (停时 T_c 之前), 需要满足 $s + y_n > x_n$, 在停时满足 $s + y_n = x_n$ 。

假设停时 $T_c = n$, 其中有 k 步满足 $x_{T_c} = k$ (发生了 k 次相向运动, $k - s$ 次反向运动, 且 $2k - s \leq n$), 这时有:

$$\begin{aligned} P(T_c = n, x_{T_c} = k) &= \frac{1}{2^{n-(2k-s)} 4^{2k-s}} C_{n-1}^{n-(2k-s)} \text{Count}(k) \\ &= \frac{1}{2^{n+2k-s}} C_{n-1}^{n-(2k-s)} \text{Count}(k) \end{aligned}$$

上述 $C_{n-1}^{n-(2k-s)}$ 的含义是除了最后一步 (因为要保证是停时), 剩下的步数中选择 $n - (2k - s)$ 步是两个粒子同时静止。

下面来求解 $\text{Count}(k)$, 将问题转换为二维平面 (x, y) 的游走问题, 起始点为 $(0, 0)$, 终点为 $(k, k - s)$, 并且保证除了最后一步要满足 $s + y > x$ 即 $y \geq x - s + 1$ 。该问题等价于从 $(0, 0)$ 到 $(k - 1, k - s)$ 且不穿过 $y = x - s + 1$ 的 (x, y) 非降游走 (每个时刻有一个加一) 方案数。

所有方案 (包括穿过) 的数量为:

$$\text{Count}_1(k) = C_{2k-s-1}^{k-1}$$

下面考虑直线 $y = x - s$, $(0, 0)$ 关于该线的镜面对称点为 $(s, -s)$, 从该点到 $(k - 1, k - s)$ 一定会穿过 $y = x - s$, 从第一个穿过的点对前面的路径做镜面对称, 则一一对应了一条从 $(0, 0)$ 到 $(k - 1, k - s)$ 且穿过 $y = x - s + 1$ 的方案, 那么穿过的数量为:

$$\text{Count}_2(k) = C_{2k-s-1}^k$$

所以 (其中 $n \geq s = \frac{a_2 - a_1}{2} > 0$):

$$\begin{aligned}
\text{Count}(k) &= \text{Count}_1(k) - \text{Count}_2(k) \\
&= C_{2k-s-1}^{k-1} - C_{2k-s-1}^k \\
P(T_c = n, x_{T_c} = k) &= \frac{1}{2^{n+2k-s}} C_{n-1}^{n-(2k-s)} (C_{2k-s-1}^{k-1} - C_{2k-s-1}^k) \\
P(T_c = n) &= \sum_{k=s}^{\lfloor \frac{n+s}{2} \rfloor} \frac{1}{2^{n+2k-s}} C_{n-1}^{n-(2k-s)} (C_{2k-s-1}^{k-1} - C_{2k-s-1}^k) \\
&= \sum_{k=s}^{\lfloor \frac{n+s}{2} \rfloor} \frac{s}{k 2^{n+2k-s}} C_{n-1}^{n-(2k-s)} C_{2k-s-1}^{k-1} \\
&= \frac{s}{n} \sum_{k=s}^{\lfloor \frac{n+s}{2} \rfloor} \frac{1}{2^{n+2k-s}} C_n^{2k-s} C_{2k-s}^k \\
&= \frac{s}{n 2^{n-s}} \sum_{k=s}^{\lfloor \frac{n+s}{2} \rfloor} \frac{1}{4^k} C_n^{2k-s} C_{2k-s}^k
\end{aligned}$$

进一步的 $\sum_{k=s}^{\lfloor \frac{n+s}{2} \rfloor} \frac{1}{4^k} C_n^{2k-s} C_{2k-s}^k$ 等价于 $f(x) = (x^2 + x + \frac{1}{4})^n$ 的 x_{n-s} 前的系数, 等价于 $f(x) = (x + \frac{1}{2})^{2n}$ 的系数, 即 $C_{2n}^{n-s} \frac{1}{2^{n+s}}$, 故原式为:

$$\begin{aligned}
P(T_c = n) &= \frac{s}{n 2^{n-s}} C_{2n}^{n-s} \frac{1}{2^{n+s}} \\
&= C_{2n}^{n-s} \frac{s}{n 2^{2n}}
\end{aligned}$$

把 $s = \frac{a_2 - a_1}{2}$ 代入, 考虑奇偶的情况, 综上所述:

$$P(T_c = n) = \begin{cases} 0, & a_2 - a_1 \text{ is odd} \\ C_{2n}^{n - \frac{a_2 - a_1}{2}} \frac{\frac{a_2 - a_1}{2}}{n 2^{2n+1}}, & a_2 - a_1 \text{ is even} \end{cases}$$

(另一种简单的考虑方法, 看到这个最后的形式想出来的, 绕了好大一圈路) 两个粒子一个时间内同时运动等价于一个运动两次, 同理用上面路径反射的方法可以以较为简单的形式直接得到同样的答案, 这里的本质在于上文提到的 $f(x) = (x + \frac{1}{2})^{2n} = (x^2 + x + \frac{1}{4})^n$ 多项式, 两个做法的区别在于如果同时考虑两个粒子运动组合数和概率会以 $x^2 + x + \frac{1}{4}$ 进行展开, 而如果只考虑一个粒子运动 $2n$ 次, 那么概率会以 $x + \frac{1}{2}$ 的形式存在, 两者方法其实本质都是一样的, 体现到形式上就是要不要把那个指数上的因子 2 提前展开到里面。

1.3.2 仿真验证

令游走次数上限为 100 步, 模拟 1000000 次, 碰撞则返回次数, 没有在上限内碰撞则计入总次数, 得到仿真和理论的曲线对比如图 5 所示:

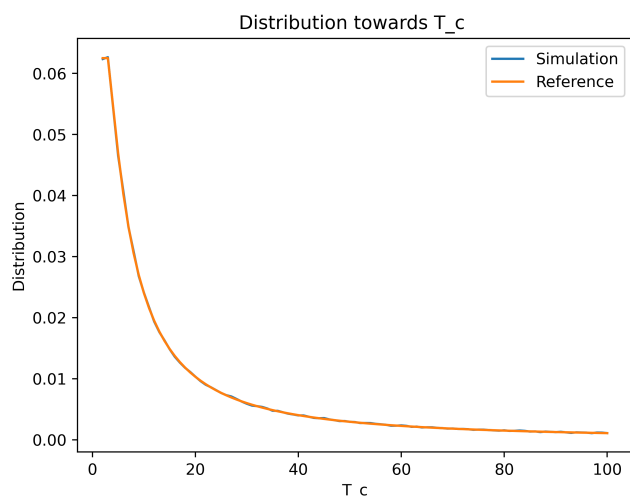


图 5: T_c 的分布

可以看到理论和仿真的结果高度一致。

2 二维随机游走

首先，同上一问我们根据对称性可以假设 $a_2 \geq a_1$ 并且 $b_2 \geq b_1$ ，因为两个粒子向四周游走的概率都是 $\frac{1}{4}$ ，该问题等价于在 $2n$ 步内单个粒子从 $(0,0)$ 首次运动到 $(a_2 - a_1, b_2 - b_1)$ 的时间的分布。

2.1 尝试

本人尝试了通过生成函数和类似第一问计数方式均没有得到准确的解析解；另一个在网络上查找过程中找到的很有意思点是，我们可以把单个点的坐标重新映射为 $(x - y, x + y)$ 这样可以证明两个维度都是独立的运动，但是问题在于无法保证首次两个维度同时碰撞（可以至少是一个维度的近似解）；同时，我们也可以考虑曼哈顿距离等方式，但根本问题也在于虽然两个距离折叠到了一维中，但实际上的运动中，曼哈顿距离的变化在粒子运动到 $(a_2 - a_1, b_2 - b_1)$ 所在的直线上（ $x = a_2 - a_1$ 和 $y = b_2 - b_1$ ）时，因为绝对值的性质，距离的分布会发生变化，这是很少部分的情况，也说明了二维的确可以用一维来近似，但也没有得到确切的解析解；我也尝试过在 MathOverflow 上提出这个问题¹，期待数学学者们给出解答。

¹<https://mathoverflow.net/questions/411878/the-distribution-of-collision-stopping-time-in-2d-random-walk>

2.2 Markov 链 (递推)

对于一个粒子的随机游走，有转移矩阵：

$$P(X_{n+1} = \vec{x}_{n+1} | X_n = \vec{x}_n) = \begin{cases} \frac{1}{4}P(X_n = \vec{x}_n), & \|\vec{x}_{n+1} - \vec{x}_n\| = 1 \text{ and } \vec{x}_n \neq (a_2 - a_1, b_2 - b_1) \\ 0, & \text{else} \end{cases}$$

$$P(X_{n+1} = \vec{x}_{n+1}) = \sum_{\vec{x}_n} P(X_{n+1} = \vec{x}_{n+1} | X_n = \vec{x}_n)$$

最后，根据之前假设有 $P(T_c = n) = P(X_{2n} = (a_2 - a_1, b_2 - b_1))$ 。

2.3 仿真验证

限定最大游走的步数为 100 步，对于理论的递推式，我们每一步只存能走到的位置，这样根据上面的递推式，可以准确算出理论值。对于仿真模拟，设置 $(a_2 - a_1, b_2 - b_1) = (2, 2)$ ，进行了 1000000 次模拟，得到如图 6 的结果，可以看到理论和仿真结果高度统一。

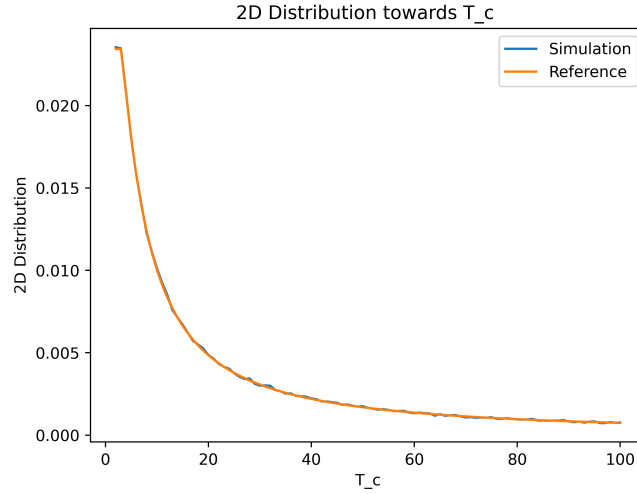


图 6: T_c 的分布

3 Brownian 运动

3.1 T_c 和 X_c 的概率分布

3.1.1 理论分析

同离散的情况，在概率分布意义上，两个粒子在时间 t 内发生的位移运动的差等价于一个粒子在 t 时间内的 Brownian 运动：

$$B_1(t) - B_2(t) \stackrel{d}{=} B(2t)$$

于是问题等价于 $B(2t)$ 首次击中 $a_2 - a_1$ 的时间 T_c 的分布。根据 Brownian 的反射性，有 CDF：

$$\begin{aligned} P(T_c \leq t) &= 2P(B(2t) \geq a_2 - a_1) \\ &= 2(1 - F(\frac{a_2 - a_1}{\sqrt{2t}})) \end{aligned}$$

其中 $F(x)$ 是标准正态分布的 CDF。

对于 X_c 的概率密度分布，假设第一个粒子的运动为 $B_1(t) + a_1$ ，第二个粒子的运动为 $B_2(t) + a_2$ ，构造运动 $\mathbf{X}(t) = (B_1(t) - B_2(t), B_1(t) + B_2(t))$ ，不难验证 $\text{Cov}(B_1(t) - B_2(t), B_1(t) + B_2(t)) = 1 - 1 = 0$ ，即 $\mathbf{X}(t)$ 为每个维度相互独立的二维布朗运动。那么问题等价于求 $B_1(t) - B_2(t)$ 首次击中 $a_2 - a_1$ 时， $B_1(T_c) + a_1$ 的取值 X_c 的分布，也就是求第二维 $B_1(T_c) + B_2(T_c)$ 的分布 $G(x)$ ，后导出 $B_1(T_c) + a_1$ 的分布为 $H(x) = G(2x - (a_1 + a_2))$ 。

不难看出 $X_1(t) = B_1(t) - B_2(t)$ 和 $X_2(t) = B_1(t) + B_2(t)$ 都等价于 $B(2t)$ ，令 $F(t) = P(T_c \leq t)$ 为停时 T_c 的 CDF，则 $Y(a_2 - a_1) = X_2(T_c)$ 的特征函数满足：

$$\begin{aligned} \phi(u) &= E[\exp(juY(x))] \\ &= \int_0^\infty E[\exp(juX_2(t))]dF(t) \end{aligned}$$

而 $X_2(t) \stackrel{d}{=} B(2t)$ ，所以：

$$\begin{aligned} E[\exp(juX_2(t))] &= \exp(-u^2t) \\ \phi(u) &= \int_0^\infty \exp(-u^2t)dF(t) \\ &= E[\exp(-u^2T_c)] \end{aligned}$$

停时 T_c 的 CDF 满足 $F(t) = 2P(B(2t) \geq a_2 - a_1)$ ，易得 PDF 为 $f(t) = \frac{(a_2 - a_1) \exp(\frac{-(a_2 - a_1)^2}{4t})}{2\sqrt{\pi t^3}}$ ，再求解 Laplace 变换同时代入 $s = u^2$ ：

$$\begin{aligned} E[\exp(-sT_c)] &= \int_0^\infty \exp(-st)f(t)dt \\ &= \int_0^\infty \exp(-st) \frac{(a_2 - a_1) \exp(\frac{-(a_2 - a_1)^2}{4t})}{4\sqrt{\pi t^3}} dt \\ &= \exp(-(a_2 - a_1)\sqrt{s}) \\ \phi(u) &= \exp(-(a_2 - a_1)|u|) \end{aligned}$$

对应参数为 $a_2 - a_1$ 的 Cauchy 分布，并导出 $B_1(T_c) + a_1$ 的分布 $H(x)$ ：

$$\begin{aligned} G(x) &= \frac{1}{\pi} \arctan(\frac{x}{a_2 - a_1}) + \frac{1}{2} \\ H(x) &= \frac{1}{\pi} \arctan(\frac{2x - (a_1 + a_2)}{a_2 - a_1}) + \frac{1}{2} \end{aligned}$$

3.1.2 仿真验证

对于 Brownian 运动的模拟，我没有使用一维随机游走来逼近，因为这样有时间要足够小才能近似成中心极限定理而导致的误差，根据 Brownian 运动的独立增量性，我们选择一个足够小的时间 Δt ，然后采样：

$$B(0), B(\Delta t), B(2\Delta t), \dots$$

同时，让计算机模拟满足标准正态分布的独立同分布变量序列 N_0, N_1, \dots ，则可以得到在采样时间点准确的 Brownian 运动：

$$B((k+1)\Delta t) = B(k\Delta t) + \sqrt{\Delta t}N_k$$

那么唯一的误差就在于所谓的碰撞，在模拟中碰撞我们定义为 $|x_1 - x_2| \leq \sqrt{\Delta t}$ ，因为相当于我们把行走的方差看成每个粒子的直径，把两个粒子看成两个圆，当圆心距离小于等于直径的时候，我们认为发生了碰撞，同时也保证了不会有互相穿过去的情况。

我们设置 $a_1 = 0, a_2 = 1$ 并且有 $\Delta t = 0.0005$ （越小误差越小），最大游走次数为 20000 次，模拟 10000 次，得到 T_c 的分布如图 7 所示。

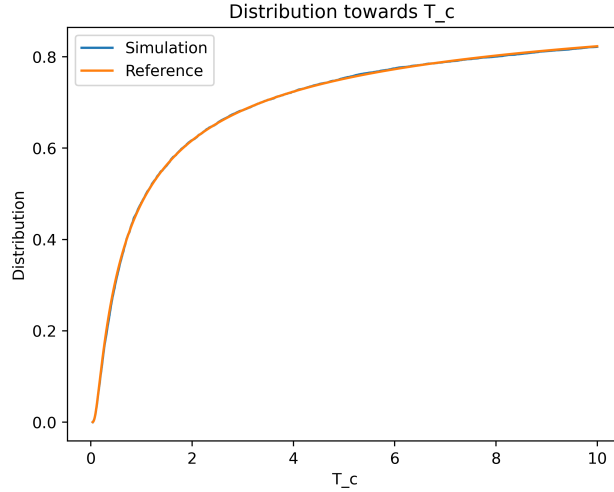


图 7: T_c 的分布 (CDF)

对于 X_c 的分布，需要注意因为我们无法判断在有限次数没有碰撞的情况最后碰撞时是停在负无穷还是正无穷的方向，所以我们需要把最大游走次数设定到非常大，同时减少步长来减少误差，也需要注意这里分布的基数是碰撞的情况，不会发生碰撞的情况不应该进行纳入总数，设置 $a_1 = 0, a_2 = 1$ 并且有 $\Delta t = 0.01$ ，最大游走次数为 200000 次，模拟 10000 次，得到 X_c 的 CDF 如图 8 所示。

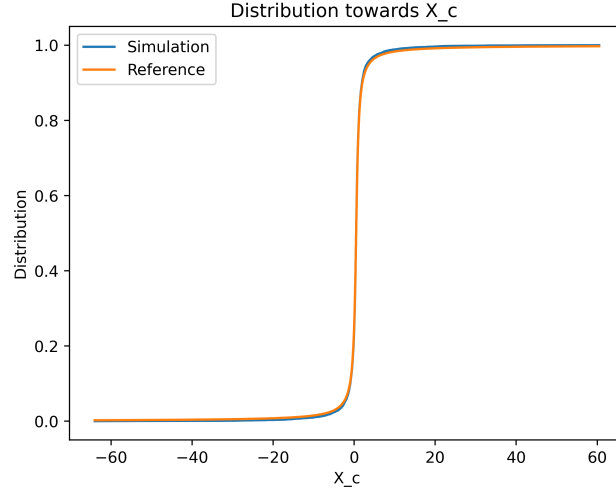


图 8: X_c 的分布 (CDF)

3.2 最小值和最大值的分布

该小问为开放式问题，我尝试了很久没有做出来，可以得到比较显然的结论是最小值一定是开始坐标更小的粒子取得，而最大值是坐标更大的取得。我们固定 $a_1 = 0, a_2 = 1$ ，设置最大游走步数为 200 步， $\Delta t = 0.01$ ，模拟 20000 次，可以先得到最小值 Min_c 的分布曲线，如图 9：

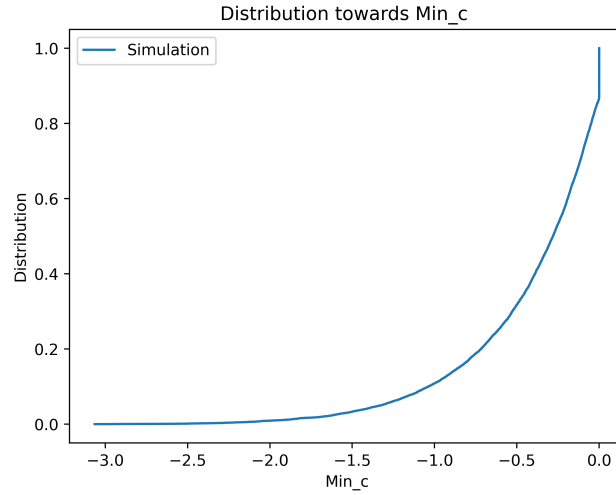


图 9: Min_c 的分布 (CDF)

可以观察到曲线的形状比较陡峭，比较像 \arctan 等函数，再加上第一问中 Cauchy 分布的联想，我们构造函数 $K(x, a, b, c, d, e, f) = a + f \arctan(bx^3 + cx^2 + dx + e)$ ，然后调用 SciPy 的优化库中 Trust Region Reflective 算法搜索参数 a, b, c, d, e, f ，得到最优解 $a = 0.32700462, b =$

$3.80120498, c = -14.88451399, d = 22.20498888, e = -11.59463005, f = 0.43500348$, 同时进行和最小值同样设置的模拟, 得到两条曲线如图 10 所示:

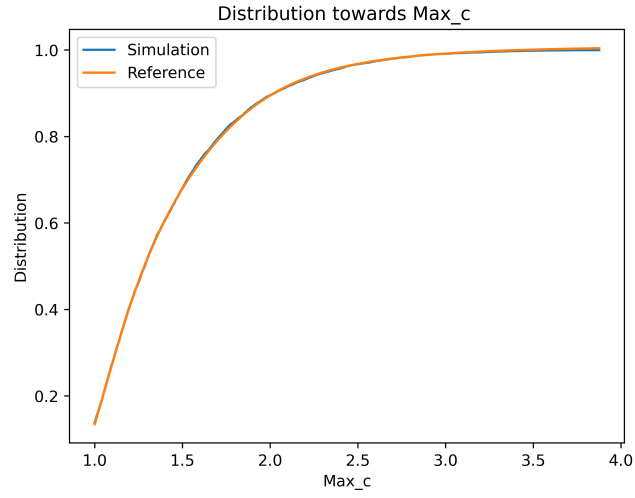


图 10: Max_c 的分布 (CDF) 和拟合曲线 (Reference)

4 代码说明

仿真代码使用 Python 实现, 具体见 main.py, 需要使用 Python 3 作为环境并安装 matplotlib、numpy 和 scipy 包。