An Alternative Method for Characterization and Comparison of Plant Root Shapes

A thesis submitted to the
College of Graduate and Postdoctoral Studies
in partial pulfillment of the requirements
for the degree of Master of Science
in the Department of School of Environment and Sustainability
University of Saskatchewan
Saskatoon

By
Yujie Pei

Contents

Existed Morphological Descriptors for Root Systems

An Alternatively Mathematical Method for Shape Description

2.1 Kac's Idea

2.2 Heat Content in Annulus

Since the heat equation, also known as the diffusion equation, defined in an annulus possesses explicit solutions, the analytical expressions of heat content, derived by the integration over the whole domain, are then accessible. In this section, the preliminary step is to solve the initial-boundary value problem (IBVP) defined in the annulus as shown in Figure 2.1. As the solution of IBVP and its related quantities are in the form of the infinite series, the numerical method will be used to approximate them to validate the research methodology in the next chapter.

2.2.1 Analytical Results

The solution to the 2-dimensional heat equation [16] Eq. 2.1 defined in the polar coordinate system, $u(r, \theta, t)$, describes the heat distribution or temperature varying in time and positions in the domain Ω shown in Figure 2.1, where r is radial coordinate, θ is the angular coordinate, t is the time, D is the diffusion coefficient, and subscripts denote derivatives with respect to the indicated independent variable.

$$u_t = D\left(u_{rr} + \frac{1}{r}u_r + \frac{1}{r^2}u_{\theta\theta}\right) \qquad \text{in } \Omega \tag{2.1}$$

$$u = 0 \qquad \text{on } \partial\Omega_1 \tag{2.2}$$

$$ru_r = 0 \qquad \text{on } \partial\Omega_2 \tag{2.3}$$

$$u(r, \theta, 0) = \frac{1}{|\Omega|} \qquad \text{in } \Omega \tag{2.4}$$

Note, $|\Omega|$ is the area of the dark annulus in Fig. 2.1 and $\partial\Omega_1$ and $\partial\Omega_2$ are disjoint subsets of the boundary of Ω.

Consider a memoryless particle, whose movement afterwards is independent of how it gets the previous positions. From the microscopic and probabilistic perspective, $u(r, \theta, t)$ gives the probability of the particles at (r, θ, t).

Solving Heat Equation

Generally, before solving the heat equation, it is convenient and efficient to generate a group of dimensionless variables by dimensional analysis. The benefit of dimensional analysis is that many physical parameters can be combined into a smaller number of unitless variables, which do not depend on the unit of the measurements and can also describe the phenomenon or system of interest [7].

Let $\mu = b/a$ be the dimensionless radius ratio, $\tau = \frac{Dt}{a^2}$ be the dimensionless time, and $\hat{r} = \frac{r}{a}$ be the unitless radius. Substitute these dimensionless variables into Eq. 2.1 and rewrite it as

$$u_\tau = (u_{\hat{r}\hat{r}} + \frac{1}{\hat{r}}u_{\hat{r}} + \frac{1}{\hat{r}^2}u_{\theta\theta}) \tag{2.5}$$

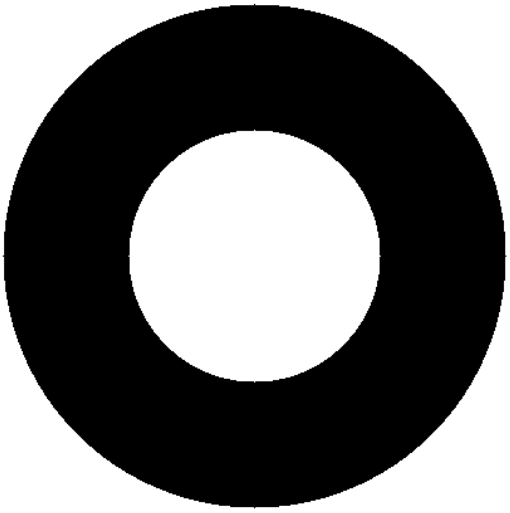


Figure 2.1: The dark annulular region, Ω, is treated as a homogeneous and isotropic medium with relecting (Neumann) inner boundary $\partial\Omega_1$ at radius a and absorbing (Dirichlet) outer boundary $\partial\Omega_2$ at radius b.

With the uniform initial condition,

$$u(\hat{r},\theta,0) = \frac{1}{|\Omega|} \tag{2.6}$$

With the homogenous boundary conditions

$$u(1,\theta,\tau) = 0 \tag{2.7}$$

$$u'(\mu,\theta,\tau) = 0 \tag{2.8}$$

After implementing the separation of variables method [16], the solutions of Eq. 2.5 with conditions Eq. 2.6, Eq. 2.7, and Eq. 2.8 are

$$u(\hat{r},\theta,\tau) = \sum_{n=1}^{\infty} c_{\tilde{0},n}\left\{ J_0(\sqrt{\lambda_{0,n}})Y_0(\sqrt{\lambda_{0,n}}\hat{r}) - Y_0(\sqrt{\lambda_{0,n}})J_0(\sqrt{\lambda_{0,n}}\hat{r}) \right\} e^{-\lambda_{0,n}\tau} \tag{2.9}$$

where

$$c_{\tilde{0},n} = \frac{1}{(\mu^2-1)} \frac{1}{\left[\frac{J_0(\sqrt{\lambda_{0,n}})}{J_0'(\mu\sqrt{\lambda_{0,n}})}\right]^2 - 1} \tag{2.10}$$

Eigenvalues $\lambda_{0,n}$ ($n \in \mathbb{N}_+$) appeared in Eq. 2.9 and Eq. 2.10 is the nth positive root of the eigenfunction Eq. 2.11, which is a cross-product of the Bessel function [75]

$$F_0(\lambda) = J_0(\sqrt{\lambda})Y_0'(\sqrt{\lambda}\mu) - J_0'(\sqrt{\lambda}\mu)Y_0(\sqrt{\lambda}) \tag{2.11}$$

Heat Content (Survival Probability)

The amount of heat contained in Ω at the moment $\tau > 0$ defined as heat content $Q_\Omega(\tau)$, which is an alternative terminology of survival probability $S(\tau)$ in some mathematical literatures [10] [70] [**?**]. $S(\tau)$ is proportional to $Q_\Omega(\tau)$ [43], which gives the probability of the particles remain diffusing in the domain Ω at time $\tau > 0$ [1]. Survival probability can be expressed by

$$\begin{aligned} S(\tau) &= \int_0^{2\pi} d\theta \int_1^{\mu} \hat{r} d\hat{r} u(\hat{r}, \theta, \tau) \\ &= \sum_{n=1}^{\infty} \frac{4}{\mu^2 - 1} \frac{1}{\lambda_{0,n} \left\{ \left[\frac{J_0(\sqrt{\lambda_{0,n}})}{J_0'(\mu\sqrt{\lambda_{0,n}})} \right]^2 - 1 \right\}} e^{-\lambda_{0,n}\tau} \end{aligned} \tag{2.12}$$

Eq. 2.12 reveals some basic properties of $S(\tau)$. Firstly, when $\tau = 0$, the survival probability is 1 since all the particles are just generated over the whole domain and not be absorbed by Ω_1. Secondly, $S(\tau)$ is a convergent series with multiexponential decay. Thirdly, $S(\tau)$ interconnects the overall geometric characteristics of Ω. For example, the decay rate of $S(\tau)$ in a short time heavily depends on the geometrical features of Ω_1, as only the particles inserted close to Ω_1 have the high probabilities of being absorbed. Finally, as the lone-time limit, $S(\tau)$ is represented by the lowest eigenvalue $\lambda_{0,1}$.

Mean First-Passage Time

The first passage phenomena play a fundamental role in stochastic processes triggered by a first-passage event [71]. In this thesis, we only focus on the stochastic evolution of $u(\hat{r}, \theta, \tau)$ until the first-passage time, at which a heat particle reaches any sites of target boundary Ω_1 for the first time. Similarly, another essential first-passage-related quantity is the first-passage probability, which is a probability of the diffusing heat particles hitting a specified site or a set of sites at a specified time for the first time [62]. All the first-passage characteristics can be expressed in terms of the first-passage probability. For example, the survival probability of heat particles at time τ calculated in the last subsection is

$$f(\tau) = -\frac{\partial S(\tau)}{\partial \tau} \tag{2.13}$$

where $f(\tau)$ is the first-passage probability to the target boundary Ω_1 at time τ regardless of particles' stop positions. By the definition, the nth moment of the exit time [62] is

$$\begin{aligned} \langle \tau^n \rangle &= \int_0^{\infty} \tau^n f(\tau) d\tau \\ &= -\int_0^{\infty} \tau^n \frac{\partial S(\tau)}{\partial \tau} d\tau \\ &= -\tau^n S(\tau)|_0^{\infty} + n \int_0^{\infty} \tau^{n-1} S(\tau) d\tau \end{aligned} \tag{2.14}$$

Substitue $n = 1$ in Eq. 2.14, the mean-first passage time $\langle\tau\rangle$, also called the average first-passage time, of heat particles implies an overall property of the system and can be expressed as

$$\begin{aligned}\langle\tau\rangle &= \int_0^\infty \tau dS \\ &= \sum_{n=1}^{\infty} \frac{4}{\mu^2 - 1} \frac{1}{\lambda_{0,n}^2 \left\{ \left[\frac{J_0(\sqrt{\lambda_{0,n}})}{J_0'(\mu\sqrt{\lambda_{0,n}})} \right]^2 - 1 \right\}}\end{aligned} \tag{2.15}$$

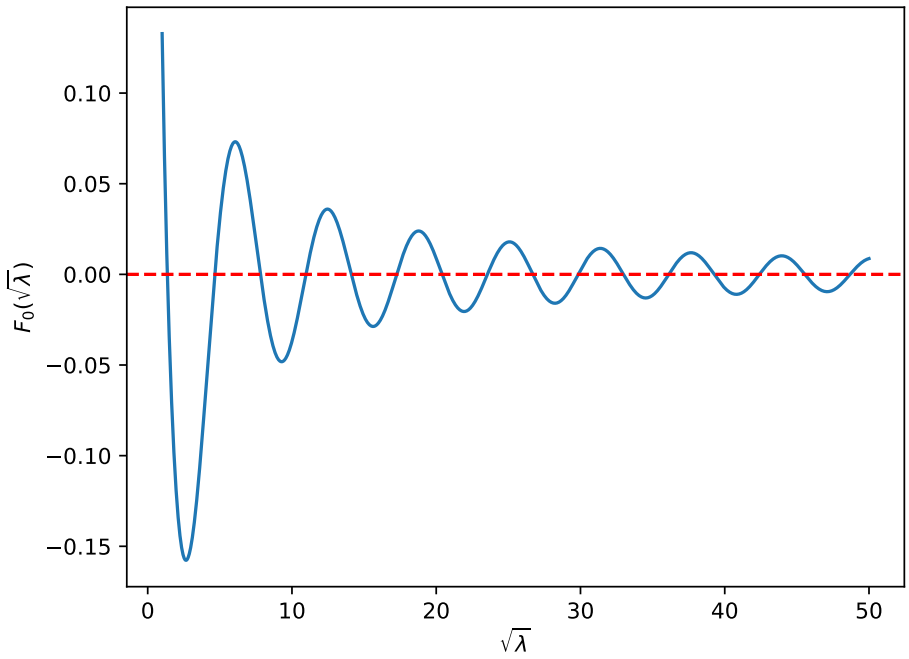


Figure 2.2: It is straightforward to evaluate the cross-product of Bessel functions Eq. 2.11 by SciPy library [73].

2.2.2 Numerical Approximation

The problem solved in the subsection 2.2.1 is defined in the continuous-space and continuous-time, and all the analytical results are in the form of the infinite series involving continuous variables, which will be approximated numerically in this subsection. Moreover, for simplicity, a specific kind of annulus, with the radius ratio $\mu = 2$, is considered.

The analytical $u(\hat{r}, \theta, \tau)$, $S(\tau)$, and $\langle\tau\rangle$ are in connection with the positive integers $\lambda_{0,n}$, so the preliminary step of the numerical evaluation is to calculate the monotonically increasing eigenvalues in Eq. 2.11. The Figure 2.2 reveals the monotonicity and periodicity of the $\lambda_{0,n}$. More precisely, the nth positive zero $\lambda_{0,n}$, as $n \to \infty$, can be bracketed in an interval $((n-1)\pi, (n+1)\pi)$ [21]. Therefore, the bisection method [73], a well-known and most reliable root finding method, is used to close in on the $\lambda_{0,n}$ by successively halving the interval until it becomes sufficiently small.

After the numerical estimation of the $n-$th positive root $\lambda_{0,n}$ of Eq. 2.11, the next step is to approximate $u(\hat{r}, \theta, \tau)$ with the first 1000 eigenvalues and the discretized radius $\hat{r}$ and τ. As illustrated in Figure 2.3, the diffusion problem experiences a rapid change in the very beginning, but then the evolution of u becomes slower and slower. Finally, the initial shape of u can not be recognized anymore.

Before the numerical approximation of the convergent non-alternating general Dirichlet series [35] $S(\tau)$, $S(0)$ is evaluated firstly, and it can be represented as

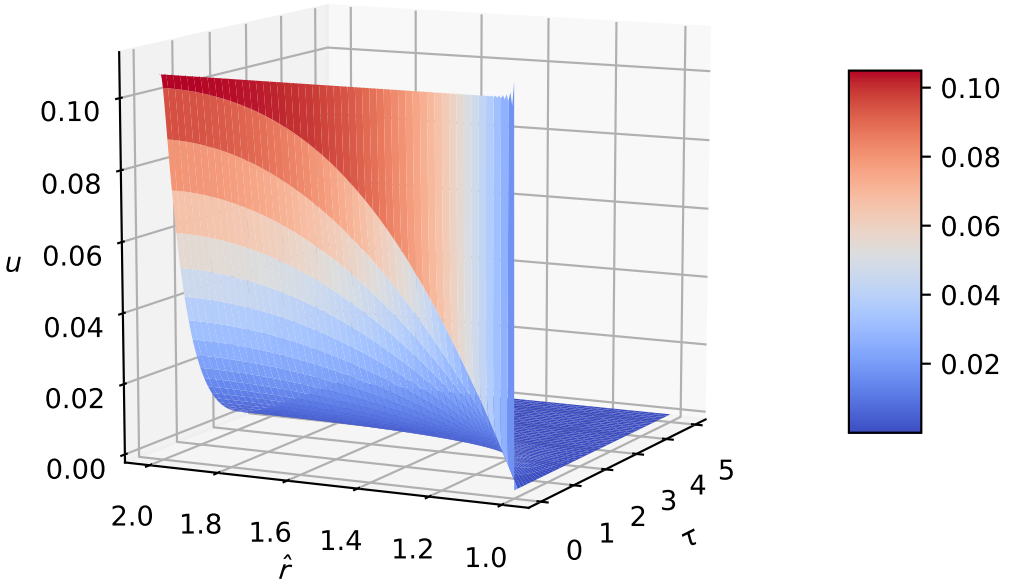


Figure 2.3: When $\tau = 0$, the approximated initial value of u is about 0.106079414 while the analytical result is $1/3\pi \approx 0.106103295$. The absolute error is 2.3881×10^{-5}. Moreover, the Gibbs phenomena [28] happened obviously near the discontinuity $\hat{r} = 0$ with the overshoots and undershoots because of the Fourier-Bessel series.

$$
\begin{aligned}
S(0) &= \sum_{n=1}^{\infty} \frac{4}{\mu^2 - 1} \frac{1}{\lambda_{0,n} \left\{ \left[\frac{J_0(\sqrt{\lambda_{0,n}})}{J_0'(\mu\sqrt{\lambda_{0,n}})} \right]^2 - 1 \right\}} \\
&= \sum_{n=1}^{\infty} c_n
\end{aligned}
\tag{2.16}
$$

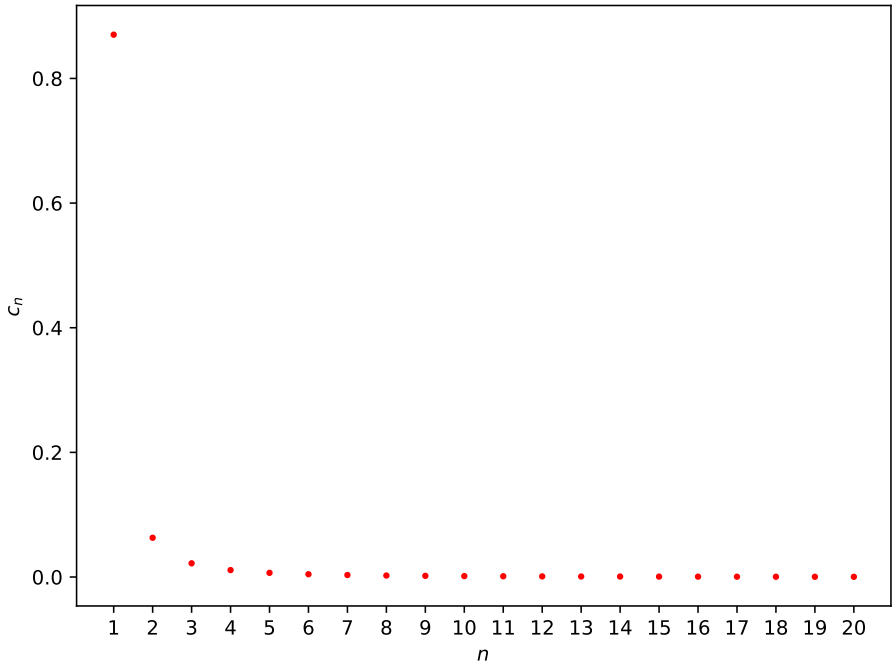
As displayed in Figure 2.4a, the individual terms of the series Eq. 2.16 are monotonically decreasing and approach to 0. Calculating the 1000−th partial sum directly, $S(0)$ is 0.9998649050990541 to 16 decimal places with 3 digit accuracy. However, Aitken's method can speed up the convergence rate of the partial sum sequence and improve the accuracy of the estimation by constructing a new series. Moreover, it does not matter whether the series is alternating or not. Aitken's acceleration is

$$
A_n^{(1)} = s_{n+1} - \frac{(s_{n+1} - s_n)^2}{s_{n+1} - 2s_n + s_{n-1}} \tag{2.17}
$$

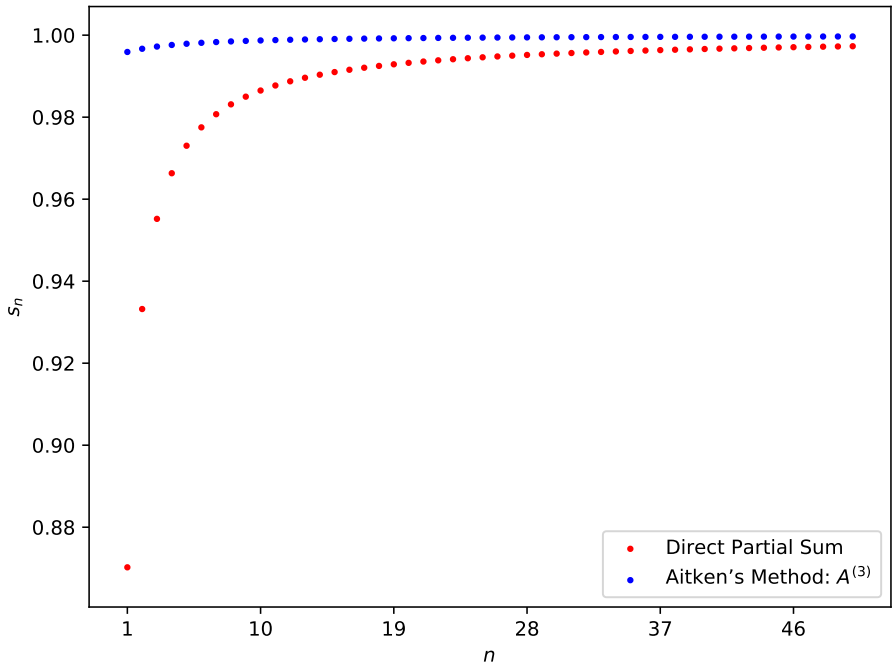
where s_n, $n = 1, 2, 3, ...$, is the n−th partial sum of Eq. 2.16, and (s_{n-1}, s_n, s_{n+1}) are three successive partial sums. The process can be repeated for the further speed-up. For example, apply Eq. 2.17 on the new series $A^{(1)}$ to generate $A^{(2)}$, on the series $A^{(2)}$ to obtain $A^{(3)}$, etc.

As displayed in Figure 2.4b, instead of adding 1000 terms of Eq. 2.16 directly, the result of sequence transformation by Aitken's method is more precise, which is 0.9999812339391291 to 16 decimal places with 4 digit accuracy. The error using Aitken's acceleration is 7.2 times smaller than simply using the partial sums.

Similarly, when $\tau > 0$, estimating $S(\tau)$ and $\langle\tau\rangle$ in Eq. 2.12 by Aitken's acceleration, shown in Figure 2.5, for validating the research methodology in the next chapter.



(a) The first 20 terms of the series Eq. 2.16 are convergent and tend to zero.



(b) The sequence of the partial sums of the series Eq. 2.16 tends to a real limit 1. Compared with calculating the sum directly, $A_n^{(3)}$ converges more rapidly and approaches closer to 1.

Figure 2.4: $S(0)$ is approximated by the direct summation and Aitken's acceleration.

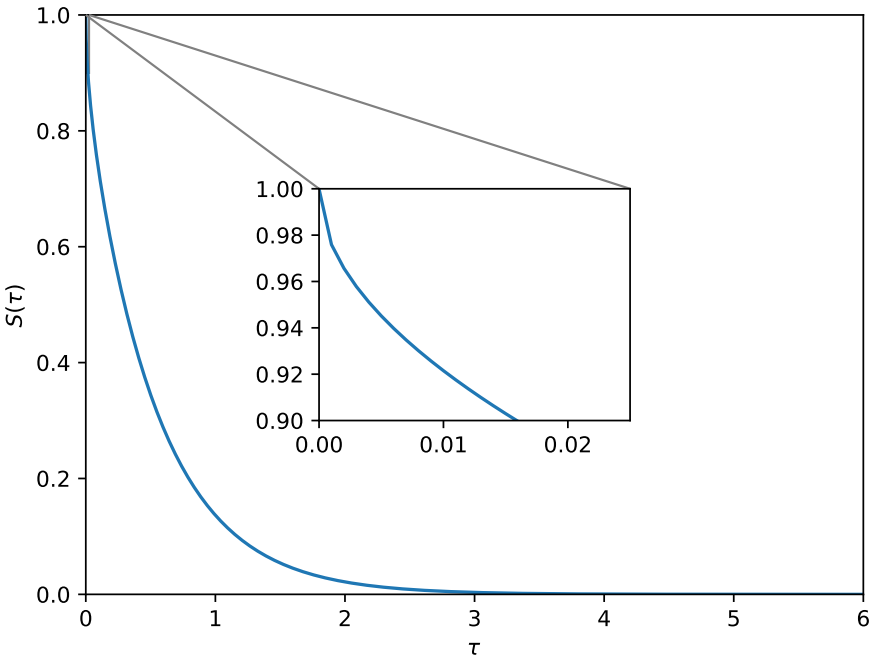


Figure 2.5: The asymptotic behaviours of survival probability $S(\tau)$ is approximated by the Aitken's acceleration. $S(\tau)$ monotonously decreases from 1 at $\tau = 0$ to 0 as τ goes to infinity. Moreover, the approximation of analytical mean first-passage time $\langle\tau\rangle$ equals 0.47339248149521174..

2.3 Monte Carlo Simulations

In this section, several generally utilized numerical methods [33][76] [27] [5] for solving the heat equation, and their limitations in practice are presented. Also, one of the non-deterministic algorithms, Monte Carlo methods (MCM) [63] [50], and its application in approximating the solution of the PDEs are proposed. As the weaknesses and challenges of applying the numerical techniques in solving $2-$dimensional heat equation defined in the real root images with millions of pixels and extremely complex root systems, the alternative fixed-time step Monte Carlo simulations, lattice random walks (LRWs) and Pearson's random walks (PRWs), are designed. The most outstanding advantage of the proposed random walk models is that the integration, named the heat content, can be approximated directly based upon the probabilistic interpretation of Brownian motion and the heat equation. Finally, the methods to analyze the output of the Monte Carlo simulations and solve the sampling-related problems in the simulations are brought up theoretically.

2.3.1 Background

In the subsection 2.2.1, the analytical solutions of the heat equation defined in the annulus with the initial and boundary conditions have been derived. The analytical heat content, calculated by the integration over Ω, implies the geometrical properties of the annulus. However, the analytical method for solving the heat equation has many restrictions, and its applications to practical problems will exhibit difficulties. Firstly, the numerical evaluation of the analytical solutions is usually by no means trivial because they are in the form of infinite series. Secondly, either irregular geometries or discontinuities lead to the complexities, so the explicit algebraic solutions are close to non-existed. Thirdly, the purely analytical techniques can apply strictly only to the linear form of the boundary conditions and to constant diffusion properties [16]. Therefore, numerical methods and computer simulations are more helpful and applicable to find solutions to the heat equation than calculating pure analytical solutions.

Numerical Methods

The techniques for solving initial-boundary value problems (IBVPs) based on numerical approximations have existed for a long time and been developed considerably including the finite-difference method (FDM), finite element method (FEM), finite volume method (FVM), boundary element method (BEM), and so forth.

FDM is frequently utilized to converting the heat equation into a system of algebraically solvable equations [33]. The basic idea is to replace the derivatives in the equation by the difference quotients. For example, the FTCS (Forward Time Centered Space) scheme [59] discretize the Laplace operator in space and the time derivative and implement the boundary conditions on the staggered grid for representing the original continuous problem, but it is conditionally stable [59]. If the spatial resolution becomes doubled, the time-step should be reduced by a factor of four to maintain the numerical stability, which causes the extremely tiny time-step in the high-resolution calculations. There are three kinds of errors needed to be considered when using FDM. First of all, in the derivation of the finite-difference equations, the higher-order terms in the Taylor series are neglected, constituting the truncation error. If the time and space interval tends to 0, the truncation errors will approach 0, or the FDM is incompatible or inconsistent with the original heat equation [16]. Another class of error appearing in FDM, called round-off error, results from the loss of precision due to the computer rounding of decimal quantities. [37]. The last type of error is the discretization error, which can be reduced by decreasing the time size, grid size, or both [16]. Moreover, DFM becomes less accurate and hard to implement when the problem is defined in the irregular geometries since the heat equation must be transformed before applying the Taylor series.

Unlike the FDM, FEM [76] divides the complicated and irregular geometries and boundaries into the union of smaller and simpler subdomains or finite elements [54], e.g. lattice, triangle, curvilinear polygons, etc. Each subdomain is locally represented by the element equation, continuous piecewise shape functions, which are finally assembled into a larger system of algebraic equations for modelling the entire problem. The numerical solution can be obtained by minimizing the associated error function to meet the certain specification of the accuracy. The smaller size of the finite element mesh, the more accurate the approximate solution. FEM has great flexibilities or adaptivity [61]. For instance, FEM can provide higher fidelity or accuracy in a specific local region and keep elsewhere identical. Nevertheless, FEM requires an amount of human involvement in building the FE model, checking the result, detecting and updating the model design. Moreover, compared with FDM, FEM demands a longer execution time and a larger amount of input data.

FVM, closely related to FEM, converts the original heat equation into the integral forms [27]. However, the accuracy of FVM is related to the numerical integration over time and space dimensions. Unlike the domain-type methods (e.g. FDM, FEM, FVM, etc.), BEM transforms the heat equation into an integral equation over the boundary of the domain using the boundary integral equation method [5]. Especially, when the domain extends to infinity or the boundary is complex, BEM is more efficient in computation than other methods because of the smaller surface or volume ratio [47] since it only discretizes the boundary and fits the boundary values into the integral equation [4]. However, it is arduous to solve the matrics generated in

[Yuge 1] The writing of this part has been revised. Dave, please give me the feedback or comments on it.

BEM, since they are generally unsymmetric and fully populated [57].

In summary, all the described numerical methods have an intrinsically similar feature - mesh discretization in the time and space dimension. In this thesis, the heat equation is defined in a 2-dimensional domain, bounded by the border of the image and that of the root system, with millions of pixels, the extremely complex roots and various boundary conditions. It may be possible to calculate the heat content contained in this domain by the numerical integration of the solutions approximated by those numerical techniques. However, some practical problems have to be taken into consideration. For instance, the far more efforts are required when applying FDM and FVM because of the complicated boundary of the roots and non-continuous issues. Although the whole 2-dimensional root image can be viewed as a discretized domain, it is still time-consuming and challenging to trace and identify the boundary of roots, label the nodes, and generate the coordinates and connectivities among the nodes in the preprocessing stage of FEM. The finer discretization, the more accurate approximated solutions of the original IBVP and the longer computational time spent by the numerical methods. More importantly, the heat content, which is the integration of the numerical solution over the space dimension, should also be approximated numerically resulting in extra effort and errors.

Probabilistic Interpretation

[Yuge 2] This writing has been revised. Dave, please give me the feedback and comments. Thanks!

In the subsection 2.2.1, the heat equation describes the temperature distribution of a homogeneous and isotropic domain [72], and its solution characterizes how the temperature changes over the position and time. From the probabilistic perspective, the heat equation and its solution can also be understood by the Brownian motion [12]. The Brownian motion also called the Wiener process, is a continuous-time and continuous-space stochastic process [46] with the continuous sample paths and stationary independent increments [40]. This process also has the Markov property: the future state depends only on the present state [9]. In the probability theory, if a large number of free particles undergoing the Brownian motion independently, the density of particles at a specific time becomes a deterministic process called diffusion, which satisfies the heat equation [41][72].

Survival Probability For simplicity, we only investigate the probabilistic interpretation of the heat equation defined in the annulus with the boundary and initial conditions as same as described in the subsection 2.2.1. Consider a particle undergoing the Brownian motion from $(\hat{r_0}, \theta_0) \in \Omega$ at $\tau = 0$ and let $P(\hat{r}, \theta, \tau | \hat{r_0}, \theta_0, 0)$ be the conditional probability of finding the particle at $(\hat{r}, \theta) \in \Omega$ at time $\tau > 0$. $P(\hat{r}, \theta, \tau | \hat{r_0}, \theta_0, 0)$ satisfies the following equations

$$\frac{\partial P(\hat{r}, \theta, \tau | \hat{r_0}, \theta_0, 0)}{\partial \tau} = \Delta P(\hat{r}, \theta, \tau | \hat{r_0}, \theta_0, 0) \qquad \text{for } (\hat{r}, \theta) \in \Omega \tag{2.18}$$

$$P(\hat{r}, \theta, \tau | \hat{r_0}, \theta_0, 0) = 0 \qquad \text{for } (\hat{r}, \theta) \in \partial\Omega_1 \tag{2.19}$$

$$\frac{\partial}{\partial \hat{r}} P(\hat{r}, \theta, \tau | \hat{r_0}, \theta_0, 0) = 0 \qquad \text{for } (\hat{r}, \theta) \in \partial\Omega_2 \tag{2.20}$$

$$P(\hat{r}, \theta, \tau | \hat{r_0}, \theta_0, 0) = \frac{1}{\hat{r}} \delta(\hat{r} - \hat{r_0}) \delta(\theta - \theta_0) \qquad \text{for } \tau = 0 \tag{2.21}$$

Note, Eq. 2.18 is the heat equation and supplemented by the initial condition Eq. 2.21 with the Dirac $\delta-$ distribution, absorbing boundary condition Eq. 2.19, and reflecting boundary condition Eq. 2.20. Similarly, the conditional probability $P(\hat{r}, \theta, \tau | \hat{r_0}, \theta_0, 0)$ can be expressed in terms of the eigenvalues and eigenfunctions.

Define the local survival probability $S(\tau | \hat{r_0}, \theta_0, 0)$ as the probability that the particle, localized initially at $(\hat{r_0}, \theta_0)$, keeps diffusing in Ω at time $\tau > 0$ without being absorbed by Ω_1. It can be calculated by

$$S(\tau | \hat{r_0}, \theta_0, 0) = \iint_\Omega \hat{r} P(\hat{r}, \theta, \tau | \hat{r_0}, \theta_0, 0) d\hat{r} d\theta \tag{2.22}$$

Thus, the survival probability $S(\tau)$ for a particle, starting with an initial distribution uniformly spread in Ω, is the average of the local survival probability over Ω. It is

$$S(\tau) = \frac{1}{|\Omega|} \iint_\Omega \hat{r_0} S(\tau | \hat{r_0}, \theta_0, 0) d\hat{r_0} d\theta_0 \tag{2.23}$$

Eq. 2.22 and Eq. 2.23 reveal that the heat content $Q_\Omega(\tau)$ in Eq. 2.12, calculated as the integration of the temperature over Ω, is proportional to the survival probability $S(\tau)$ [43].

Mean First-Passage Time The first passage phenomena play a fundamental role in the stochastic processes triggered by a first-passage event [71]. In this thesis, the first-passage event refers to the first time when the Brownian motion is stopped because of the present of the Dirichlet boundary condition. One of the essential first-passage-related quantities is the first-passage time or the first-hitting time [62], which is the time taken by particle undergoing the Brownian motion from an initial position to any sites of Ω_1 for the first time. Particles' mean first-passage time $\langle\tau\rangle$, also called the average first-passage time, has a closed relationship with the survival probability $S(\tau)$ [62]

$$\begin{aligned}\langle\tau\rangle &= \int_0^\infty \tau dS(\tau) \\ &= \sum_{n=1}^{\infty} \frac{4}{\mu^2-1} \frac{1}{\lambda_{0,n}^2 \left\{ \left[\frac{J_0(\sqrt{\lambda_{0,n}})}{J_0'(\mu\sqrt{\lambda_{0,n}})} \right]^2 - 1 \right\}}\end{aligned} \tag{2.24}$$

From Eq. 2.24 and Eq. 2.11, it is clear that $\langle\tau\rangle$ implies an overall property of the annulus since it only depends on the radius ratio of annulus μ.

Brownian Motion and Random Walks Brownian motion, the irregular motion of individual particles, has been existed for a long time before the random-walk theory was developed. At the beginning of the twentieth century, the term, random walk, was initially proposed by Karl Pearson [58]. He utilized the isotropic planar random flights to model how mosquitoes migrate and invade randomly in the cleared jungle regions. At each time step, the mosquito moves to a random direction with a fixed step length. Rayleigh [60] answered Pearson's question in the same year, that is, the distributions of mosquitos after many steps have been taken is identical to superposition the sound vibrations with unit amplitude and arbitrary phase [19]. At almost the same time, Louis Bachelier [6] designed a model for the financial time series based on the random walks. Louis Bachelier also explored the relationship between discrete random walks and the continuous heat equation. During the development of random-walk theory, many other scientific fields, including the random processes, random noise, spectral analysis, and stochastic equations, were developed by some physicists [24] [23] [68]. The continuous Brownian motion is the scaling limit of the discrete random walks as the time and space increments approach zero [52][72].

Lattice Random Walks (LRWs) Let us consider a particle performing the simple random walk on the d-dimensional integer grid $\mathbb{Z}^d$. It is a discrete-space and discrete-time symmetric hopping process [62] on the lattice. At each time step, the particle moves to one of its $2d$ nearest neighbours with probability $\frac{1}{2d}$. If $d \leq 2$, the random walk is recurrent, which means the particle will return to its origin infinitely often with the probability 1. If $d \geq 3$, the random walk is transient, which indicates the particle will return to its origin only finitely often with the probability 1 [39] [52]. Only 2-dimensional lattice random walks (LRWs) and its connection with the heat equation are introduced since this thesis aims to explore and characterize the shape of roots in the $2-$dimensional images.

Let $\triangle l$ be the distance between two sites in the lattice and δ be the time step. Let $p(i,j,n)$ be the probability of finding a particle to be in position (i,j) after n steps. There has

$$p(i,j,n) = \frac{p(i-\triangle l,j,n-1)+p(i+\triangle l,j,n-1)+p(i,j-\triangle l,n-1)+p(i,j+\triangle l,n-1)}{4} \tag{2.25}$$

$$\begin{aligned} p(i,j,n)-p(i,j,n-1) = &\frac{1}{4}\left(p(i-\triangle l,j,n-1)-2p(i,j,n-1)+p(i+\triangle l,j,n-1)\right) \\ &+\frac{1}{4}\left(p(i,j-\triangle l,n-1)-2p(i,j,n-1)+p(i,j+\triangle l,n-1)\right) \end{aligned} \tag{2.26}$$

$$x = i\triangle l \tag{2.27}$$

$$y = j\triangle l \tag{2.28}$$

$$t = n\delta \tag{2.29}$$

Eq. 2.27 and Eq. 2.28 can be used to definde particle's position in the $x-y$ coordinate system and Eq. 2.29 is the walking time taken by the particle after n steps.

$$p(i,j,n)-p(i,j,n-1) = p(x,y,\frac{t}{\delta})-p(x,y,\frac{t-\delta}{\delta}) \tag{2.30}$$

$$\approx \delta\frac{\partial p(x,y,t)}{\partial t} \tag{2.31}$$

Similarly,

$$p(i-\triangle l,j,n-1)-2p(i,j,n-1)+p(i+\triangle l,j,n-1) \approx (\triangle l)^2\frac{\partial^2 p(x,y,t)}{\partial x^2} \tag{2.32}$$

$$p(i,j-\triangle l,n-1)-2p(i,j,n-1)+p(i,j+\triangle l,n-1) \approx (\triangle l)^2\frac{\partial^2 p(x,y,t)}{\partial y^2} \tag{2.33}$$

Aftering substitue Eq.2.31, Eq. 2.32, and Eq. 2.33 into Eq. 2.26, $p(x,y,t)$, the probability of particle at (x,y) at time t, satisfies

$$\frac{\partial p(x,y,t)}{\partial t} = \frac{(\triangle l)^2}{4\delta}\left(\frac{\partial^2 p(x,y,t)}{\partial x^2}+\frac{\partial^2 p(x,y,t)}{\partial y^2}\right) \tag{2.34}$$

where $D = \frac{(\triangle l)^2}{4\delta}$ is the diffusion coefficent. This above derivation shows a tight relationship between the $2-$ dimensional discrete lattice random walks and the heat equation.

Pearson's Random Walks (PRWs) Based on Pearson's problem and Rayleigh's answer, Stadje [69] and Masoliver et al. [55] considered a two-dimensional continuous-time and continuous-space random walk,

defined as Pearson's random walks (PRWs) in this thesis. In PRWs, particle moves with constant speed and with random directions distributed uniformly in $[0, 2\pi)$. Moreover, the lengths of the straight-line paths and the turn angles are stochastically independent. If the mean step length approaches zero and the walking time is big enough, the behaviours of particles in PRWs weakly converges to the Wiener Process [69], which satisfies the traditional heat equation.

Monte Carlo Simulations For Solving PDEs

Monte Carlo methods (MCMs), the commonly used computational techniques, aim to generate samples from a given probability distribution, estimate the functions' expectations under this distribution, and optimize the complicated objective functions by using random numbers [50] [63]. MCMs can be used to solve the IBVPs by generating the random numbers to simulate the successive positions of the trajectory of a stochastic process at fixed instants [51][48], since the original continuous problem can be represented by the probabilistic interpretation and the solution can be approximated by the expectation of some functional of the trajectories of a stochastic process [31][64]. Therefore, unlike the numerical techniques proposed in the subsubsection 2.3.1, the nondeterministic Monte Carlo simulations are grid-free on the domain, boundary, and the boundary conditions of the problem [31].

Monte Carlo simulations have been applied frequently in solving the elliptic partial differential equation, for example, the Laplace's and Poisson's equations [34] [11] [56]. For example, let $u(P_0)$ be the value of the solution of the elliptic partial differential equation at a specific point P_0 in a bounded domain. $u(P_0)$ can be estimated by a point P_1, which is sampled uniformly on the largest circle C_0 centered at P_0 with radius r_0 lying entirely in the domain. If P_1 gets closed to the target bounday within an error, $u(P_1)$ is known, and it can be considered as a particle's estimate of $u(P_0)$ by multiplying the particle's statistical weight. If not, $u(P_1)$ is should be estimated in the same way as $u(P_0)$, that is, P_2 is sampled uniformly on the largest circle C_1 centered at P_1 with radius r_1 lying entirely in the domain. Check the position of P_2, and the procedure will be repeated until the simulation is terminates on the traget boundary, which is defined as one particle's estimate of $u(P_0)$. Finally, averaging a larger number of one-particle estimates, $u(P_0)$ will be more accurate.

However, Monte Carlo simulations are barely applied in solving parabolic and hyperbolic partial differential equations, such as the $1-$ dimensional and $2-$ dimensional time-dependent heat problem. As introduced in the papers [65] [29], after obtaining the probabilistic interpretation of the finit-difference approximation of the heat equation, the solution of the heat equation at a specific space-time point can be approximated by averaging a large number of random-walking particles, whose trajectories are simulated by the Monte Carlo methods untill they hit the any sites of the target boundary. However, the drawbacks of this method are obvious. Firstly, there is the error appeared in the finit-difference approximation. Secondly, there has the statistical sampling errors inherent in the Monte Carlo simulations. Last but not least, it will be time-comsuming to evaluate the solution defined in the whole domain, the real root images with millions pixels in this thesis, since this method can only be used to approximate the solution of the heat equation at one point at a time.

2.3.2 Algorithms of Random Walks

In the subsection 2.3.1, the practical challenges of solving the heat equation defined in the $2-$ dimensional domain with the complicated root systems by numerical methods and Monte Carlo simulations are revealed, and the probabilistic interpretation of the heat equation, survival probability, and random walks are introduced. From Eq. 2.12, Eq. 2.23, and Eq. 2.15, the final goal in this thesis is to approximate the heat content, or the survival probability, defined as the integrals, which only depends on the time variable. In other words, our interest is only related to the time when the first-passage event happens in the stochastic process, that is, the first-passage time. Therefore, in this subsection, two random walks algorithms are designed to mimic particles' first-passage time by $2-$ dimensional fixed-time step Monte Carlo simulations in the real root images for approximating the integration as expressed in Eq. 2.23 and Eq. 2.15.

Lattice Random Walks

input :
output:

Algorithm 1: Lattice Random Walks (LRWs)

Pearson's Random Walks

input :
output:

Algorithm 2: Pearson's Random Walks (PRWs)

2.3.3 Output Analysis

The output of the fixed-time step Monte Carlo simulations are particles' first passage time t, which is the number of steps taken by the particles hitting any positions of the target for the first time. Since first-hitting-time models are a sub-class of survival analysis in statistics [3], it is straightforwad to use the Kaplan-Meier estimator to estimate the survival function $S(t)$ [49] of the numerical simulation, which provide the probability that a particle remains wandering beyound any specified time. Moreover, the pointwise upper and lower confidence interval can also calculatated by the Greenwood's exponential formula [38]. In this subsection, the Kaplan-Meier estimator and confidence interval of $S(t)$ are introduced theoretically. However, in practice, the existed Python module, lifeline [18], will be used to implement the estimation. After obtaining the estimated survival function of the fixed-time step Monte Carlo simulations, the scaling relationship between t and τ is derived for the valiation of research methodology in the next chapter.

Kaplan-Meier Estimator

The general definition of the survival time is the time starting from a specified point to the occurrence of a given event [8], such as death, pregnancy, job loss, etc. Also, the analysis of the group of survival data is called survival analysis [3]. In the survival analysis, three kinds of situations will affect the subjects' survival time [30]. Firstly, the subjects are uncooperative and refused to continue to participate in the research. Secondly, some subjects do not experience the event before the end of the study, but they would have experienced the event if they keep being observed. Finally, the researchers lose touch with the subjects in the middle of the investigation. In practice, since these subjects have partial information about survival, the scientists will label the above circumstances as censored observations [8] instead of ignoring them and decreasing sample size.

In clinical trials or community trials, Kaplan-Meier Estimator [44], a non-parametric analysis, is a commonly applied statistical method in the survival analysis for the measurement of the fraction of the survival time after the treatment [1] and for generating the corresponding survival curve. It also works well with the mentioned three difficult situations. With various assumptions [25] [30], the Kaplan-Meier survival curve can be created and provides the probability of surviving in a given length of time while considering the time in many small intervals [3].

Let $0 < t_1 < t_2 < ...$ be the distinct increasing observed times, or the number of steps taken by the particle countering the absorbing boundary, in the sample. Let n_i be the number of particles who either have not yet stopped moving up to time t_i or else who are absorbed on the target boundary at time t_i in the simulations. Let d_i the number of particles hitting the target boundary at time t_i. The Kaplan-Meier or product-limit estimator $\widehat{S}(t)$ of the survival function of the numerical simulation $S(t)$ is [1]

$$\widehat{S}(t) = \prod_{i:t_i \leq t} (1 - \frac{d_i}{n_i}) \tag{2.35}$$

Confidence Interval

The upper and lower $(1-\alpha)\times 100\%$ confidence intervals of the survival function $S(t)$ for a fixed time t was firstly proposed and derived by Greenwood in 1926 [32],

$$\widehat{S}(t) \pm z_{\alpha/2}\sqrt{\widehat{Var}[\widehat{S}(t)]} \qquad \text{where} \tag{2.36}$$

$$\widehat{Var}[\widehat{S}(t)] = \widehat{S}(t)^2 \sum_{t_i \le t} \frac{d_i}{n_i(n_i - d_i)} \tag{2.37}$$

Note, z_α is the $\alpha-$th quantile of the normal distribution.

In 1999, Hosmer and Lemeshow [38] developed the exponential Greenwood formula based on the earlier works of Kalbfleisch and Prentice [42], which provides an asymmetric confidence interval for $S(t)$

$$e^{-e^{c_+(t)}} < S(t) < e^{-e^{c_-(t)}} \qquad \text{where} \tag{2.38}$$

$$c_\pm(t) = log(-log\widehat{S}(t)) \pm z_{\alpha/2}\sqrt{\widehat{V}} \qquad \text{and} \tag{2.39}$$

$$\widehat{V} = \frac{1}{(log\widehat{S}(t))^2} \sum_{t_i \le t} \frac{d_i}{n_i(n_i - d_i)} \tag{2.40}$$

Note, if $c1 < c2$, there has $e^{-e^{c_2}} < S(t) < e^{-e^{c_1}}$.

Compared with the traditional Greenwood confidence intercal calculation, the exponential Greenwood formula will make sure that the endpoints in Eq. 2.38 lie in $(0, 1)$, while the endpoints in Eq. 2.36 could be negative or larger than 1 [66].

Relationship between t and τ

Particles' average one-step displacement $\triangle l$ in the fixed-time step Monte Carlo simulations, LRWs and PRWs, are associated with the time step is δ:

$$\triangle l = 2\sqrt{D\delta} \tag{2.41}$$

where D is the diffusion coefficent.

Eq. 2.41 implies that the time step δ must be designed small enough to make sure that $\triangle l$ is shorter than the smallest geometrical features of the boundaries. Thus, $\triangle l$ should equal or be less than one-pixel size in the simulations. Furthermore, the δ is regarded as a fundamental bridge between particles' number of steps t and unitless continuous-time τ,

$$\tau = t\delta = \frac{(\triangle l)^2 t}{4D} \tag{2.42}$$

where D is 1.

When running the LRWs in the annulus, $\triangle l$ is always $\frac{1}{100}$ since particle's step length is as same as one-pixel size. However, $\triangle l$ is related to the particle's step length in PRWs. If particle's step length is 0.5, a half of a pixel, then $\triangle l$ equals $\frac{1}{100} \times \frac{1}{2} = \frac{1}{200}$. Similiarly, when the step length in PRWs is 0.1, then $\triangle l$ is $\frac{1}{1000}$.

2.3.4 Sample Size Determination

Based on the law of large number (LLN) [20], as the sample size approaches infinity, the sample mean tends to get closer to the true population mean with the high probability. Moreover, the central limit theorem (CLT) [20] illustrates how the sampling distribution of the mean changes as a function of the sample size and how much more reliable a large experiment is. On the downside, the increasing number of trials results in the higher cost of performing the simulation, which is the major drawback of the fixed time-step Monte Carlo simulations. Therefore, it is necessary to conduct the minimum number of simulation runs to achieve a desired degree of precision.

General Method

There are five parts of the standard way to determine the sample size in the Monte Carlo simulations. Firstly, simulating with a certain amount of samples. Secondly, repeating the simulation several times with the sample size as the same as the first step. Thirdly, increasing the number of samples and implementing the first two steps again. Fourthly, running a regression analysis of the variability of the sample statistic as a function of sample size. Finally, estimating the sample size that will result in any desired level of convergence by some probabilistic inequalities, including Chebyshev's inequality [15], Cantelli's inequality [13], Vysochanskij–Petunin inequality [74], etc.

Dvoretzky–Kiefer–Wolfowitz (DKW) inequality

Although the sample size estimated by the general method does not depend on the geometries characterized by the random walk models, how long the simulation will run is unknown, and the final calculated sample size will be larger than the really necessary value sometimes. Therefore, an alternative method, named the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality [22], is proposed for the sample size determination by generating the simultaneous confidence interval of the survival function without simulating random walk models and considering the shape of objects in the images.

Let $F_N(x)$ denote the empirical distribution functions (empirical CDF) for a sample of N real-valued *i.i.d.* random variables, $X_1, ..., X_N$, with continuous cumulative distrbution function (CDF) $F(x)$. The DKW inequality, as expressed in Eq. 2.43, bounds the probability that the random function $F_N(x)$ differs from the true $F(x)$ by more than a given constant ε [22].

$$Pr(\sup_{x\in\mathbb{R}} |F_N(x) - F(x)| > \varepsilon) \leq 2e^{-2N\varepsilon^2} \quad for\ every\ \varepsilon > 0 \tag{2.43}$$

The equally spaced confidence bounds or simultaneous band around the F_N encompassing the entire $F(x)$ can be expressed by

$$F_N(x) - \varepsilon \leq F(x) \leq F_N(x) + \varepsilon \tag{2.44}$$

On the other hand, assume the simultaneous band produced by Eq. 2.44 containing the $F(x)$ at a given confidence level $1 - \alpha$, the interval ε can be calculated by

$$\varepsilon = \sqrt{\frac{\ln \frac{2}{\alpha}}{2N}} \tag{2.45}$$

Given a converge probability α and constant ε, it is straightforward to estimate the sample size N in the fixed-time step Monte Carlo simulations in any images by Eq. 2.45.

2.4 Two-sample Statistical Tests

The differences between the survival curves generated by the Kaplan-Meier estimator are visible sometimes. However, the dissimilarities won't be easily detected by eyes if the survival curves are overlapping over some parts or crossing at some time points. Since the Kaplan-Meier estimator does not provide any information on whether two groups of survival data are statistically similar or different, some popular statistical tests used specially in the survival analysis course are presented in this section. Which test should be selected in a specific circumstance is always debated because there is a fine line between the statistical tests in the survival analysis. Therefore, acknowledging the data in hand and identifying the assumptions well is a prerequisite to determine the tests appropriately.

Before listing the pros and cons of several statistical tests, the censored survival times will be recalled firstly, which indicates the time at which a subject is unobserved and the time to the event of a subject is not recorded [26]. In this thesis, it is possible to appear the censoring observation in the beginning or at any other moment during the Monte Carlo simulations. If the simulation finished, but the particle did not reach the target boundary, the particle will be regarded as a right-censored. When the particle is abandoned and not been observed during the simulations, it is termed the random right censoring [26]. Another cause of a deficient observation of particles' survival times is the left censoring, which hints that the particles had stopped diffusing before the simulation began. For instance, the particle is generated in or on the pixels of roots. As mentioned in the last section, the Kaplan-Meier method can still cope well with the right-censored and left-censored observations in output.

In survival analysis, as the time interval gets close to 0, the instantaneous hazard rate can be calculated by limiting the number of events per unit time divided by the number of events at risk [14]. The hazard ratio is an estimate of the hazard rate in one group relative to that in another group [67]. If the survival curves are parallel with the identical shape, the hazard ratio is constant at any interval of time. In this situation, the log-rank tests, also named the Mantel-Haenszel, are reliable [17].

If the hazard ratio does not satisfy the assumption, the log-rank test will not be powerful to detect the differences in the survival functions. In such a case, the Gehan-Breslow-Wilcoxon test, also called Gehan's generalized Wilcohon procedure, should be considered alternatively [2]. Also, under the constant hazard ratio assumption, the Wilcoxon tests might be more reliable than the log-rank tests [17]. The former one gives more weight to the early failures, but the latter one is more suitable for comparing the later events in the data [17]. Generally, some general non-parametric tests, based on the rank ordering (e.g. Mann-Whitney U test, Kruskal-Wallis, etc.), are not always feasible in censoring survival data [2]. However, Gehan's generalized Wilcohon test is still robust when the censoring rates are low, and the censoring distributions of groups are equal [45].

Neither log-rank test nor Gehan's generalized Wilcohon test can work well when the survival curves cross while the Tarone-Ware test should be chosen [53]. It pays more attention to the failures happening

somewhere in the middle of study [25]. Moreover, there is no limitation of the number of groups when the Tarone-Ware test is applied [17]. Similarly, the Fleming-Harrington test is also accessible and robust for testing the differences between two or more survival curves in the right-censored data based on the counting process [36].

2.5 Research Design

...

2.5.1 Methodology

2.5.2 Idea

Fixed-time Step Monte Carlo Simulations on Artificial Images

3.1 Background

...

3.1.1 Purposes

3.1.2 Shape Design

3.1.3 Branching Structures

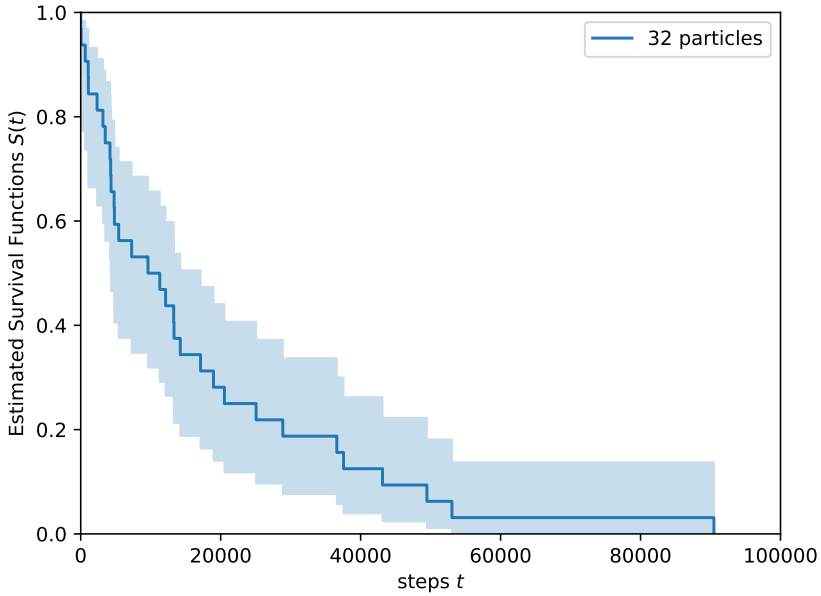
3.2 Methodology Validation

In the last chapter, the survival probability $S(\tau)$ and the mean first-passage time has been calculated by solving the heat equation and approximated by the numerical methods. Lattice Random Walks (LRWs) and Pearson's Random Walks (PRWs) are implemented in the annulus image, as shown in Figure 2.1, in Python. This section aims to validate the research methodology by comparing the estimated survival functions $S(t)$ of the numerical data with the analytical solutions $S(\tau)$ where t is the number of steps taken by the particles in the fixed-time step Monte Carlo simulations and τ denotes the unitless time.

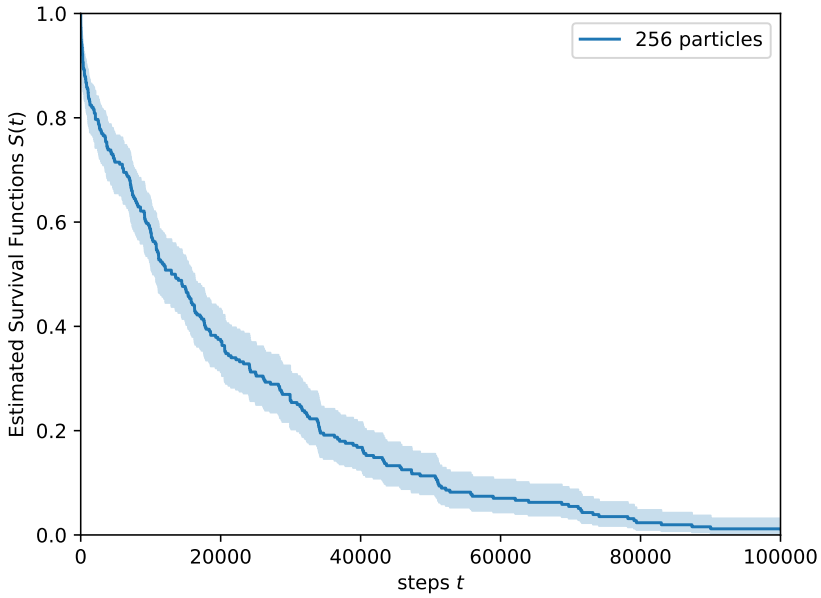
3.2.1 Statistical Fluctuation Analysis

Fixed-time step Monte Carlo simulations, LRWs and PRWs, are the nondeterministic numerical representations of the original statistical problem defined in the continuous-time and continuous-space with numerous inputs and discrete-time trajectories. In the simulations, the initial positions of the enormous number of particles and their moving directions at each time step are determined by the randomly uniform sampling. Thus, it is inevitable to appear the statistical fluctuations, also called variance, defined as a measure of the discrepancies between the estimate and the true solution. The brute-force way to reduce the statistical fluctuations is to increase the sample size.

The first kind of error stems from the sampling. As shown in Figure 3.1a and Figure 3.1b, the larger sample size in the simulations, the estimated survival functions will be more precise. LRWs are used to mimic the continuous-time and continuous-space diffusion process by generating the discrete random trajectories in the discrete time, which results in the time-discretization and space-discretization errors. Although PRWs is a model defined in continuous-time and continuous-space, the random paths demand much longer time simulation as shown in Figure 3.2.



(a) Survival curve for the LRWs with 32 particles.



(b) Survival curve for the LRWs with 256 particles.

Figure 3.1: Estimated survival curves and 95% confidence intervals for Monte Carlo simulations of partical diffusion on an annulus. As the number of particles increasing, the uncertainty of the LRWs simulation are lower since the confidence band of the estimated survival function becomes narrower.

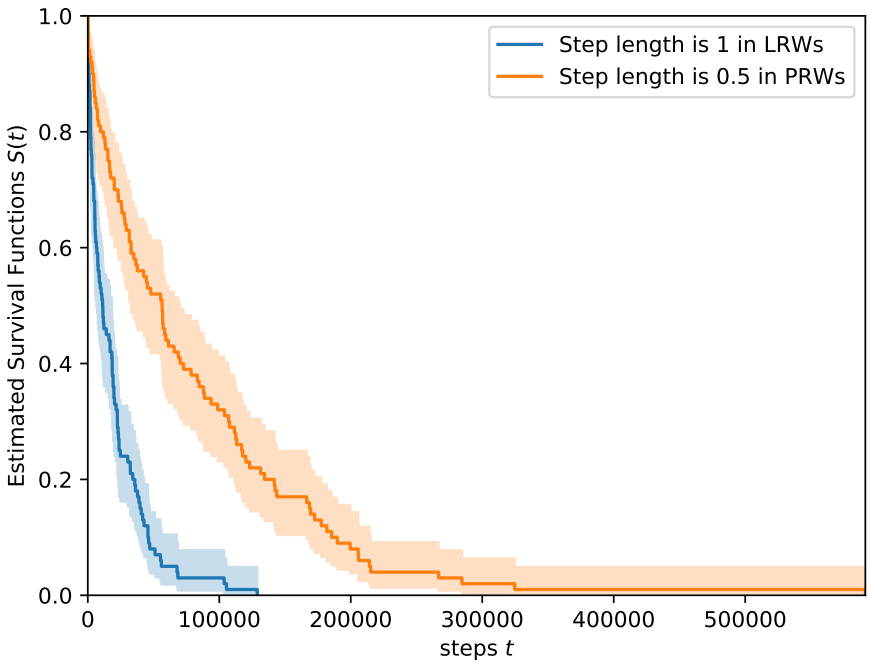


Figure 3.2: When run LRWs and PRWs in the annulus with 100 particles, the finer discretization step results in the longer simulation time.

Test Methods (standard nonparametric)	Statistics	P Values
Logrank	0.017679	0.894223
Fleming-Harrington	0.742536	0.388850
Gehan-Breslow	0.742536	0.388850
Tarone-Ware	0.499418	0.479756

Table 3.1: The estimated survival function of 1338643 particles in the LRWs is statistically similar to the analytical survival function.

3.2.2 Sample Size Determination and Evaluation

In the last chapter, two approaches used to determine the appropriate sample size in the fixed-time step Monte Carlo simulations have been proposed. One of them is based on inferential statistics [**?**], which infers and estimates the unknown population parameters from the sample statistics. Another one is simpler since it does not need any simulations.

Chebyshev's inequality

According to LLN [20], the unknown population mean first-passage time $\bar{X}$ can be estimated by sample mean $\bar{X}_N$ when N is big enough. Chebyshev's inequality [15] is a probabilistic inequality that can be applied to any probability distribution of a random variable with the finite expected value and non-zero variance. This inequality provides an upper bound to the probability that the absolute deviation of a random variable from its mean will exceed a given threshold.

In the Figure 3.3, the number of steps t in the numerical simulations have been converted into the unitless time τ by the Eq.(2.16). Thus, given a predesignated error ϵ, the required number of particles N can be determined by

[Yuge 3] No hard number

$$Pr(|X_N - \bar{X}| \geq \epsilon) = Pr(|X_N - \bar{X}_N| \geq \epsilon) \leq \frac{\sigma_N^2}{\epsilon^2} \approx \frac{2^b N^k}{\epsilon^2} = 0.01 \tag{3.1}$$

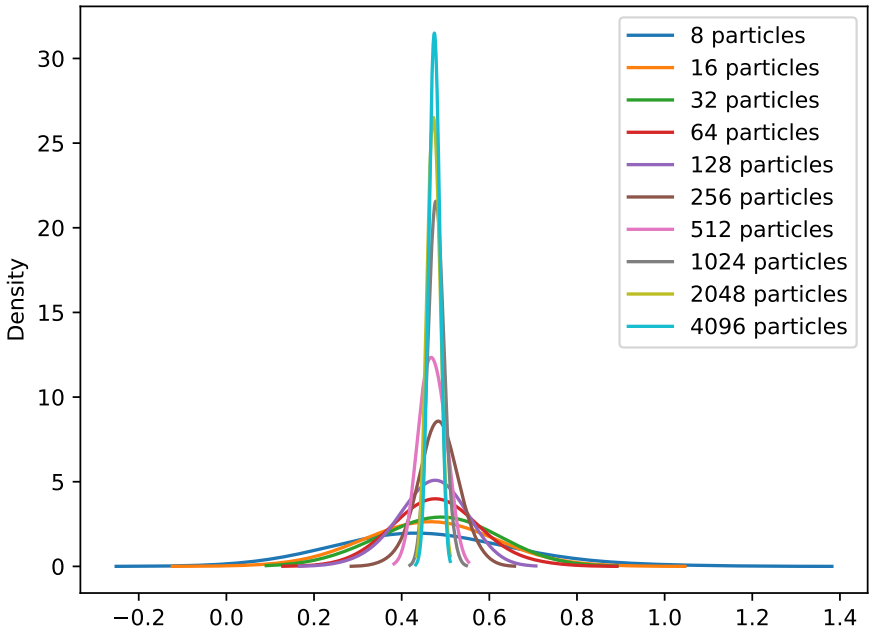
where $\epsilon = 0.01\bar{X}$.

The required number of particles in LRWs and PRWs can be estimated by Eq. 3.1, which is

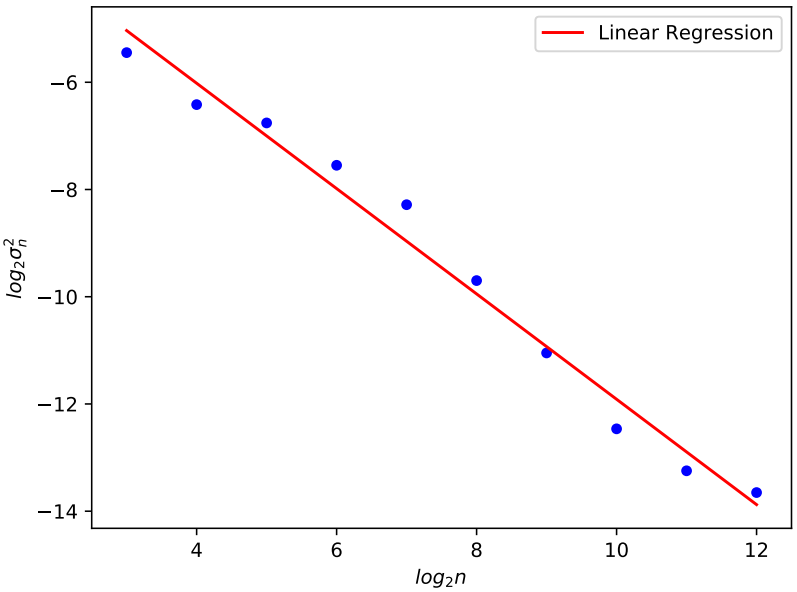
$$N \geq (\frac{0.01\epsilon^2}{2^b})^{\frac{1}{k}} \approx 1338643 \tag{3.2}$$

where $\epsilon \approx 0.004744$, $b \approx -2.088495$, and $k \approx -0.982400$. Therefore, the number of particles should be at least 1338643 to make sure that there is no more than 1% chance of X_N to be outside $[0.46865856, 0.47812641]$.

From the visualized comparison in Figure 3.4 and the results of the two-sample statistical tests in Table 3.1 and Table 3.2, the fixed-step Monte Carlo simulations' results converge to the analytical outcomes. Therefore, the integral of the solutions of heat equations can be approximated by the Monte Carlo simulations without calculating manually. As mentioned in the last chapter, the integral, $S(\tau)$, indicates the annulus' geometrical

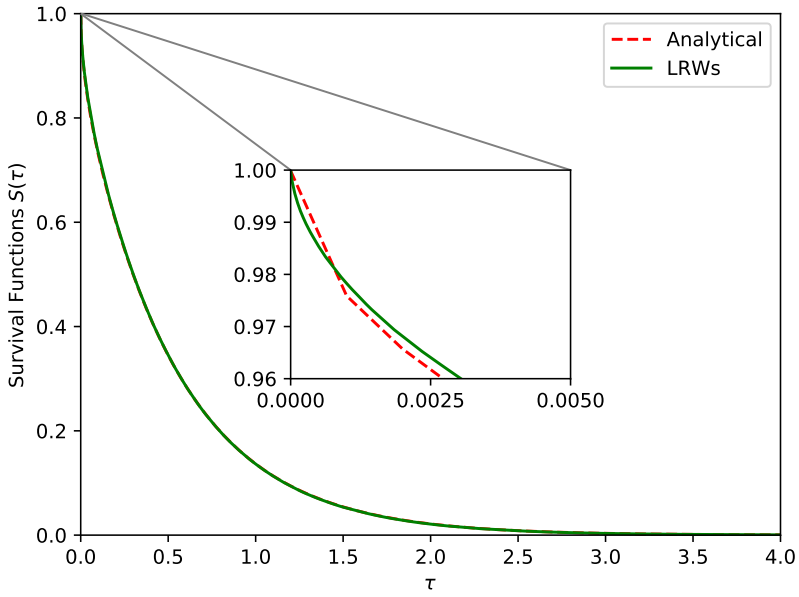


(a) Running the LRWs in the annulus with $N = 2^i$ particles and calculating the mean first-passage time X_N, where $i = 3, 4, 5, ..., 12$. For each N, replicating the simulation 50 times and recalculating the mean of the mean first-passage time $\bar{X}_N$ and the variance σ_N^2. As the sample size N increase, the distribution of the sample means X_N becomes narrower and approximately normal.

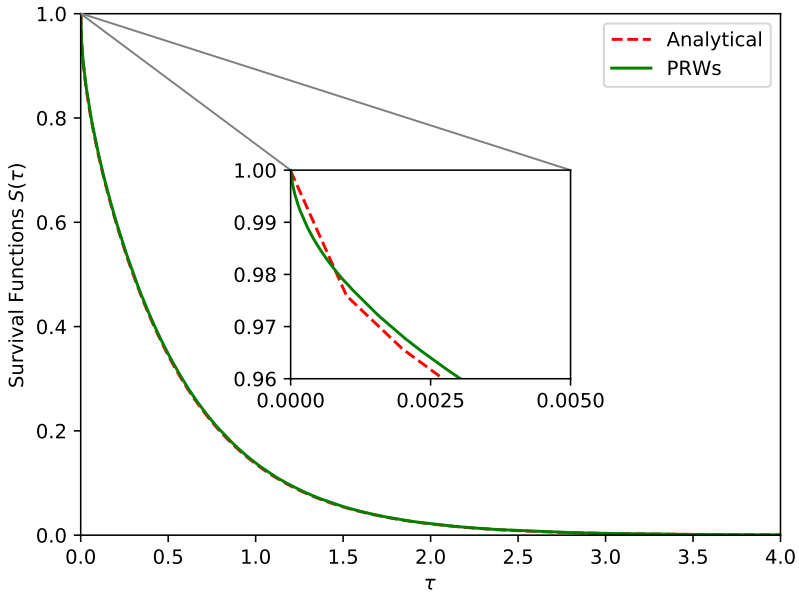


(b) A fitted linear regression model is used to explore the scaling relationship between $log_2(\sigma_N^2)$ and $log_2 N$. $log_2(\sigma_N^2) \approx b + klog_2 N$, where k and b are the estimated model parameters, slope and intercept, respectively.

Figure 3.3



(a) Survival curve for LRWs.



(b) Survival curve for PRWs.

Figure 3.4: Running PRWs and LRWs in the annulus with 1338643 particles determined by the Eq. 3.2 illustrating that the short and long time asypototic behaviors of the estimated survival functions of particles undergoing LRWs and PRWs are consistent with the analytical result in Eq Yuge

Test Methods (standard nonparametric)	Statistics	P Values
Logrank	0.039142	0.843168
Fleming-Harrington	0.083388	0.772757
Gehan-Breslow	0.083388	0.772757
Tarone-Ware	0.010582	0.918069

Table 3.2: The estimated survival function of PRWs, which has 1338643 particles with step length 0.5, is not statistically different from the analytical survival function.

information. Therefore, the fixed-time step Monte Carlo simulation can describe the shape of an object without using the rulers. However, the number of particles in the numerical simulations estimated by Eq. 3.1 is abundant, which causes a high computational cost because each random trajectories of each particle are simulated in LRWs and PRWs till they reach the inner boundary of the annulus.

Dvoretzky–Kiefer–Wolfowitz (DKW) inequality

Chebyshev's inequality can be applied to any probability distribution, but it is also weaker than other inequalities. DKW inequality is more efficient since the confidence band is generated without running any simulations. For example, let $F(\tau)$ be the true cumulative distribution function (CDF) of the first passage time, a continuous unitless random variable on the interval $[0, \infty]$. $F(\tau)$ has a relationship with $S(\tau)$, which is

$$F(\tau) = 1 - S(\tau) \tag{3.3}$$

The true CDF is known by Eq. 3.3, which can also be approximated numerically. A simple example of generating the CDF-based confidence bounds by DKW inequality is shown in Figure 3.5. The empirical distribution function $F_{256}(\tau)$ is estimated by the lifeline module in Python [18]. Thus, the interval ε contains the entire $F(\tau)$ with the probability 95% can be calculated by Eq.(2.20) Yuge

$$\varepsilon = \sqrt{\frac{\ln \frac{2}{0.05}}{2 * 256}} \approx 0.084881 \tag{3.4}$$

Moreover, the sample size estimated by DKW inequality Eq.(2.20) Yuge does not depend on the geometries or the kind of simulations because the simultaneous confidence bounds always contain the true cumulative distribution at a specific confidence level. For instance, assume the probability, that the maximum distance between $F_N(\tau)$ and $F(\tau)$ is bigger than 0.005, is smaller than 0.01, the minimum required number of particles should meet the inequality

$$Pr(\sup_{x \in \mathbb{R}} |F_N(\tau) - F(\tau)| > 0.005) \leq 2e^{-2N0.005^2} = 0.01 \tag{3.5}$$

Thus, after the transformation of Eq. 3.5, the sample size can be determined by

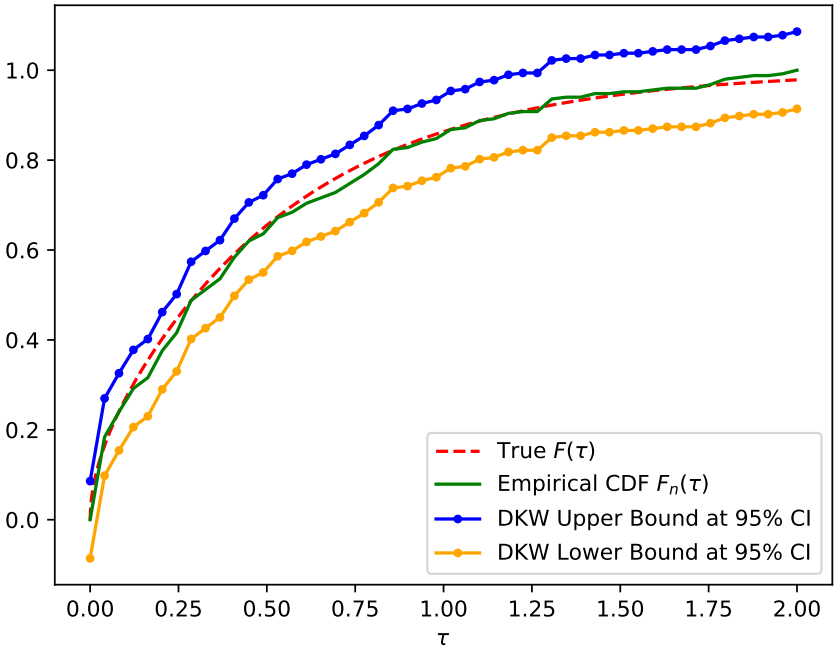


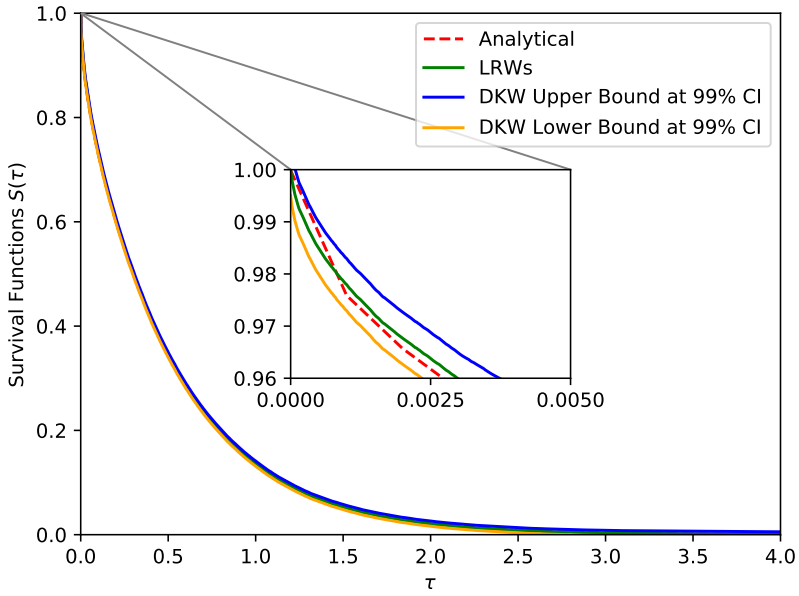
Figure 3.5: The simultaneous band around $F_{256}(\tau)$ with interval 0.084881 calculate by Eq. 3.4 encompasses the entire $F(x)$ at 95% confidence level.

$$N \geq \frac{\ln(\frac{0.01}{2})}{-2 \times 0.005^2} \approx 105966 \tag{3.6}$$

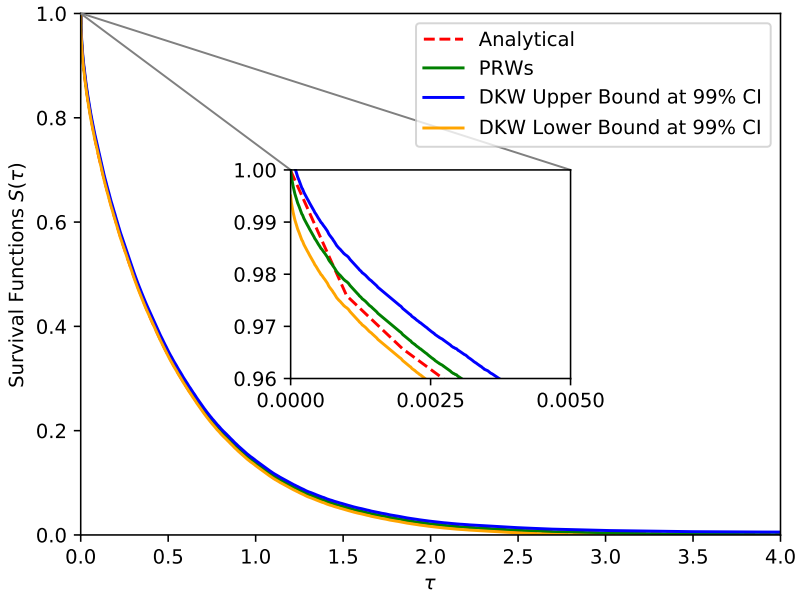
From the Figure 3.6, Table 3.3, and Table 3.4, although the sample size in the LRWs and PRWs determined by DKW inequality is about 10 times smaller than that by Chebyshev's inequality, the estimated survival functions of the numerical simulations converge to the analytical results within the amount of statistical uncertainty.

Test Methods (standard nonparametric)	Statistics	P Values
Logrank	1.532224	0.215779
Fleming-Harrington	1.630358	0.201654
Gehan-Breslow	1.630358	0.201654
Tarone-Ware	1.619530	0.203157

Table 3.3: The estimated survival function of 105966 particles in the LRWs is not statistically different to the analytical survival function.



(a) The simultaneous confidence bands, generated by Eq. 3.5, of the estimated survival function $S(t)$ for LRWs contain the whole ture analytical $S(\tau)$.



(b) The simultaneous confidence bands, generated by Eq. 3.5, of the estimated survival function $S(t)$ for PRWs contain the whole ture analytical $S(\tau)$.

Figure 3.6: PRWs and LRWs are implemented in the annulus with 105966 particles determined by the Eq. 3.6. (a) and (b) show that the DKW simultaneous confidence bands of estimated survival function with the interval 0.005 encompass the entire analytical $S(\tau)$ at 99% confidence level.

Test Methods (standard nonparametric)	Statistics	P Values
Logrank	2.645624	0.103835
Fleming-Harrington	1.473674	0.224767
Gehan-Breslow	1.473674	0.224767
Tarone-Ware	1.986810	0.158675

Table 3.4: The estimated survival function of PRWs, which has 105966 particles with step length 0.5, is statistically similar to the analytical survival function.

3.2.3 Conclusion

Instead of calculating the asymptotic expansion of the heat content manually as $\tau \to 0^+$, the total heat energy β [**?**] for time $\tau > 0$ can be approximated by the fixed-time step Monte Carlo simulations for describing the full-scale geometrical features of the annulus. Moreover, the required number of particles in the simulations determined by the DKW inequality is much smaller than the superabundant value estimated by Chebyshev's inequality.

3.3 Assumption Verification

....

3.3.1 Circle and Rectangular

...

3.3.2 Branching Structures

...

3.3.3 Conculsion

3.4 Conclusion

...

References

[1] Odd Aalen, Ornulf Borgan, and Hakon Gjessing. *Survival and Event History Analysis: A Process Point of View*. Springer Science & Business Media, 2008.

[2] Girdhar Gopal Agarwal. Statistics for surgeons–understanding survival analysis. *Indian journal of surgical oncology*, 3(3):208–214, 2012.

[3] Douglas G Altman. *Practical statistics for medical research*. CRC press, 1990.

[4] Whye-Teong Ang. *A beginner's course in boundary element methods*. Universal-Publishers, 2007.

[5] Dorothy C Attaway. The boundary element method for the diffusion equation: a feasibility study. In *Boundary Integral Methods*, pages 75–84. Springer, 1991.

[6] Louis Bachelier. Théorie de la spéculation. In *Annales scientifiques de l'École normale supérieure*, volume 17, pages 21–86, 1900.

[7] Grigory Isaakovich Barenblatt, Grigorii Isaakovĭc Barenblatt, and Barenblatt Grigory Isaakovich. *Scaling, self-similarity, and intermediate asymptotics: dimensional analysis and intermediate asymptotics*, volume 14. Cambridge University Press, 1996.

[8] Viv Bewick, Liz Cheek, and Jonathan Ball. Statistics review 12: survival analysis. *Critical care*, 8(5):389, 2004.

[9] Albert T Bharucha-Reid. *Elements of the Theory of Markov Processes and their Applications*. Courier Corporation, 2012.

[10] Garrett Birkhoff and Jack Kotik. Note on the heat equation. *Proceedings of the American Mathematical Society*, 5(1):162–167, 1954.

[11] Thomas E Booth. Exact monte carlo solution of elliptic partial differential equations. *Journal of Computational Physics*, 39(2):396–404, 1981.

[12] ROBERT Brown. Microscopical observations. *Philos. Mag*, 4:161–173, 1828.

[13] Francesco Paolo Cantelli. Sui confini della probabilita. In *Atti del Congresso Internazionale dei Matematici: Bologna del 3 al 10 de settembre di 1928*, pages 47–60, 1929.

[14] L Douglas Case, Gretchen Kimmick, Electra D Paskett, Kurt Lohman, and Robert Tucker. Interpreting measures of treatment effect in cancer clinical trials. *The oncologist*, 7(3):181–187, 2002.

[15] Pafnutii Lvovich Chebyshev. Des valeurs moyennes. *J. Math. Pures Appl*, 12(2):177–184, 1867.

[16] John Crank. *The mathematics of diffusion*. Oxford university press, 1979.

[17] Ruvie Lou Maria Custodio Martinez. Diagnostics for choosing between log-rank and wilcoxon tests. 2007.

[18] Cameron Davidson-Pilon. lifelines: survival analysis in python. *Journal of Open Source Software*, 4(40):1317, 2019.

[19] Alessandro De Gregorio and Enzo Orsingher. Flying randomly in rd with dirichlet displacements. *Stochastic processes and their applications*, 122(2):676–713, 2012.

[20] Frederik Michel Dekking, Cornelis Kraaikamp, Hendrik Paul Lopuhaä, and Ludolf Erwin Meester. *A Modern Introduction to Probability and Statistics: Understanding why and how*. Springer Science & Business Media, 2005.

[21] *NIST Digital Library of Mathematical Functions.* http://dlmf.nist.gov/, Release 1.0.26 of 2020-03-15. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds.

[22] Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pages 642–669, 1956.

[23] Albert Einstein. On the theory of the brownian movement. *Ann. Phys*, 19(4):371–381, 1906.

[24] Albert Einstein et al. On the electrodynamics of moving bodies. *Annalen der physik*, 17(10):891–921, 1905.

[25] İ Etikan, S Abubakar, and R Alkassim. The kaplan-meier estimate in survival analysis. *Biom Biostatistics Int J*, 5(2):00128, 2017.

[26] I Etikan, K Bukirova, and M Yuvali. Choosing statistical tests for survival analysis. *Biom. Biostat. Int. J*, 7:477–481, 2018.

[27] Robert Eymard, Thierry Gallouët, and Raphaèle Herbin. Finite volume methods. *Handbook of numerical analysis*, 7:713–1018, 2000.

[28] Temple H Fay and P Hendrik Kloppers. The gibbs' phenomenon for fourier-bessel series. *International Journal of Mathematical Education in Science and Technology*, 34(2):199–217, 2003.

[29] Jozef Gembarovic. Using monte carlo simulation for solving heat conduction problems, 03 2017.

[30] Manish Kumar Goel, Pardeep Khanna, and Jugal Kishore. Understanding survival analysis: Kaplan-meier estimate. *International journal of Ayurveda research*, 1(4):274, 2010.

[31] Denis Grebenkov. Efficient monte carlo methods for simulating diffusion-reaction processes in complex systems. In *First-Passage Phenomena and Their Applications*, pages 571–595. World Scientific, 2014.

[32] M Greenwood. The natural duration of cancer. london: His majesty's stationery office; 1926. *Reports on public health and medical subjects*, (33).

[33] Christian Grossmann, Hans-Görg Roos, and Martin Stynes. *Numerical treatment of partial differential equations*, volume 154. Springer, 2007.

[34] Abdolhossein Haji-Sheikh. Application of monte carlo methods to thermal conduction problems. 1967.

[35] Godfrey Harold Hardy and Marcel Riesz. *The general theory of Dirichlet's series.* Courier Corporation, 2013.

[36] David P Harrington and Thomas R Fleming. A class of rank test procedures for censored survival data. *Biometrika*, 69(3):553–566, 1982.

[37] Joe D Hoffman and Steven Frankel. *Numerical methods for engineers and scientists.* CRC press, 2018.

[38] David W Hosmer Jr, Stanley Lemeshow, and Susanne May. *Applied survival analysis: regression modeling of time-to-event data*, volume 618. John Wiley & Sons, 2011.

[39] Barry D Hughes. Random walks and random environments. *Bulletin of the American Mathematical Society*, 35(4):347–349, 1998.

[40] Kiyosi Itô, P Henry Jr, et al. *Diffusion processes and their sample paths.* Springer Science & Business Media, 2012.

[41] Mark Kac. Random walk and the theory of brownian motion. *The American Mathematical Monthly*, 54(7P1):369–391, 1947.

[42] John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, 2011.

[43] Pavol Kalinay, Ladislav Šamaj, and IGOR TRAVĚNEC. Survival probability (heat content) and the lowest eigenvalue of dirichlet laplacian. *International Journal of Modern Physics B*, 25(15):1993–2007, 2011.

[44] Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

[45] Pinar Gunel Karadeniz, Ilker Ercan, et al. Examining tests for comparing survival curves with right censored data. *Stat Transit*, 18(2):311–28, 2017.

[46] Samuel Karlin. *A first course in stochastic processes*. Academic press, 2014.

[47] John T Katsikadelis. *Boundary elements: theory and applications*. Elsevier, 2002.

[48] Gilbert W King. Monte-carlo method for solving diffusion problems. *Industrial & Engineering Chemistry*, 43(11):2475–2478, 1951.

[49] David G Kleinbaum and Mitchel Klein. Competing risks survival analysis. *Survival Analysis: A self-learning text*, pages 391–461, 2005.

[50] Dirk P Kroese, Tim Brereton, Thomas Taimre, and Zdravko I Botev. Why the monte carlo method is so important today. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6(6):386–392, 2014.

[51] AA Kronberg. Solution of two boundary value problems by the monte carlo method. *USSR Computational Mathematics and Mathematical Physics*, 16(1):153–161, 1976.

[52] Gregory F Lawler. *Random walk and the heat equation*, volume 55. American Mathematical Soc., 2010.

[53] E Leton and P Zuluaga. Equivalence between score and weighted tests for survival curves. *Communications in Statistics-Theory and Methods*, 30(4):591–608, 2001.

[54] Daryl L Logan. *A first course in the finite element method*. Cengage Learning, 2011.

[55] Jaume Masoliver, Josep M Porra, and George H Weiss. Some two and three-dimensional persistent random walks. *Physica A: Statistical Mechanics and its Applications*, 193(3-4):469–482, 1993.

[56] Mervin E Muller et al. Some continuous monte carlo methods for the dirichlet problem. *The Annals of Mathematical Statistics*, 27(3):569–589, 1956.

[57] Muhammad Mushtaq, Nawazish Ali Shah, and Ghulam Muhammad. Advantages and disadvantages of boundary element methods for compressible fluid flow problems. *Journal of American Science*, 6(1):162–165, 2010.

[58] Karl Pearson. The problem of the random walk. *Nature*, 72(1867):342–342, 1905.

[59] Richard H Pletcher, John C Tannehill, and Dale Anderson. *Computational fluid mechanics and heat transfer*. CRC press, 2012.

[60] Lord Rayleigh. The problem of the random walk. *Nature*, 72(1866):318, 1905.

[61] Junuthula Narasimha Reddy. An introduction to the finite element method. *New York*, 27, 1993.

[62] Sidney Redner. *A guide to first-passage processes*. Cambridge University Press, 2001.

[63] Reuven Y Rubinstein and Dirk P Kroese. *Simulation and the Monte Carlo method*, volume 10. John Wiley & Sons, 2016.

[64] Karl K Sabelfeld and Nikolai A Simonov. *Random walks on boundary for solving PDEs*. Walter de Gruyter, 2013.

[65] MNO Sadiku, CM Akujuobi, and SM Musa. Monte carlo analysis of time-dependent problems. In *Proceedings of the IEEE SoutheastCon 2006*, pages 7–10. IEEE, 2006.

[66] S Sawyer. The greenwood and exponential greenwood confidence intervals in survival analysis. *Applied survival analysis: regression modeling of time to event data*, pages 1–14, 2003.

[67] Ritesh Singh and Keshab Mukhopadhyay. Survival analysis in clinical trials: Basics and must know areas. *Perspectives in clinical research*, 2(4):145, 2011.

[68] M von Smoluchowski. Drei vortrage uber diffusion, brownsche bewegung und koagulation von kolloidteilchen. *ZPhy*, 17:557–585, 1916.

[69] Wolfgang Stadje. The exact probability distribution of a two-dimensional random walk. *Journal of statistical physics*, 46(1-2):207–216, 1987.

[70] M Van den Berg. Heat content and brownian motion for some regions with a fractal boundary. *Probability theory and related fields*, 100(4):439–456, 1994.

[71] Nicolaas Godfried Van Kampen. *Stochastic processes in physics and chemistry*, volume 1. Elsevier, 1992.

[72] SR Srinivasa Varadhan, Pl Muthuramalingam, and Tara R Nanda. *Lectures on diffusion problems and partial differential equations*. Springer Berlin/New York, 1980.

[73] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake Vand erPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1. 0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[74] DF Vysochanskij and Yu I Petunin. Justification of the 3σ rule for unimodal distributions. *Theory of Probability and Mathematical Statistics*, 21(25-36), 1980.

[75] George Neville Watson. *A treatise on the theory of Bessel functions*. Cambridge university press, 1995.

[76] Miloš Zlámal. On the finite element method. *Numerische Mathematik*, 12(5):394–409, 1968.