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 Heat Content in Annulus

Since the heat (diffusion) equation defined in an annulus possesses explicit solutions, the analytical expressions of heat content, derived by the integration over the whole domain, are accessible. In this section, the preliminary step is to solve the initial-boundary value problem (IBVP) defined in the annulus. The numerical methods will then be used to approximate that heat content, which helps validate the research methodology in the next chapter.

2.1.1 Analytical Results

The heat equation Eq.(2.1) defined in the polar coordinate system describes the heat distribution varying in time and Ω, as shown in Figure 2.1. The solution $u(r, \theta, t)$ implies the temperature over spatial positions and time. In this subsection, the primary purpose is to calculate the total amount of heat contained in Ω, heated at some nonzero temperature, in which $\partial\Omega_1$ is cooled to the zero temperature and $\partial\Omega_2$ is perfectly insulated.

$$u_t = D\big(u_{rr} + \frac{1}{r}u_r + \frac{1}{r^2}u_{\theta\theta}\big) \tag{2.1}$$

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

$$u' = 0 \quad on\ \partial\Omega_2 \tag{2.3}$$

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

where D is the diffusion coeffiecnt, r is radial coordinate, θ is the angular coordinate, and t is the time.

$u(r, \theta, t)$ is a probability density function, which gives the value of heat particles at (r, θ) at time t. Eq.(2.2) and Eq.(2.3) indicate the Dirichlet boundary condition on $\partial\Omega_1$ and the Neumann boundary condition on $\partial\Omega_2$, respectively. In other words, the heat particles will be absorbed when they encounter with $\partial\Omega_1$ and be reflecting when they reach $\partial\Omega_2$. Eq.(2.4) states that the heat particles distribute uniformly over the whole domain at time $t = 0$, where $|\Omega|$ equals the total area of the annulus.

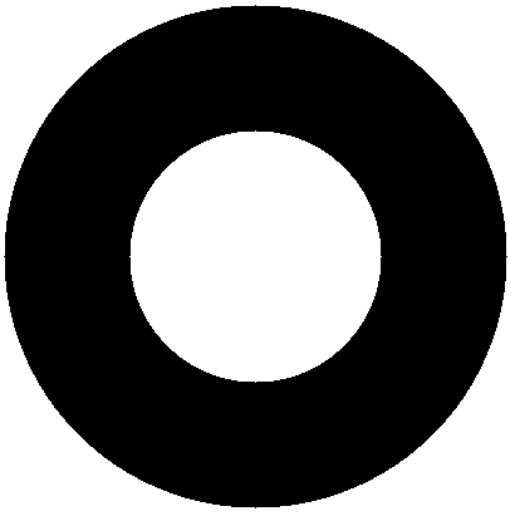


Figure 2.1: Assume the annulus Ω is a homogeneous and isotropic medium and has the inner boundary $\partial\Omega_1$ with radius a and outer boundary $\partial\Omega_2$ with radius b.

Solving Heat (Diffusion) Equation

Before solving the heat equation, it is convenient and efficient to generate 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 interestt [**?**].

Define $\mu = b/a$ as the dimensionless radius ratio, $\tau = \frac{Dt}{a^2}$ as the dimensionless time, and $\hat{r} = \frac{r}{a}$ as the unitless radius. Substitute the dimensionless variables and rewrite Eq.(2.1) as

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

With the uniform initial condition,

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

With the boundary conditions

$$u\big(1, \theta, \tau\big) = 0$$

$$u'\big(\mu, \theta, \tau\big) = 0$$

After implementing the separation of variables method [**?**], the solutions of Eq.(2.5) are

$$u(\hat{r},\theta,\tau)=\sum_{n=1}^{\infty}c_{\tilde{0},n}\Big\{J_0\big(\sqrt{\lambda_{0,n}}\big)Y_0\big(\sqrt{\lambda_{0,n}}\hat{r}\big)-Y_0\big(\sqrt{\lambda_{0,n}}\big)J_0\big(\sqrt{\lambda_{0,n}}\hat{r}\big)\Big\}e^{-\lambda_{0,n}\tau} \tag{2.6}$$

where

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

Eigenvalues $\lambda_{0,n}$ ($n\in\mathbb{N}_+$) is nth positive root of the cross-product of Bessel functions [**?**]

$$F_0(\lambda)=J_0(\sqrt{\lambda})Y_0'(\sqrt{\lambda}\mu)-J_0'(\sqrt{\lambda}\mu)Y_0(\sqrt{\lambda}) \tag{2.7}$$

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 in some mathematical literatures [**?**] [**?**] [**?**]. Survival probability $S(\tau)$ is proportional to $Q_\Omega(\tau)$ [**?**], which gives the probability of the particles remain diffusing in the domain Ω at time $\tau>0$ [**?**]. Survival probability can be expressed by

$$\begin{aligned}S(\tau)&=\int_0^{2\pi}d\theta\int_1^{\mu}\hat{r}d\hat{r}u(\hat{r},\theta,\tau)\\&=\sum_{n=1}^{\infty}\frac{4}{\mu^2-1}\frac{1}{\lambda_{0,n}\left\{\left[\frac{J_0(\sqrt{\lambda_{0,n}})}{J_0'(\mu\sqrt{\lambda_{0,n}})}\right]^2-1\right\}}e^{-\lambda_{0,n}\tau}\end{aligned} \tag{2.8}$$

From Eq.(2.8), we can find 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 [**?**]. One of the essential first-passage-related quantities is the first-passage probability, which is a probability of diffusing particle or a random-walk hitting a specified site or a set of sites at a specified time for the first time [**?**]. All the first-passage characteristics can be expressed in terms of first-passage probability itself. For example, the survival probability at time τ calculated in the last subsection has

$$F(\tau) = -\frac{\partial S(\tau)}{\partial \tau} \tag{2.9}$$

where $F(\tau)$ is the first-passage probability to the target boundary at time τ. By the definition, the nth moment of the exit time [**?**] is

$$\begin{aligned} \langle \tau^n \rangle &= \int_0^\infty \tau^n F(\tau) d\tau \\ &= -\int_0^\infty \tau^n \frac{\partial S(\tau)}{\partial \tau} d\tau \\ &= -\tau^n S(\tau)|_0^\infty + n \int_0^\infty \tau^{n-1} S(\tau) d\tau \end{aligned} \tag{2.10}$$

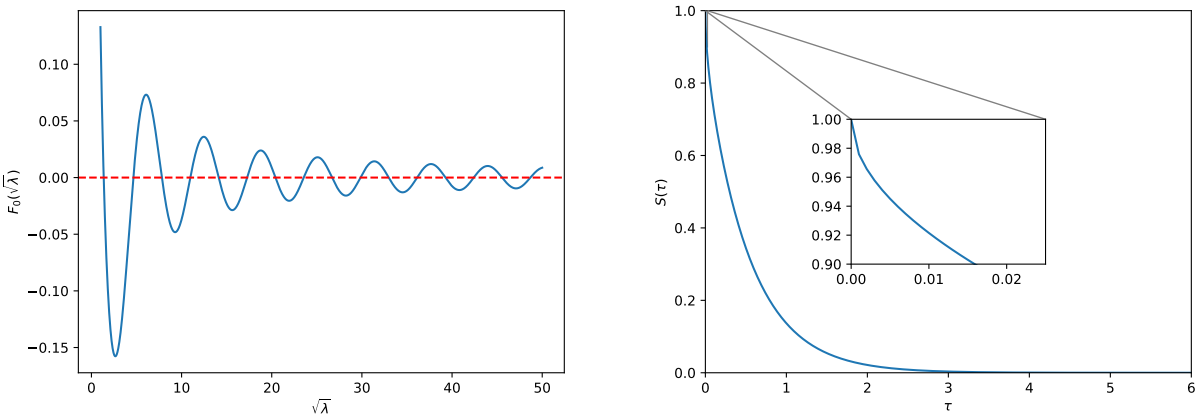
Substitue $n = 1$ in Eq.(2.10), the mean-first passage time $\langle \tau \rangle$, also called the average first-passage time, of a group of particles implies an overall property of the system is

$$\begin{aligned} \langle \tau \rangle &= \int_0^\infty \tau dS \\ &= \sum_{n=1}^{\infty} \frac{4}{\mu^2 - 1} \frac{1}{\lambda_{0,n}^2 \left\{ \left[\frac{J_0(\sqrt{\lambda_{0,n}})}{J_0'(\mu\sqrt{\lambda_{0,n}})} \right]^2 - 1 \right\}} \end{aligned} \tag{2.11}$$

2.1.2 Numerical Analysis

Numerical analysis is an area of mathematics and computer science that creates, analyzes, and implements algorithms for approximating numerical solutions to problems involving continuous variables [**?**]. For simplicity, a specific kind of annulus is considered, which has the radius ratio $\mu = 2$.

The preliminary step of the evaluation of $S(\tau)$, a general Dirichlet series, and $\langle\tau\rangle$ is to compute the monotonically increasing $\lambda_{0,n}$ of Eq.(2.7). The nth positive zero $\lambda_{0,n}$, as $n \to \infty$, can be bracketed in an interval $((n-1)\pi, (n+1)\pi)$ [**?**]. Bisection method [**?**], a well-known and most reliable root finding method. It can be used to close in on the root by successivly halving the interval until it becomes sufficiently small.



(a) **(b)**

Figure 2.2: (a) It is straightforward to evaluate the cross-product of Bessel fucntions by SciPy library [**?**]. (b) The asymptotic behaviors of survival probability $S(\tau)$ are approximated by the numerical method with the first 1000 eigenvalues. $S(\tau)$ monotonously descrese 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 0.47339248.

2.2 Monte Carlo Simulations

In this section, several widely used numerical methods for solving partial differential equations (PDEs), such as the diffusion equation, and their limitations in practice are introduced. From the deterministic perspective, one of the probabilistic algorithms, Monte-Carlo (CM) methods, is proposed. However, discretization errors also affect the accuracy of the solutions. Finally, based on the theories of two random walk models, Pearson's random walks (PRWs) and lattice random walks (LRWs), two fixed-time step Monte Carlo simulations are designed. Those simulations can approximate the integration of the heat equations' solutions, named survival functions, which describes the geometrical properties of the shape.

2.2.1 Background

In the last section in this chapter, the exact analytical solutions of the heat equation defined in the annulus with the initial and boundary conditions have been derived. And then, the analytical survival probability can be calculated by integration on the annulus. Nevertheless, either irregular geometries or discontinuities cause the complication in solving the PDEs in practice, so the explicit algebraic solutions are close to non-existed. Therefore, numerical methods and computer simulations are more useful and accessible in solving differential equations than calculating pure analytical solutions.

Deterministic Numerical Methods

The techniques for solving initial-boundary value problems (IBVPs) based on numerical approximations have existed for a long time. The finite-difference methods (FDM) are frequently used and easily implemented in solving the differential equations by converting them into a system of algebraically solvable equations [**?**]. The basic idea is to replace the derivatives in the equations by difference quotients. For example, the FTCS (Forward Time Centered Space) scheme [**?**] aims to approximate the numerical solutions of the heat equation and other similar parabolic PDEs. FTCS discretizes the Laplace operator in space and the time derivative and implements boundary conditions on the staggered grid to represent the original continuous problem, but it is numerically stable if and only if it satisfies a specified condition [**?**]. Generally, there are two classes of errors appearing in FDM called round-off error and truncation error. The former one results from the loss of precision due to the computer rounding of decimal quantities. The latter one is caused by discarding the remainders in the difference approximation and highly depends on the time and space step. In other words, the smaller the step size, the longer the simulation duration and higher data quality [**?**].

Compared with FDM, the finite element method (FEM) is particularly solving PDEs with relatively higher quality when the problems are defined in two or three space dimensions. FEM divides the physical system's complicated geometries and boundaries into a certain amount of smaller and simpler subdomains [**?**] (eg. lattice, triangle, curvilinear polygons, etc.). Every subdomain is locally approximated by a finite set of element equations (piecewise shape functions) assembled into a larger system of equations for modelling the entire problem finally [**?**]. The goal of FEM is to approximate a stable solution by minimizing the associated error function with the variational calculus [**?**]. The significant advantages of FEM are representing the complex geometries accurately, exploring the local characters of the approximation, and expressing the solutions in a unified formulation [**?**]. However, FEM needs an amount of mathematical skill that requires human involvement, such as converting the original equation into the equivalent weak formulation, choosing and changing the variational formulation and discretization strategy in a particular problem, etc. In this thesis, the heat equations defined in 2-dimensional images with millions of pixels, the extremely complex root systems and various boundary conditions, which makes it time-consuming and challenging to trace and identify the problems' geometries, label the nodes, and generate the coordinates and connectivities among

the nodes in the preprocessing stage of FEM.

In contrast, the finite volume method (FVM) converts the 2-dimensional diffusion equations into a linear equations system, which can be solved by the direct method [**?**]. However, the accuracy of FVM is related to the integration with respect to time and space. Unlike the domain-type methods (e.g. FDM, FEM, FVM, etc.), the boundary element method (BEM) is an alternative numerical computational technique for solving PDEs formulated as an integral equation based on the boundary. 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 [**?**] since it only discretizes the boundary and fits the boundary values into the integral equation [**?**].

However, all of the mentioned numerical methods have an intrinsically similar feature - mesh discretization in the time and space dimension. They need to deal with the problem in defining the extremely complicated boundary of the root systems. Moreover, if the mesh becomes finer and the internal points number becomes larger, the discrete problem's solutions will converge to the original IBVPs with a higher level of accuracy, but the computational time will be longer. Furthermore, after applying the numerical methods, it will sometimes take a much longer time to decide whether the solution is right and adjust the schemes by understanding the systems qualitatively. More importantly, the extra efforts are demanded based on the approximated solutions of heat equations since the aimed numerical approximation is the survival function.

Monte Carlo Techniques

Monte Carlo (MC) methods, 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 [**?**]. Compare with the deterministic numerical methods, applying Monte Carlo methods for solving PDEs is grid-free on the domain, boundary, and the boundary conditions of the problem [**?**].

From the deterministic perspective, sloving PDEs by Monte Carlo procedure is based on the classical probabilistic representations, e.g. Feynman-Kac representations [**?**], in the form of Wiener or diffusion path integral [**?**]. For example, when solving the d-dimensional heat equation, the Brownian motion generated by the second-order differential operator is simulated using the numerical methods for solving the stochastic differential equation [**?**]. However, the discretization step should be small enough to achieve the desired accuracy, which will result in the long simulation.

Another straightforward approach for solving PDEs by Monte Carlo method based on the probabilistic interpretation of integral equations equivalent to the original continuous problem, whose solutions can be represented as expectation of some functions of the trajectories of a stochastic process [**?**][**?**][**?**]. Those trajectories generated by Monte Carlo techniques approximate the expectations and the solutions. For instance, the Markov process is proposed for estimating the expectations by simulating the successive positions of the trajectory of the process at fixed instants with step $\triangle t$ [**?**][**?**].

2.2.2 Models for Continious Diffuison Process

Borwnian Motion and Random Walks

The heat equation describes the temperature distribution of a certain homogeneous and isotropic domain that does not exist any heat sources [**?**]. In the domain, the heat spreads randomly in all directions at some rate, so it is easier to understand diffusion (heat) equations by considering the heat as the individual random particle [**?**]. The spreading of the heat 'particles' is a diffusion process defined in continuous time with a continuous state space and continuous sample paths [**?**]. Brownian motion [**?**], also called the Wiener process, is a case of a diffusion process with the Markov properties: future states depend only on the present states [**?**]. From a probabilistic perspective, the probability density of Brownian particles satisfies the heat (diffusion) equation [**?**][**?**].

Brownian motion [**?**], the irregular motion of individual pollen particles, has been existed for a long time before the random-walk theory. At the beginning of the twentieth century, the term, random walk, was initially proposed by Karl Pearson [**?**]. He described a simple example of 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 [**?**] stated that Pearson's question, the distributions of mosquitos after many steps have been taken, is the same as superpositioning the sound vibrations with unit amplitude and arbitrary phase [**?**].

At almost the same time, Louis Bachelier [**?**] developed a model of the financial time series based on the random walk and explored the connection between discrete random walks and continuous diffusion (heat) equation. During the development of random walk theory, many important fields, such as random processes, random noise, spectral analysis, and stochastic equations, were developed by some physicists [**?**] [**?**] [**?**]. The continuous Brownian motion can be considered as the limit of discrete random walk as the time and space increments go to zero [**?**].

Theory of Lattice Random Walks (LRWs)

Let us consider the heat particles perform the simple random walk on the d-dimensional integer grid $\mathbb{Z}^d$. It is a discrete-space and discrete-time symmetric hopping process [**?**] on the lattice. At each time step, a random walker moves to one of its $2d$ nearest neighbours with probability $\frac{1}{2d}$. If $d \leq 2$, the random walk is recurrent, in which the particle will returns to its origin infinitely often with the probability 1. If $d \geq 3$, the random walk is transient in which the particle will return to its origin only finitely often with probability 1 [**?**].

For simplicity, we start from considering 2-dimensional lattice random walks (LRWs) [**?**]. Let $\triangle l$ be the distance between two sites and δ be the time step. Define $p(x, t|x_S)$ as the conditional probability of a particle to be in position x at time t starting from x_S at time $t = 0$. The temperature at a site is given by the amonunt of heat going in from nerighboring site. Thus, $p(x, t + \delta|x_S)$ is defined as probability of a particle to be in position x at time $t + \delta$ starting from x_S at time $t = 0$,

$$p(x,t+\delta|x_S) = \frac{p(x-\triangle l e_x,t|x_S)+p(x+\triangle l e_x,t|x_S)+p(x-\triangle l e_y,t|x_S)+p(x+\triangle l e_y,t|x_S)}{4} \tag{2.12}$$

where e_x and e_y are unit vectors of $x-$ axis and $y-$ axis, respectively.

When $\delta \to 0$, $\triangle l \to 0$, and $\delta \sim (\triangle l)^2$,

$$p(x,t|x_S)+\delta p(x,t|x_S) = p(x,t|x_S)+\frac{(\triangle l)^2}{4}(\frac{\partial^2 p(x,t|x_S)}{\partial x^2}+\frac{\partial^2 p(x,t|x_S)}{\partial y^2}) \tag{2.13}$$

Finally, the 2-dimensional heat equation is

$$\frac{\partial p(x,t|x_S)}{\partial t} = \frac{(\triangle l)^2}{4\delta}(\frac{\partial^2 p(x,t|x_S)}{\partial x^2}+\frac{\partial^2 p(x,t|x_S)}{\partial y^2}) \tag{2.14}$$

where $D = \frac{(\triangle l)^2}{\delta}$ is the diffusion coefficent.

Theory of Pearson's Random Walks (PRWs)

Based on Pearson's problem and Rayleigh's answer, Stadje [**?**] and Masoliver et al. [**?**] considered a two-dimensional continuous-time and continuous-space random walk, defined as Pearson's random walks (PRWs) in this thesis, moving with constant speed and with random directions distributed uniformly in $[0, 2\pi)$. Moreover, in PRWs, the lengths of the straight-line paths and the turn angles are stochastically independent. If the mean step length approaches zero and the time is big enough, the behaviours of particles in PRWs weakly converges to the Wiener Process [**?**], which satisfies the traditional heat equation.

2.2.3 Output Analysis

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:

$$\triangle l = 2\sqrt{D\delta} \tag{2.15}$$

where D is the diffusion coefficent.

Eq.(2.15) 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,

$$\tau = t\delta = \frac{(\triangle l)^2 t}{4D} \tag{2.16}$$

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 $\frac{1}{1000}$.

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 [**?**], such as death, pregnancy, job loss, etc. Also, the analysis of the group of survival data is called survival analysis [**?**]. In the survival analysis, three kinds of situations will affect the subjects' survival time [**?**]. 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 [**?**] instead of ignoring them and decreasing sample size.

In clinical trials or community trials, Kaplan-Meier Estimator [**?**], 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 [**?**] and for generating the corresponding survival curve. It also works well with the mentioned three difficult situations. With various assumptions [**?**] [**?**], 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 [**?**].

The output of random walk models is particles' wandering time, also defined as the number of steps, before encountering the absorbing boundary. Particle's behaviours are affected by the boundary conditions, and each of them carries some partial geometrical information of the edges. In this thesis, the numerical survival functions, which can be estimated by [**?**]

$$\widehat{S}(t) = \prod_{i:t_i \leq t} (1 - \frac{d_i}{n_i}) \tag{2.17}$$

where t_i is a time when at least one particle reaches the absorbing boundary, d_i the number of particles encountering with the target boundary at time t_i, and n_i is the number of particles which have not yet stopped moving up to time t_i.

2.2.4 Sample Size Determination

Based on the law of large number (LLN) [**?**], 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) [**?**] 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 [**?**], Cantelli's inequality [**?**], Vysochanskij–Petunin inequality [**?**], 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 [**?**], is proposed for the sample size determination 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.(2.18), bounds the probability that the random function $F_N(x)$ differs from the true $F(x)$ by more than a given constant ε [**?**].

$$Pr(\sup_{x\in\mathbb{R}} |F_N(x) - F(x)| > \varepsilon) \leq 2e^{-2N\varepsilon^2} \quad for\ every\ \varepsilon > 0 \tag{2.18}$$

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

$$F_N(x) - \varepsilon \leq F(x) \leq F_N(x) + \varepsilon \tag{2.19}$$

On the other hand, assume the simultaneous band produced by Eq.(2.19) containing the $F(x)$ at a given confidence level $1 - \alpha$, the interval ε can be calculated by

$$\varepsilon = \sqrt{\frac{\ln \frac{2}{\alpha}}{2N}} \tag{2.20}$$

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.(2.20).

2.3 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 [**?**]. 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 [**?**]. 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 expressed by limiting the number of events per unit time divided by the number at risk [**?**]. The hazard ratio is an estimate of the hazard rate in one group relative to that in another group [**?**]. 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 [**?**].

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 [**?**]. Also, under the constant hazard ratio assumption, the Wilcoxon tests might be more reliable than the log-rank tests [**?**]. 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 [**?**]. 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 [**?**]. However, Gehan's generalized Wilcohon test is still robust when the censoring rates are low, and the censoring distributions of groups are equal [**?**].

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 [**?**]. It pays more attention to the failures happening somewhere

in the middle of study [**?**]. Moreover, there is no limitation of the number of groups when the Tarone-Ware test is applied [**?**]. 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 [**?**].

Fixed-time Step Monte Carlo Simulations on Artificial Images

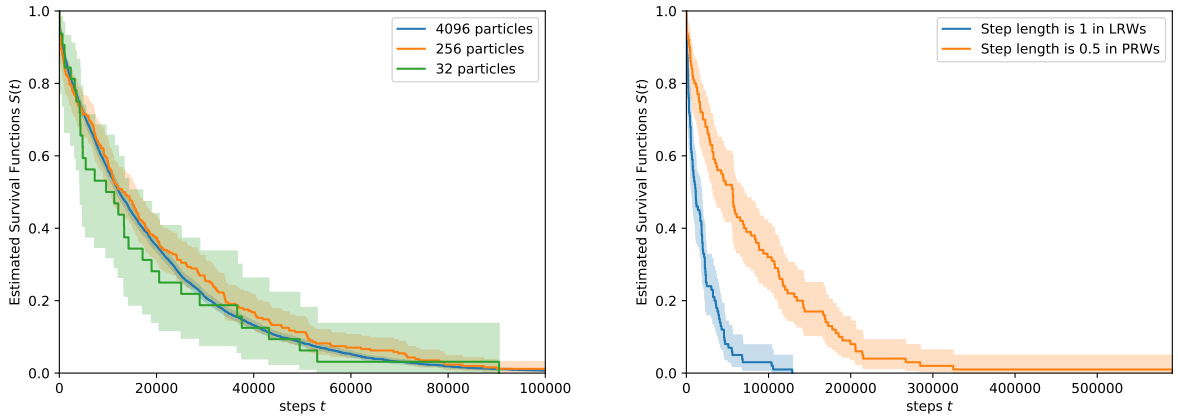
3.1 Methodology Validation

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.

3.1.1 Statistical Fluctuation Analysis

Fixed-time step Monte Carlo simulations, LRWs and PRWs, are the virtual representations of the original statistical problem defined in the continuous-time and continuous-space with numerous inputs and discrete-time trajectories sampled randomly over and over again, which results in the statistical fluctuation.

The first kind of error stems from the sampling. As shown in Figure 3.1(a), 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 3.1(b).



(a) **(b)**

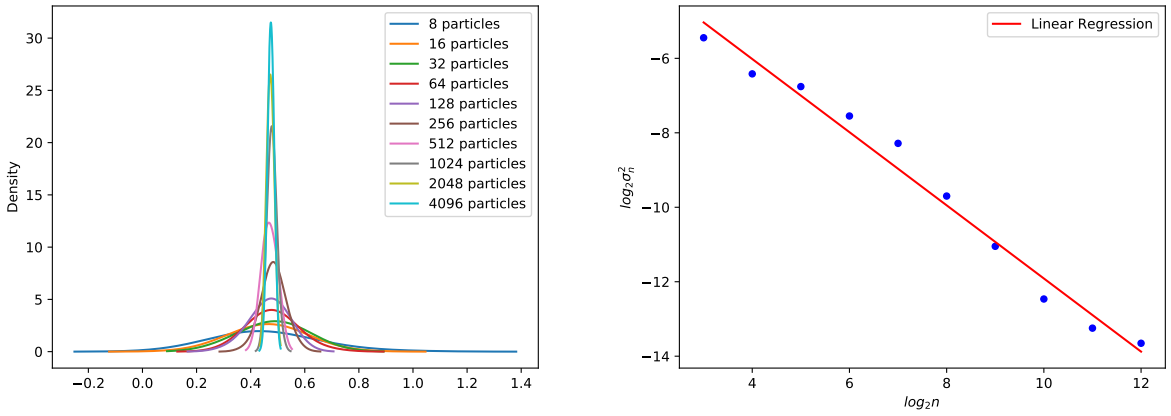
Figure 3.1: (a) As the number of particles increasing, the uncertainty of the LRWs simulation are lower since the confidence band of the estimated survival function becomes narrower. (b) When run LRWs and PRWs in the annulus with 100 particles, the finer discretization step results in the longer simulation time.

3.1.2 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 [**?**], 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 [**?**] 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.



(a) **(b)**

Figure 3.2: (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 $log_2 N$. $log_2(\sigma_N^2) \approx b + k log_2 N$, where k and b are the estimated model parameters, slope and intercept, respectively.

In the Figure 3.2, 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

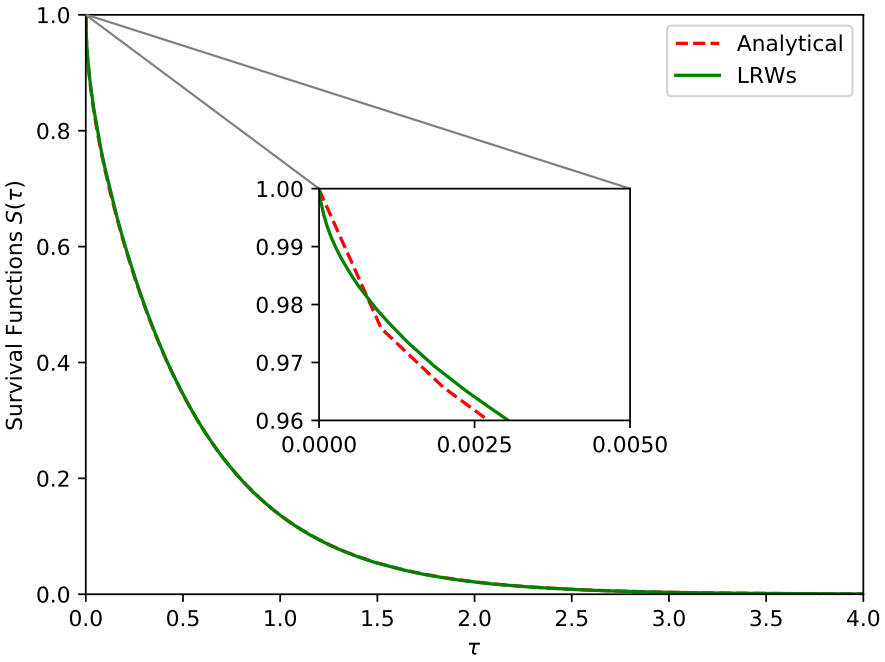
$$Pr(|X_N - \bar{X}| \geq \epsilon) = Pr(|X_N - \bar{X}_N| \geq \epsilon) \leq \frac{\sigma_N^2}{\epsilon^2} \approx \frac{2^b N^k}{\epsilon^2} = 0.01 \tag{3.1}$$

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

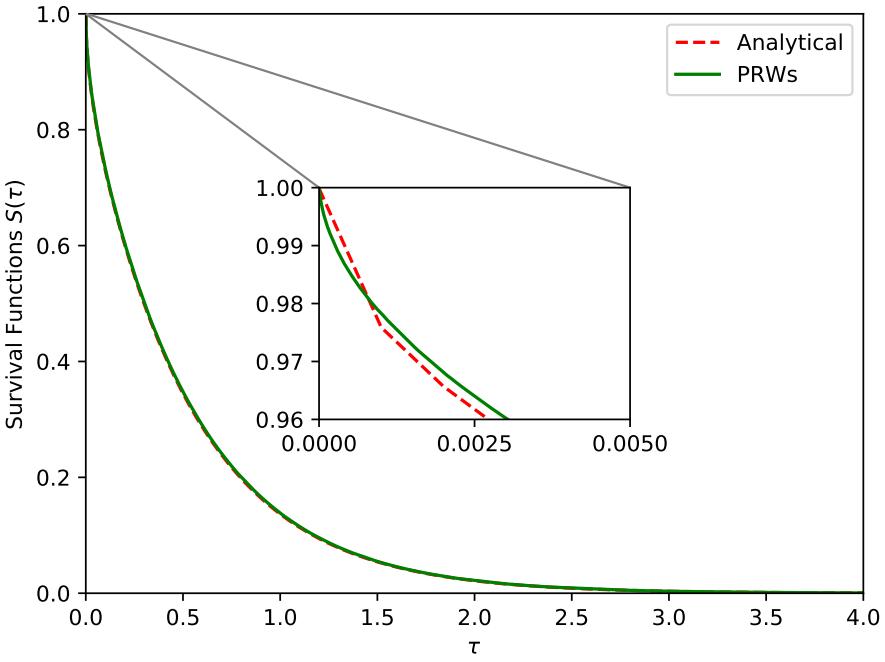
The required number of particles in LRWs and PRWs can be estimated by Eq.(3.3), which is

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

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 $[0.46865856, 0.47812641]$.



(a)



(b)

Figure 3.3: Running PRWs and LRWs in the annulus with 1338643 particles determined by the Eq.(3.2). (a) and (b) illustrate that the short and long time asypototic behaviors of the estimated survival functions of particles' diffusing times in LRWs and PRWs are consistent with the analytical result.

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.

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.

From the visualized comparison in Figure 3.3 and the results of the two-sample statistical tests, 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 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 Fig.3.4. The empirical

distribution function $F_{256}(\tau)$ is estimated by the lifeline module in Python [**?**]. Thus, the interval ε contains the entire $F(\tau)$ with the probability 95% can be calculated by Eq.(2.20)

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

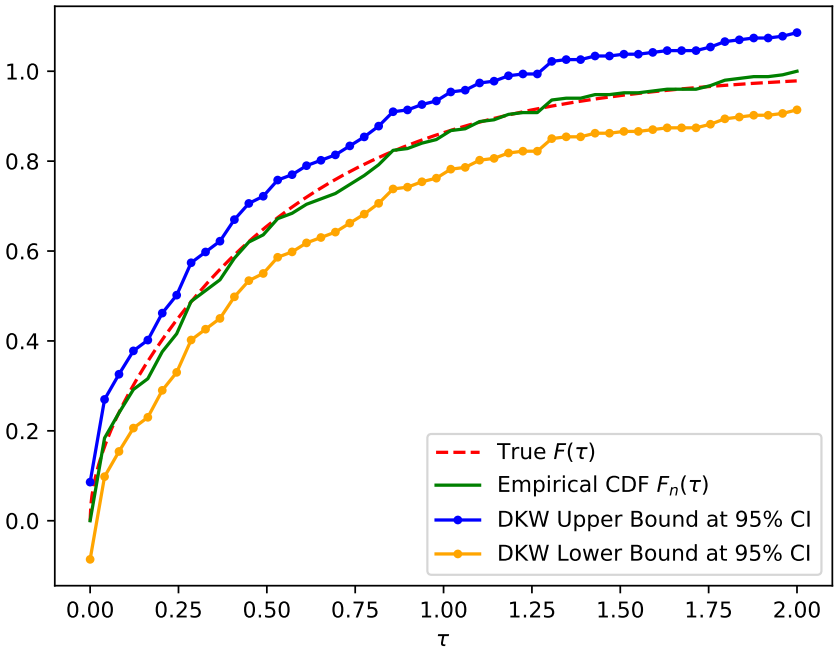


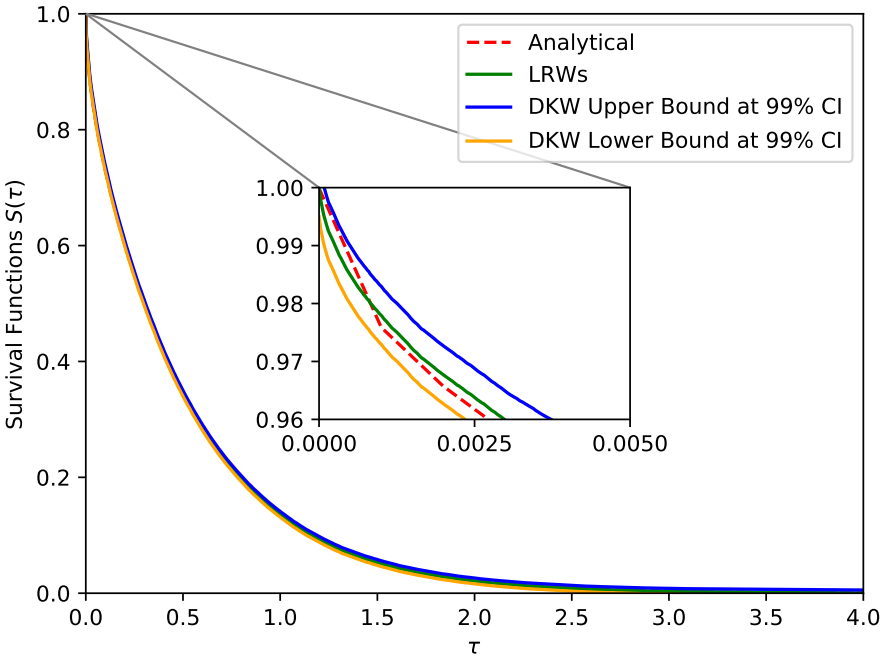
Figure 3.4: The simultaneous band around $F_{256}(\tau)$ with interval 0.084881 calculate by Eq.(3.4) encompasses the entire $F(x)$ at 95% confidence level.

Moreover, the sample size estimated by DKW inequality Eq.(2.20) 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

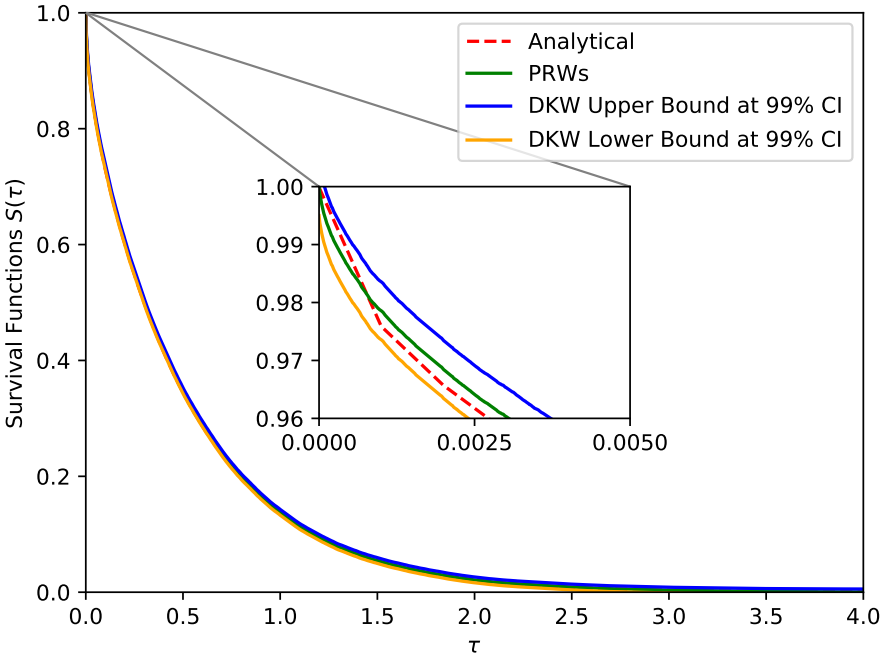
$$Pr(\sup_{x \in \mathbb{R}} |F_N(\tau) - F(\tau)| > 0.005) \leq 2e^{-2N0.005^2} = 0.01 \tag{3.5}$$

Thus

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



(a)



(b)

Figure 3.5: PRWs and LRWs are implemented in the annulus with 105966 particles determined by the Eq.(3.6). (a) and (b) show that the simultaneous confidence bands of estimated survival function with the interval 0.005 contain the entire analytical $S(\tau)$ at 99% confidence level.

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.

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.

3.1.3 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.