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

[Yuge 1] Dave, please have a look at this section and give me feedback. Thanks!

[Yuge 2] Dave, please have a look at this section and give me feedback

[Yuge 8] Dave, please have a look at this section and give me feedback. Thanks!

Existing Morphological Descriptors for Root Systems

1.1 Background

...

1.1.1 Importance of Roots

1.1.1.1 Agriculture

...

1.1.1.2 Environment and Sustainability

...

1.1.2 Importance of Research

...

1.2 Summary of Existed Descriptors

...

1.2.1 Metric

...

1.2.2 Non-Metric

...

1.3 Problem Statements

...

1.3.1 Limitation of Data

...

1.3.2 Incompleteness and Low Efficiency

...

1.3.3 Incorrectness

...

An Alternative Mathematical Method for Shape Description

2.1 Kac's Idea: Can One Hear the Shape of a Drum?

...

2.1.1 Interpretation

...

2.1.2 Summarize Kac's Idea

...

2.2 Extended Works of Kac's Idea: Heat Content

...

2.2.1 Mathematical Formula

...

2.2.2 Exploration of Geometrical Information

...

2.2.3 Difficulties in Application

...

2.3 Numerical Methods for Solving Parabolic Partial Differential E

[Yuge 9] Dave, please have a look at this section and give me feedback. Thanks!

The heat equation is a crucial time-dependent parabolic partial differential equations (PDE) characterizing how a quantity diffuses through a given domain over time. From the physical interpretation, its solution describes the heat distribution or temperature varying in time and positions and can be obtained uniquely by considering specific initial and boundary conditions. When the diffusion coefficient is constant, the general solution of the heat equation is in one of the two standard forms [12]. One is constituted of a series of error functions or related integrals, which helps evaluate the diffusion behaviour numerically in the early stage. Another one is in the form of a trigonometric series, which converges rapidly for a long time. If the heat equation is defined in a cylinder, a series of Bessel functions will replace the trigonometric series.

As introduced in the section 2.2, the heat content $Q_\Omega(t)$, defined as the total amount of heat in a bounded domain Ω, is the integration of the solution to the heat equation over the space-dimension. The geometric features of Ω can be obtained from the coefficients of the asymptotic expansion of $Q_\Omega(t)$. However, so far, only the first few terms are available explicitly. Furthermore, the traditional analytical techniques 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 pure analytical techniques can apply strictly only to the linear form of the boundary conditions and to constant diffusion properties [12].

Therefore, numerical techniques and simulations are more practical and applicable to find solutions to the partial differential equations (PDEs) than the pure analytical methods. The methods 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.

2.3.1 Finite Difference Method

[Yuge 10] Dave, please have a look at this section and give me feedback. Thanks!

FDM is a frequently utilized technique to convert the heat equation into a system of algebraically solvable equations [25]. The basic idea is to replace the derivatives in the equation by the difference quotients. For example, the FTCS (Forward Time Centered Space) scheme [49] discretizes the Laplace operator in space and the time derivative. The discretization results in the integer number of points in space and an integer number of times. At a specific space-time point, the field variables can be calculated by considering the initial and boundary conditions. In the following case, field variable is the temperature.

Let $u(x, y, t)$ be the temperature distribution at position (x, y) and time t in a $2-$dimensional homogeneous and isotropic domain Ω. It is well-known that without any internal heat sources in Ω, $u(x, y)$ satisfies the heat equation.

$$u_t = D\left(u_{xx} + u_{yy}\right) \tag{2.1}$$

$$u(x, y, t) = 0 \qquad \text{for } (x, y) \in \partial\Omega \tag{2.2}$$

$$u(x, y, 0) = \frac{1}{|\Omega|} \qquad \text{for } t = 0 \tag{2.3}$$

Note, D is a constant diffusion coefficient. u_t inidicates partial derivative with respect to time t, while u_{xx} and u_{yy} indicate second partial derivative with respect to x and y repectively. Eq. 2.2 is the Dirichlet boundary condition, also called the absorbing boundary condition, implying the temperature at the boundary of the domain is cooled down to 0. Eq. 2.3 denotes the constant initial temperature distribution.

Before applying of FTCS, let descrtize Ω along the $x-$axis and $y-$axis as a regular lattice. Both the range of x and y are divided into equal intervals $\triangle l$. Also, the time is devided into equal interval δ. Let the corrdinates of a representative grid point (x, y, t) be $(i\triangle l, j\triangle l, n\delta)$, where $\triangle l$ is the distance between two neighboring sites of the lattice and δ is the time step. For simplicity, we denote the value of u at the point $(i\triangle l, j\triangle l)$ at time $n\delta$ by $u(i, j, n)$.

By applying the Taylor's series [60], the difference formula for time derivative is

$$u_t = \frac{u(i, j, n+1) - u(i, j, n)}{\delta} + \mathcal{O}(\delta) \tag{2.4}$$

Similarly, the difference formula for the spatial derivaive of x and y are

$$u_{xx} = \frac{u(i-1, j, n) - 2u(i, j, n) + u(i+1, j, n)}{(\triangle l)^2} + \mathcal{O}\left((\triangle l)^2\right) \tag{2.5}$$

$$u_{yy} = \frac{u(i, j-1, n) - 2u(i, j, n) + u(i, j+1, n)}{(\triangle l)^2} + \mathcal{O}\left((\triangle l)^2\right) \tag{2.6}$$

Dropping the error terms $\mathcal{O}(\delta)$ and $\mathcal{O}((\triangle l)^2)$ and substituting the Eq. 2.4, Eq. 2.5, and Eq. 2.6 into original heat equation Eq. 2.1, there will have

$$\frac{u(i,j,n+1)-u(i,j,n)}{\delta} = D\left(\frac{u(i-1,j,n)+u(i+1,j,n)-4u(i,j,n)+u(i,j-1,n)+u(i,j+1,n)}{(\triangle l)^2}\right) \tag{2.7}$$

Rearranged Eq. 2.7 as

$$u(i,j,n+1) = \frac{D\delta}{(\triangle l)^2}\Big(u(i-1,j,n)+u(i+1,j,n)-4u(i,j,n)+u(i,j-1,n)+u(i,j+1,n)\Big)+u(i,j,n) \tag{2.8}$$

Finally, the value of $u(i,j,n+1)$ can be expressed explicitly in terms of $u(i-1,j,n)$, $u(i+1,j,n)$, $u(i,j-1,n)$, $u(i,j+1,n)$, and $u(i,j,n)$ as

$$u(i,j,n+1) = \beta\Big(u(i-1,j,n)+u(i+1,j,n)+u(i,j-1,n)+u(i,j+1,n)\Big)+(1-4\beta)u(i,j,n) \tag{2.9}$$

Where

$$\beta = \frac{D\delta}{(\triangle l)^2} \tag{2.10}$$

The FTCS scheme is conditionally stable [49] because the explicit formula in Eq. 2.9 is stable if and only if $\beta \leq \frac{1}{4}$, which means

$$\delta \leq \frac{(\triangle l)^2}{4D} \tag{2.11}$$

Eq. 2.11 reveals the relationship between the spatial resolution $\triangle l$ and the time-step δ, which implies that the high spatial resolution calculations demand an extremely tiny time-step to maintain numerical stability.

After discretizing the time and space, defining the condition for stability and initializing the solution's matrix, the next step of FTCS is to fill in initial and boundary conditions. Based on Eq. 2.9, the field values at the finite number of space-time points can be estimated by the iterations or solving the linear algebraic equations. Finally, visualizing the numerical result and checking the approximation based on the known knowledge of the systems.

There are three kinds of errors that should be considered when using FDM. First of all, the truncation error appears in the process of ignoring the higher-order terms in the Taylor series to obtain the finite-difference equations. 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 [12]. Another kind of error in FDM is round-off error, which results from the loss of precision due to the computer rounding of decimal quantities. [28]. The last type of error is the discretization error, which can be reduced by decreasing the time size, grid size, or both of them [12]. Finally, DFM becomes inaccurate and arduous in the practical application when the problem is defined in the irregular geometries since the heat equation must be transformed before applying the Taylor series.

2.3.2 Finite Element Method

Unlike the FDM, the finite element method (FEM) [64] divides the complicated geometries, irregular shapes, and boundaries into the union of smaller and simpler subdomains (e.g. lattice, triangle, curvilinear polygons, etc.), which are called finite elements [45]. The smaller size of the finite element mesh, the more accurate approximate for the solution. Moreover, FEM has great flexibilities or adaptivity [52]. For example, FEM can provide higher fidelity or higher accuracy in a local region and keep other subdomains the same. Each subdomain is locally represented by the element equation, the continuous piecewise shape functions, which are finally assembled into a system of algebraic equations for modelling the entire problem. For a considerable number of elements, the parallelling computation can approximate the solution numerically by minimizing the associated error function. Nevertheless, FEM heavily relies on numerical integration, where the quadrature rules sometimes cause difficulties. FEM requires an amount of human involvement in building the FE model, checking the result, detecting and updating the model design. Moreover, FEM demands a longer execution time and an enormous amount of input data compared with FDM.

[Yuge 11] Dave, please have a look at this section and give me feedback. Thanks!

2.3.3 Other Numerical Techniques

Another method closely related to the FEM is the finite volume method (FVM) since its fundamental idea is to divide the computational region into a system of independent control volumes. Around each grid point of the whole domain, there has a control volume. A set of discrete equations will be obtained by integrating the PDE over each volume [20]. However, the accuracy of FVM is related to the integration over time and space. Unlike the domain-type methods (e.g. FDM, FEM, FVM, etc.), the boundary element method (BEM) transforms the heat equation into an integral equation over the boundary of the domain using the boundary integral equation method [5]. When the region extends to infinity or its boundary is complex, BEM is more efficient in computation than other methods because of the smaller surface or volume ratio [38] since it only discretizes the boundary and fits the boundary values into the integral equation [4]. The reduction of the dimensionality of the problem is the basic advantage of BEM. However, the matrics generated in the BEM are generally unsymmetric and fully populated, which are difficult to be solved [47].

[Yuge 12] Dave, please have a look at this section and give me feedback. Thanks!

2.3.4 Limitations in Practice

In this thesis, the heat equation defined in 2-dimensional domain Ω, which is bounded by the border of the image and the whole root system. Ω has millions of pixels, the extremely complex roots and various boundary conditions. For calculating the heat content contained in Ω, the numerical computational techniques can be used to approximate the solutions of the heat equation, but some practical difficulities have to be considered since all of them have an intrinsically similar feature - mesh discretization in the time and space dimension. For instance, the far more efforts are required in sovling heat equation by FDM and FVM because of the complicated boundary of the roots and non-continuous issues. Although the whole 2-dimensional root image can be regarded 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 approximation of the original IBVP and the higher cost of computation time in the numerical methods. Moreover, the final goal of this research is not the numerical approximation of the temperature distribution over time and space, but the heat content $Q_\Omega(t)$, defined as an integration over the space dimension, which results in the extra effort and errors.

[Yuge 13] Dave, please have a look at this section and give me feedback. Thanks!

2.4 Monte Carlo Methods in Solving Partial Differential Equations Yuge

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 [41] [54]. 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 [42][39], 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 [23][55].

Monte Carlo methods are barely applied in solving parabolic partial differential equations, such as the $1-$ dimensional and $2-$ dimensional time-dependent heat problem. For example, Eq. 2.9 is a Smoluchowski equation [59] and provides a probabilistic interpretation of the Brownian particle, which is moving randomly one step to $u(i-1,j)$, $u(i+1,j)$, $u(i,j-1)$, or $u(i,j+1)$, with probability β, or remaining at the same position (i,j), with probability $(1-4\beta)$. As introduced in the papers [56] [21], given any $\beta \leq \frac{1}{4}$, the solution of the $2-$dimensional heat equation at a specific space-time point can be approximated by averaging the value of a large number of random-walking particles, whose trajectories or directions for each step are determined by the random numbers generated by the Monte Carlo techniques, on the any sites of the target boundary. However, the drawbacks of this method are obvious. Firstly, there is the error appeared in the finite-difference approximation. Secondly, there have statistical sampling errors inherent in the Monte Carlo simulations. Thirdly, each particle's detailed trajectory will be simulated until it is absorbed by the boundary, so it will be time-consuming in the simulation with a high spatial resolution. Moreover, this method can only estimate the solution to the heat equation at a particular space-time point.

Without mimicking the detailed trajectories, another efficient random process in the continuous time and continuous space simulated by MCMs has been applied frequently in solving the elliptic partial differential equation, for example, the Laplace's and Poisson's equations [26] [9] [46], and steady-state diffusion equation [61]. For example, let $v(\boldsymbol{s_0})$ be the value of the solution to an elliptic PDE at a specific point $\boldsymbol{s_0}$ in the domain Ω with initial and boundary conditions. $v(\boldsymbol{s_0})$ can be estimated by a point $\boldsymbol{s_1}$, which is sampled uniformly on the largest circle C_0 centered at $\boldsymbol{s_0}$ with radius r_0 lying entirely in Ω. If $\boldsymbol{s_1}$ gets closed to the target bounday within an error, $v(\boldsymbol{s_1})$ is known, and it can be considered as one particle's estimate of $v(\boldsymbol{s_0})$ by multiplying the particle's statistical weight. If not, $v(\boldsymbol{s_1})$ should be estimated in the same way as $v(\boldsymbol{s_0})$, that is, $\boldsymbol{s_2}$ is sampled uniformly on the largest circle C_1 centered at $\boldsymbol{s_1}$ with radius r_1 lying entirely in the domain. Check the position of $\boldsymbol{s_2}$, and the procedure will be repeated until the simulation is terminates on the traget boundary, which is defined as one particle's estimate of $v(\boldsymbol{s_0})$. Finally, averaging a larger number of one-particle estimates, $v(\boldsymbol{s_0})$ will be more accurate. However, this process has not been used in solving the non-steady-state diffusion problem.

2.5 Monte Carlo Simulation for Approximating Heat Content $Q_\Omega(t$

In the section 2.3, several generally utilized numerical methods [25][64] [20] [5] for approximating the solutions to the heat equation and their limitations in practice are presented. In this section, instead of understanding the heat equation from the macroscopic view, two microscopic probabilistic interpretations are introduced. Moreover, the Monte Carlo method (MCM) [54] [41], one of the non-deterministic techniques, and its application in solving PDEs numerically are summarized. Moreover, an alternative fixed-time step Monte Carlo simulations, lattice random walks (LRWs), is designed. The most outstanding advantage of the proposed random walk model is that the heat content can be approximated directly based upon the probabilistic interpretation of 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.

[Yuge 14] Dave, please have a look at this section and give me feedback. Thanks!

2.5.1 Background

[Yuge 15] Dave, please have a look at this section and give me feedback. Thanks!

Instead of understanding the heat equation Eq. 2.1 and its solution from the physical point of view at the macroscopic scale, it will be easier to think about it from the microscopic peospective: a large amount of randomly moving particles.

2.5.1.1 Brownian Motion Yuge

In the early 19th century, the French scientist Jean Baptiste Joseph Fourier presented the remarkable heat equation describing the heat conduction in a solid and proposed a technique, named the separation of variable, to solve the equation subject to the initial and boundary conditions [7]. In 1827, the Scottish botanist Robert Brown discovered the microscopic motion of the pollen grains, which suspended in the water going in an erratic and highly irregular zigzag pattern [10]. Based on his report, other scientists verified the existence of the Brownian motion, defined as the irregular motion of individual particles, in other fields. Originally, in 1905, Albert Einstein [17] developed a satisfactory explanation of the Brownian motion from the jostling of the particle by the molecules of water. He also utilized a probabilistic model showing that the Brownian motion, in a certain sense, provided a solution to Fourier's heat equation. In 1918, Norbert Wiener began considering the Brownian motion as a mathematical random process with the required properties in a rigorous way [63]. Therefore, Brownian motion, also called the Wiener process, is a continuous-time and continuous-space stochastic process with the continuous sample paths and stationary independent increments [31] [37]. The stationary increment of the Brownian motion is a statement of time homogeneity; that is, the distribution of the particle's displacement in a time interval depends only on the length of the time interval. Brownian particles' independent increments imply the Markov property since the particle's displacement is independent of its past displacement. Therefore, Brownian motion can be considered as one of the fundamental Markov processes.

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 the diffusion process, which satisfies the diffusion equation [32][62] or heat equation. Let $P(\boldsymbol{s}, t|\boldsymbol{s_0})$ be the probability density for Brownian motion at a point $\boldsymbol{r} \in \Omega$ at time $t > 0$ started from $\boldsymbol{s_0} \in \Omega$ at $t = 0$ satisfying the following equations

$$\frac{\partial P(\boldsymbol{s}, t|\boldsymbol{s_0})}{\partial t} = \Delta D P(\boldsymbol{s}, t|\boldsymbol{s_0}) \tag{2.12}$$

$$P(\boldsymbol{s}, t|\boldsymbol{s_0}) = \delta(\boldsymbol{s} - \boldsymbol{s_0}) \qquad \text{for } t = 0 \tag{2.13}$$

$$P(\boldsymbol{s}, t|\boldsymbol{s_0}) = 0 \qquad \text{for } t > 0 \text{ and } \boldsymbol{s} \in \partial\Omega \tag{2.14}$$

Note, D is the diffusion constant. To ensure the existence and uniqueness of the solution, the heat equation Eq. 2.12 has to be supplemented by the initial condition Eq. 2.13 with the Dirac $\delta-$ distribution

and the Dirichlet boundary condition Eq. 2.14, which indicates the absorption and disappearance of the particle hitting the boundary $\partial\Omega$.

2.5.1.2 Random-Walk Theory Yuge

Brownian motion has existed for a long time before the considerable development of the random-walk theory. At the beginning of the twentieth century, the term, random walk, was initially proposed by Karl Pearson [48]. He utilized the isotropic planar random flights to model the erratic migration and invade of mosquitoes in the cleared jungle regions. In his model, the mosquito moves to a random direction with a fixed step length at each time step. Moreover, he asked a question in a paper about the distribution of mosquitos after many steps have been taken. Actually, earlier in 1880, John William Strutt and Baron Rayleigh have used the similar concept of random walk in analyzing a certain random vibration problem [50]. Therefore, in 1905, Rayleigh [51] answered Pearson's question, that is, the distribution is identical to superposition the sound vibrations with unit amplitude and arbitrary phase [15]. At almost the same time, Louis Bachelier 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 [6]. While Rayleigh only studied the symmetric random walk, the nonsymmetric case was considered by Mark Kac in 1947 [32]. Kac provided a proof that the nonsymmetric random walk is the mathematical analogy to the Smoluchowski's diffusion with drift [59]. Moreover, 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 [17] [16] [59].

2.5.1.3 Lattice Random Walks (LRWs) Yuge

The continuous Brownian motion is the scaling limit of the discrete random walks as the time and space increments approach zero [43][62]. 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 [53] 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 denotes the particle will return to its origin only finitely often with the probability 1 [30] [43].

From the probalistic view, only 2-dimensional lattice random walks (LRWs) and its connection with the heat equation are introduced here since this thesis aims to explore and characterize the shape of roots in the $2-$dimensional images. In the random walk, let $\triangle l$ be the distance between two sites in the lattice and δ be the time between two steps. Let $P(x, y, t+\delta|\boldsymbol{s_0})$ be the conditional probability of a particle to be in position (x, y) at time $t+\delta$ starting from $\boldsymbol{s_0}$ at time $t = 0$. Without hitting the boundary $\partial\Omega$ of the domain Ω, $P(x, y, t+\delta|\boldsymbol{s_0})$ can be expressed by the stochastic difference equation

$$P(x,y,t+\delta|\boldsymbol{s_0}) = \frac{P(x-\triangle l,y,t|\boldsymbol{s_0}) + P(x+\triangle l,y,t|\boldsymbol{s_0}) + P(x,y-\triangle l,t|\boldsymbol{s_0}) + P(x,y+\triangle l,t|\boldsymbol{s_0})}{4} \tag{2.15}$$

After rewriting the master Eq. 2.15 by subtracting $P(x,y,t|\boldsymbol{s})$ and dividing by δ,

$$\begin{aligned}\frac{P(x,y,t+\delta|\boldsymbol{s_0}) - P(x,y,t|\boldsymbol{s_0})}{\delta} = \frac{1}{4\delta}\Big(&P(x-\triangle l,y,t|\boldsymbol{s_0}) + P(x+\triangle l,y,t|\boldsymbol{s_0}) \\ &+ P(x,y-\triangle l,t|\boldsymbol{s_0}) + P(x,y+\triangle l,t|\boldsymbol{s_0}) - 4P(x,y,t|\boldsymbol{s_0})\Big)\end{aligned} \tag{2.16}$$

As $\delta \to 0$, the left side of Eq. 2.16 is the definition of a derivative repect to time

$$P(x,y,t+\delta|\boldsymbol{s_0}) \approx P(x,y,t|\boldsymbol{s_0}) + \frac{\partial P(x,y,t|\boldsymbol{s_0})}{\partial t}\delta \tag{2.17}$$

For the right side of Eq. 2.16, expanding the conditional probability P as Taylor series and ignoring the higher order terms, there has

$$P(x \pm \triangle l,y,t|\boldsymbol{s_0}) \approx P(x,y,t|\boldsymbol{s_0}) \pm \frac{\partial P(x,y,t|\boldsymbol{s_0})}{\partial x}\triangle l + \frac{1}{2}\frac{\partial^2 P(x,y,t|\boldsymbol{s_0})}{\partial x^2}(\triangle l)^2 \tag{2.18}$$

$$P(x,y \pm \triangle l,t|\boldsymbol{s_0}) \approx P(x,y,t|\boldsymbol{s_0}) \pm \frac{\partial P(x,y,t|\boldsymbol{s_0})}{\partial y}\triangle l + \frac{1}{2}\frac{\partial^2 P(x,y,t|\boldsymbol{s_0})}{\partial y^2}(\triangle l)^2 \tag{2.19}$$

Substituting Eq. 2.18 and Eq. 2.19 into right side of Eq. 2.16 and implementing all the obvious cancellations, the Eq. 2.16 is rearranged as

$$\frac{\partial P(x,y,t|\boldsymbol{s_0})}{\partial t} \approx \frac{(\triangle l)^2}{4\delta}\Big(\frac{\partial^2 P(x,y,t|\boldsymbol{s_0})}{\partial x^2} + \frac{\partial^2 P(x,y,t|\boldsymbol{s_0})}{\partial y^2}\Big) \tag{2.20}$$

When $\tau \to 0$, $\triangle l \to 0$, and $\tau \sim (\triangle l)^2$, Eq. 2.20 becomes an exact relation.

$$\frac{\partial P(x,y,t|\boldsymbol{s_0})}{\partial t} = D\Big(\frac{\partial^2 P(x,y,t|\boldsymbol{s_0})}{\partial x^2} + \frac{\partial^2 P(x,y,t|\boldsymbol{s_0})}{\partial y^2}\Big) \tag{2.21}$$

where $D = \frac{(\triangle l)^2}{4\delta}$ is the diffusion coefficent. This above derivation shows a relationship between a very simple microscopic discrete random walk model and a well-known macroscopic probalistic partial differential equation.

2.5.2 Monte Carlo Integration: Survival Probability $S(t)$

[Yuge 16] Dave, please have a look at this section and give me feedback. Thanks!

Brownian motion is a continuous mathematical model for the heat flow by averaging the value of a large number of particles. Also, the Brownian motion interpretation provides a solution formula for the heat equation Eq. 2.1 by $P(\boldsymbol{s}, t|\boldsymbol{s_0})$

$$u(\boldsymbol{s}, t) = \int_\Omega P(\boldsymbol{s}, t|\boldsymbol{s_0})\delta(\boldsymbol{s} - \boldsymbol{s_0})d\boldsymbol{s_0} \tag{2.22}$$

In Eq. 2.22, $P(\boldsymbol{s}, t|\boldsymbol{s_0})$, also called the foundamental solution or heat kernel of Ω, can be expressed in terms of the eigenvalues and eigenfunctions of the Dirichlet Laplacian

$$P(\boldsymbol{s}, t|\boldsymbol{s_0}) = \sum_{j=1}^{\infty} \phi_j(\boldsymbol{s_0})\phi_j(\boldsymbol{s})e^{-D\lambda_j t} \tag{2.23}$$

where the ϕ_j form an orthonormal basis of $L^2(\Omega)$ of real valued eigenfunctions corresponding to the $\lambda_j \in \mathbb{R}^+$ satisfying

$$-\Delta\phi_j = \lambda_j\phi_j \tag{2.24}$$

Given any starting point $\boldsymbol{s_0} \in \Omega$ at $t = 0$, let $S(t|\boldsymbol{s_0})$ be the probability of the Brownian motion keeps diffusing in Ω at time $t > 0$ without being absorbed by $\partial\Omega$, which is

$$S(t|\boldsymbol{s_0}) = \int_\Omega d\boldsymbol{s}P(\boldsymbol{s}, t|\boldsymbol{s_0}) \qquad \text{where} \tag{2.25}$$

$$S(t|\boldsymbol{s_0}) = 0 \qquad \text{for } \boldsymbol{s_0} \in \partial\Omega \tag{2.26}$$

If the initial locations spread uniformly in Ω, the survival probability $S(t)$ for the Brownian motion wil be the average of the local survival probability over Ω.

$$S(t) = \frac{1}{|\Omega|}\int_\Omega d\boldsymbol{s_0}S(t|\boldsymbol{s_0}) = \sum_{j=1}^{\infty}\gamma_j^2 e^{-D\lambda_j t} \qquad \text{where} \tag{2.27}$$

$$\gamma_j = \frac{1}{\sqrt{|\Omega|}}\int_\Omega d\boldsymbol{s}\phi_j(\boldsymbol{s}) \tag{2.28}$$

Eq. 2.25 and Eq. 2.27 reveal that the heat content $Q_\Omega(t)$, calculated as the integration of the solution to the heat equation over Ω, is proportional to the survival probability $S(t)$ [34].

2.5.3 Monte Carlo Simulation for Diffusing Particles: Lattice Random Walks

...

2.5.3.1 Algorithm of Lattice Random Walks

...

2.5.3.2 Methods for Sample Size Determination

...

2.5.4 Output Analysis

Particles' first-passage time t, also called the first-hitting time, is the output of the lattice random walks (LRWs). More specifically, t is the number of steps taken by the particle encountering any positions of the target boundary for the first time. Since the first-hitting-time model is a sub-class of survival analysis in statistics [3], it is straightforward to estimate the survival function $S(t)$ of the numerical simulation by the Kaplan-Meier estimator $\widehat{S}(t)$ [40]. $S(t)$ provides the probability that the particle remains simulating in the LRWs beyond any specified time. Moreover, the pointwise upper and lower confidence interval of $S(t)$ can be estimated by the Greenwood's exponential formula [29]. In this subsection, the Kaplan-Meier estimator and pointwise confidence intervals are presented theoretically. However, in practice, the existing Python module, lifeline [14], helps implement the estimation. Based on the scaling relationship between the number of steps t and the unitless time τ, non-parametric two-sample statistical tests are used to compare $S(t)$ and $S(\tau)$ for validating the research methodology.

2.5.4.1 Kaplan-Meier Estimator

The survival time is defined as the time starting from a specified point to the occurrence of a given event [8], such as death, pregnancy, job loss, etc. The analysis of the group of survival time data is called survival analysis [3]. In the survival analysis, three kinds of situations will affect the subjects' survival time [22]. Firstly, the subjects are uncooperative and naturally dropout or withdrawal from the research. Secondly, some subjects do not experience the event by 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 carry partial information about survival, the scientists will label the above circumstances as censored observations [8] instead of ignoring them and decreasing the sample size.

In clinical trials, Kaplan-Meier estimator [35], 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 [3]. Kaplan-Meier estimator also works well with the mentioned three difficult situations under the consideration of various assumptions [18] [22]. First of all, the event status only consists of two mutually exclusive states: censored (0) or failure (1). Secondly, the censored observations have identical survival prospects as those who keep being observed. Thirdly, the objects have equal survival probabilities no matter they participate in the research early or late. Lastly, the survival time, the time to an event or censorship, should be clearly defined and precisely recorded in the study.

Let $0 < t_1 < t_2 < ...$ be the monotonously increasing observed times, or the number of steps taken by the particle hitting the absorbing boundary, in the numerical simulations. Let n_i be the number of particles, who have not yet stop the simulation up to t_i or be absorbed exactly at t_i. 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

$S(t)$ of the numerical simulation is [1]

$$\widehat{S}(t) = \prod_{i:t_i \leq t} (1 - \frac{d_i}{n_i}) \tag{2.29}$$

2.5.4.2 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 by Greenwood in 1926 [24],

$$\widehat{S}(t) \pm z_{\alpha/2} \sqrt{\widehat{Var}[\widehat{S}(t)]} \qquad \text{where} \tag{2.30}$$

$$\widehat{Var}[\widehat{S}(t)] = \widehat{S}(t)^2 \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)} \tag{2.31}$$

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

In 1999, Hosmer and Lemeshow [29] developed the exponential Greenwood formula based on the earlier works [33], which provides an asymmetric confidence interval for $S(t)$

$$e^{-e^{c_+(t)}} < S(t) < e^{-e^{c_-(t)}} \qquad \text{where} \tag{2.32}$$

$$c_\pm(t) = log(-log\widehat{S}(t)) \pm z_{\alpha/2}\sqrt{\widehat{V}} \qquad \text{and} \tag{2.33}$$

$$\widehat{V} = \frac{1}{(log\widehat{S}(t))^2} \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)} \tag{2.34}$$

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

Compared with the traditional calculations of the Greenwood confidence interval, the exponential Greenwood formula will make sure that the endpoints in Eq. 2.32 lie in $(0, 1)$, while the endpoints in Eq. 2.30 could be negative or larger than 1 [57].

2.5.4.3 Scaling Relationship

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

$$\triangle l = 2\sqrt{D\delta} \tag{2.35}$$

where D is the diffusion coefficent.

Eq. 2.35 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,

$$\tau = t\delta = \frac{(\triangle l)^2 t}{4D} \tag{2.36}$$

where D is 1.

For example, 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.

2.5.4.4 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 [19]. 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 [19]. 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 [11]. The hazard ratio is an estimate of the hazard rate in one group relative to that in another group [58]. 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 [13].

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 [13]. 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 [13]. 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 [36].

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 [44]. It pays more attention to the failures happening somewhere in the middle of study [18]. Moreover, there is no limitation of the number of groups when the Tarone-Ware test is applied [13]. 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 [27].

Method Validation in Annulus

3.1 Analytical Results

...

3.1.1 Solving Heat Equation

...

3.1.2 Heat Content $Q_\Omega(\tau)$

...

3.2 Numerical Approximation

...

3.2.1 Eigenvalue Estimation

...

3.2.2 Approximation of Solution

...

3.2.3 Approximation of Heat Content

3.3 Lattice Random Walks in Annulus

...

3.3.1 Algorithm

...

3.3.2 Sampling Errors

...

3.3.3 Sample Size Determination

3.4 Comparison of Numerical and Analytical Results

...

3.4.1 Sample Size Evaluation

...

3.4.2 Comparison of $Q_\Omega(\tau)$ and $S(t)$

...

3.4.3 Conclusion

Lattice Random Walks on Artificial Images

4.1 Shape Design

...

4.1.1 Simple Shapes

...

4.1.1.1 Shape Description

...

4.1.1.2 Purpose

...

4.1.2 Complicated Shapes

....

4.1.2.1 Shape Description

...

4.1.2.2 Purpose

...

4.2 Assumption Verification

...

4.2.1 Circle and Rectangular

...

4.2.2 Artifical Branching Structures

...

4.2.3 Conclusion

...

4.3 Conclusion

...

Lattice Random Walks in Real Root Images

5.1 Image Description

...

5.2 Output Analysis

...

5.2.1 Distance Matrices

...

5.2.2 Statistical Tests for Distance Matrices

...

Conclusion

6.1 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] Jean Baptiste Joseph Baron Fourier. *The analytical theory of heat*. The University Press, 1878.

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

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

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

[11] 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.

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

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

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

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

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

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

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

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

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

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

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

[23] 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.

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

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

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

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

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

[29] 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.

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

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

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

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

[34] 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.

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

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

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

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

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

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

[41] 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.

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

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

[44] 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.

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

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

[47] 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.

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

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

[50] Lord Rayleigh. Xii. on the resultant of a large number of vibrations of the same pitch and of arbitrary phase. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 10(60):73–78, 1880.

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

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

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

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

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

[56] 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.

[57] 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.

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

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

[60] Brook Taylor. *Methodus incrementorum directa & inversa. Auctore Brook Taylor, LL. D. & Regiae Societatis Secretario*. typis Pearsonianis: prostant apud Gul. Innys ad Insignia Principis in ..., 1715.

[61] S Torquato and In Chan Kim. Efficient simulation technique to compute effective properties of heterogeneous media. *Applied physics letters*, 55(18):1847–1849, 1989.

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

[63] Norbert Wiener. Differential-space. *Journal of Mathematics and Physics*, 2(1-4):131–174, 1923.

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