Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/6/20230818

The Pólya's Random Walk Theorem

Jingyi Lu

University of California, Santa Barbara, Santa Barbara, United States

jingyi_lu@ucsb.edu

Abstract. Since there are a variety of uncertainties and probabilities in the actual world, mathematicians and other researchers have found numerous answers to these common problems. They have also developed theories, such as the probability theory, to help them organize their significant discoveries. A drunk person will eventually find their way home, but a drunk bird might never find it, as said by Shizuo Kakutani. A random walk, in its most basic definition, is just a path on a network or lattice where each step is decided at random using some probability distribution, such a coin or fair dice. A motion, such as going left or right on a number line or up or down on a plane, would correspond to each outcome of tossing a coin, heads or tails. Recurrent random walks are those that end up back at the beginning after a predetermined number of steps. Any non-recurrent random walk is transitory.

Keywords: Random Walk, Probabilities, Stochastic Process

1. Introduction

There are always different uncertainties and probabilities in real life, so mathematicians and other scholars have discovered many resolutions to such daily dilemmas, and they also created theories, which are called the probability theory, to organize their important findings. Stochastic process, usually containing a sequence of random variables, is a statistical study which is one of the central subjects of the probability theory. Mentioning the random variable, it has two different types: discrete and continuous random variables. For the discrete random variable, probability distribution and cumulative distribution function can be: $\sum_i p(x_i) = 1$ and $F(x) = P(X \leq x)$. For the continuous random variable, X is a continuous random variable if there exists f, which is the density function and satisfies $P(X \in A) = \int_A f(x)dx$ for any reasonable set A. When scholars contribute their efforts to observe the random variables, they also study the relationship (or they call it as correlation) between the random variables they have set. For example, they called the random variables independent from each other when the variables are found to be uncorrelated. The scholars clarify that two random variables X and Y are independent if for all C, $D \in \mathcal{R}$, $P(X \in C, Y \in D) = P(X \in C)P(Y \in D)$, and then the events set $A = \{X \in C\}$ and $B = \{Y \in D\}$ are also independent. Back to the discussion of the stochastic process, one of the examples of the stochastic process can be the well-known Bernoulli process. In the Bernoulli distribution, a random variable only takes two values: 0 or 1. The probability of 1 is always set as p, so the probability of 0 is set as $1 - p$. Relating this Bernoulli process to the real life, the game of flipping a coin can be a good representation: for example, when players flip a coin, they can have a random variable X. In the beginning, X is equal to 0. Then, they can add 1 to X each time when they get a head and add 0 to X when they get a tail. If people want to compute the probability of getting some number k as X's final value in the end, they can use the formula: $(X = k) = \binom{n}{k} p^k(1-p)^{n-k}$. In addition, in the case of flipping a coin, people can set p as $\frac{1}{2}$, so the probability of getting a tail is also $\frac{1}{2}$. The value of p varies based on different situations. From the example of flipping coins, it is clear to see what the stochastic process is in more details. It is a mathematical and statistical study of randomness and probability of different outcomes. Figuring out the probabilities relating to the result of flipping a coin may be helpful if people want to use this simple game as a gamble. Clearly, in real life, there are always uncertainties surrounding people, and people sometimes have to work to get the possibilities of different outcomes to make wise decisions. For example, people figure out a way, building probabilistic reasoning systems, to deal with these uncertainties. Basically, what people do is to create a model which has all relevant knowledge of their domain in quantitative and probabilistic terms. They can supply some general knowledge first, and then more specific information, which is called evidence, to draw the conclusions. Users can tell the system what kinds of outcomes they want (this is called queries). Probabilistic reasoning systems are flexible to answer queries regardless of people's different situations, and they can predict what kinds of events will happen in the future, infer what causes those events, and better predict the future based on the learnings from the past events. Another thing people can do with probabilistic reasoning system is to improve the general knowledge (this is what people basically know without including many details and considerations of some specific situations) by learning from the past experiences. Different from learning from past events to better predict the future, this way improves the model itself. To make such improvements, people need a learning algorithm, whose goal is to produce a new model instead of answering queries. The system will eventually predict the outcomes and return the answer as probabilities, and people can utilize such methods[1-18]. Back again to the problem of stochastic process, this essay wants to focus on a very important example of the stochastic process, random walk. Just like how people thrive to work on probabilistic reasoning systems, mathematicians and statistics experts apply many different probability theories and concepts to compute different probabilities of the random walk's behavior and observe the overall patterns. This essay will demonstrate those experts' efforts on diving more deeply in the areas of stochastic process and random walk.

2. Main body

Random walk is always an interesting and worthy subject to be studied on because it can help solve stochastic differential equations and make contributions to epidemiological modeling, problems of aggregation, protein folding, population dynamics in biology, and so on. It can be assumed that there is a point, which scholars call it as a random walker, making consecutive steps in a space at equal intervals of time. The special point is about the randomness: the direction the random walker moves at each step and the length the random walker moves at each step are selected randomly and independently. About random walk, people may always consider whether the random walker will return to its origin or not. If it does, what is the possibility of returning to the starting point and does that relates to the dimension?

The Pólya's Random Walk Theorem gives some huge insights about the questions listed in the end of the last paragraph. According to this theorem, the random walker can return to the origin, which is the place it starts to walk, in one or two dimensions with the possibility of 1; however, in a dimension greater than or equal to 3, the random walker may or may not return to the origin. The previous case is called recurrent, and the later one is called transient. There are many ways to prove such a theorem in statistical ways, and Jonathan Novak's proof is valuable and outstanding. In this essay, the focal point is to see how his proof works in details.

First of all, E can be set as the event representing the simple random walker returning to the initial position, and E_n as the event walker returning to the origin for the first time after n steps. Also, events in the set E_n are mutually exclusive to each other despite of n with different values. A key point to understand is that when n is equal to 0, E_n should be 0 because if the random walker stays at the origin without moving, it will not be counted as a return. At the same time, we set p as the possibility of the

DOI: 10.54254/2755-2721/6/20230818

event E happening, and p_n as the possibility of the event E_n happening. The relationship between p and p_n can be expressed as this: $p = \sum_{n\geq 0} p_n$. Then, some new concepts can be introduced: for example, in Novak's proof, ℓ_n is used to stand for the number of loops of length n based at a particular point in $\mathbb{Z}^d$, and r_n is used to represent the number of loops of length n, which are indecomposable. The relationship between ℓ_n and r_n can be expressed as: $\ell_n = \sum_{k=0}^{n} r_k \ell_{n-k}$ for all $n \geq 1$. Then, both sides can be divided by $(2d)^n$, which shows the total walks of length n the random walker goes from a particular point in $\mathbb{Z}^d$, and the result equation can be written as: $q_n = \sum_{k=0}^{n} p_k q_{n-k}$ for all $n \geq 1$. Making a clarification, this equation, $\frac{\ell_n}{(2d)^n}$, which is q_n, shows the possibility of the random walker returning to the origin after n steps. $\frac{r_n}{(2d)^n}$, which is p_n, represents the possibility of the random walker returning to the origin for the first time after moving n steps. Therefore, based on the two possibilities, the generating functions can be written as the following: $P(z) = \sum_{n=0}^{\infty} p_n z^n$ and $Q(z) = \sum_{n=0}^{\infty} q_n z^n$. According to the algebra of formal power series, the relationship between P(z) and Q(z) can be expressed as: $P(z)Q(z) = Q(z) - 1$. In this equation, it is necessary to know that the interval for z is $[0,1)$, and then divide both sides of the equation by Q(z) so that this equation can be obtained: $P(z) = 1 - \frac{1}{Q(z)}$, $z \in [0,1)$. Because it is easy to prove that $P(1) = \sum_{n=0}^{\infty} p_n = p$, according to the Abel's power series theorem (when the power series converges at $z = 1$ in this case, which is just proved in the previous lines, the power series' value is the limit of the value when z is set as $z \to 1$); therefore, this equation can be get: $p = \lim_{\substack{z\to 1 \\ z\in[0,1)}} P(z) = 1 - \frac{1}{\lim_{\substack{z\to 1 \\ z\in[0,1)}} Q(z)}$. From this equation, it is found that the value of p can be either 1 (recurrent case) or less than 1 (transient case), and this reflect the Pólya's Random Walk Theorem.

In this paragraph, the theorem will be attempted to be proved more thoroughly by showing how the recurrent and transient cases relate to the specific dimension numbers. First, an expression for loop generating function like this is needed: $L(z) = \sum_{n=0}^{\infty} \ell_n z^n$, and then $Q(z) = L\left(\frac{z}{2d}\right)$. From this equation, it must be noted that d represents the number of dimensions. Since better comprehension of L(z) is necessary, making an exponential loop generating function is important for further proofs: $E(z) = \sum_{n=0}^{\infty} \ell_n \frac{z^n}{n!}$. Then, more new connotations are introduced here: $\ell_n{}^{(d)}$ is the number of length n loops in $\mathbb{Z}^d$, and $E_d(z)$ is the exponential generating function while studying the loops in d-dimension. For example, this essay can focus on the two-dimensional case, which means d is set as 2. In this case, the random walker on $\mathbb{Z}^2$ has length n loop which consists of steps from horizontally and vertically. It can be set that there are some number k horizontal steps and $n - k$ vertical steps (the length is n, and the random walker moves unit steps). Therefore, the number of length n loops in $\mathbb{Z}^2$ can be expressed as: $\binom{n}{k}\ell_k{}^{(1)}\ell_{n-k}{}^{(1)}$. Because knowing when the random walker will take the k horizontal steps will determine when it will take the $n - k$ vertical steps, it can be concluded the number of length n loops on $\mathbb{Z}^2$ is: $\ell_n{}^{(2)} = \sum_{k=0}^{n}\binom{n}{k}\ell_k{}^{(1)}\ell_{n-k}{}^{(1)}$. From the equation above, the relationship between $E_1(z)$ and $E_2(z)$ can be expressed as: $E_2(z) = E_1(z)^2$. Using the same logics as above, it can be concluded that a general rule can be obtained: $E_d(z) = E_1(z)^d$. Now, turning the focus toward the one-dimensional case, similar to the 2D case, the loops can be counted and the exponential generating function can be determined. In $\mathbb{Z}$, there are k positive steps and k negative steps ($k \geq 0$); therefore, these expressions can be get: $\ell_n{}^{(1)} = \begin{cases} \binom{2k}{k}, & \text{if } n = 2k \text{ is even} \\ 0, & \text{if } n \text{ is odd} \end{cases}$ and $E_1(z) = \sum_{k=0}^{\infty}\binom{2k}{k}\frac{z^{2k}}{(2k)!} = \sum_{k=0}^{\infty}\frac{z^{2k}}{k!k!}$. Then, it is found that the exponential generating function for the one-dimensional case is a modified Bessel function of the first kind, so the differential equation can be generated as: $\Big(z^2\frac{d^2}{dz^2} + z\frac{d}{dz} -$

$(z^2+\alpha^2)\Big)F(z)=0,\ \ \alpha\in\mathbb{C}$. From the differential equation, a series representation can be obtained, which was denoted by $I_\alpha(z)$: $I_\alpha(z)=\sum_{k=0}^{\infty}\frac{(\frac{z}{2})^{2k+\alpha}}{k!\Gamma(k+\alpha+1)}$. Based on this equation, it can be concluded that $E_1(z)=I_0(2z)$; therefore, since $E_d(z)=E_1(z)^d$, this expression can be get: $E(z)=I_0(2z)^d$.

In the next step, Borel transform is basically used to help with the proof. Now, there is already a representation of the exponential generating function $E(z)$, but a representation of the ordinary generation function $L(z)$ is still required. First, the integral transform like this is necessary: $(\beta f)(z)=\int_0^\infty f(tz)e^{-t}dt$. With this equation, the transform method can be used to get the ordinary generation function of number of length n loops on $\mathbb{Z}^d$: $L(z)=\beta E(z)=\beta I_0(2z)^d=\int_0^\infty I_0(2tz)^d e^{-t}dt$. Also, from the beginning of this paragraph, it is already known that $Q(z)=L\left(\frac{z}{2d}\right)$; therefore, $Q(z)$ can be expressed as the following: $Q(z)=L\left(\frac{z}{2d}\right)=\int_0^\infty I_0(\frac{tz}{d})^d e^{-t}dt$. To prove the Pólya's Random Walk Theorem (recurrence and transience), it is necessary to see whether the integral representation of $Q(z)$, which is just obtained, converges or diverges when $z\to 1$. To reach this goal, the tail integral is needed to be used like this: $\int_N^\infty I_0(\frac{tz}{d})^d e^{-t}dt$, $N\gg 0$ (N is very large). Then, observing the behavior of the integrand while setting $t\to\infty$, this formula can be obtained: $I_0\left(\frac{tz}{d}\right)=\frac{1}{\pi}\int_0^\pi e^{tf(\theta)}d\theta$, and $f(\theta)=\frac{z}{d}\cos\theta$. To estimate this integrand while $t\to\infty$, it is needed to use a method called the Laplace's method, which is useful for determining the asymptotic behavior. It is easy to note that $f'(0)=0$ and $f''<0$. Then, the quadratic Taylor approximation of $f(\theta)$ can be used so that this can be get: $f(\theta)\approx f(0)-|f''(0)|\frac{\theta^2}{2}$. With this finding, replace $f(\theta)$ in $\int_0^\pi e^{tf(\theta)}d\theta$ so that it can be found to be approximately equal to this expression: $e^{tf(0)}\int_0^\pi e^{-t|f''(0)|\frac{\theta^2}{2}}d\theta$. What can be done is to make the integral expand to the right to infinity so that such equation is obtained: $\int_0^{+\infty} e^{-t|f''(0)|\frac{\theta^2}{2}}d\theta=\frac{\pi}{2t|f''(0)|}^{1/2}$. Then, this result can be put back to the integral and get the approximation: $\int_0^\pi e^{tf(\theta)}d\theta\approx e^{tf(0)}\frac{\pi}{2t|f''(0)|}^{1/2}$. This integral approximation's accuracy can increase when $t\to\infty$. Based on all the findings, which are already listed above, the asymptotic formula can be made as the following: $I_0(\frac{tz}{d})^d\ e^{-t}\sim c*e^{t(z-1)}(tz)^{-\frac{d}{2}}, t\to\infty$; (c is a constant). Then, this equation can be obtained by applying the monotone convergence theorem: $\lim_{\substack{z\to 1\\ z\in[0,1)}}\int_N^\infty e^{t(z-1)}(tz)^{-\frac{d}{2}}dt=\int_N^\infty \lim_{\substack{z\to 1\\ z\in[0,1)}}\int_N^\infty e^{t(z-1)}(tz)^{-\frac{d}{2}}dt=\int_N^\infty t^{-\frac{d}{2}}dt$.

Therefore, it can be concluded that observing $\int_N^\infty t^{-\frac{d}{2}}dt$, if the integral diverges, the simple random walk will be recurrent. If the integral converges, the random walk will be transient instead. When $d=1$ or 2, which means while observing the one-dimensional or two-dimensional cases, the integral $\int_N^\infty t^{-\frac{d}{2}}dt$ diverges; however, when the dimension number is greater than or equal to 3, the integral converges. Hence, the Pólya's Random Walk Theorem is proved.

3. Conclusion

People around the world always have strong incentives to study stochastic process (random walk) because such studies always make a lot of contributions to different areas and then help with improvements in people's real life. The Pólya's Random Walk Theorem is always a good topic to work on. Among different proofs and comprehensions, Novak's works shine; therefore, this essay mainly focuses on how Novak explain why the random walker returns to the origin for sure in one- or two-dimensional space, but it does not in higher-dimensional space. The process of proving is quite advanced but clear. First, clarify some events and probabilities, which are needed, to work with and denote them with different symbols. Second, find the relationship between ℓ_n and r_n, and find the relationship between $P(z)$ and $Q(z)$ based on the previous findings. Third, find the expression for loop generating function and exponential loop generating function, and the general rule $E_d(z) = E_1(z)^d$ can be found. Fourth, from the series representation of the modified Bessel function, it can be concluded that $E(z) = I_0(2z)^d$. Fifth, use the Borel transform to make the integral representation of $Q(z)$. Finally, use the Laplace principle to determine whether the integral representation converges or diverges so that a complete proof can be achieved. In general, this way of proof is complete and beautiful, but there are many hard theorems used in the procedure of proof (like monotone convergence theorem, modified Bessel equation, and so on). It might be complicated for new statistics or mathematics learner; therefore, in the future, it is hoped that there will be easier ways to prove the Pólya's Random Walk Theorem. Also, back to the probabilistic reasoning systems mentioned in the introduction, such systems can also be attempted to build to work on random walk problems. For example, computer codes can be written, and the general patterns of random walker's behaviors can be recorded (application of the Pólya's Random Walk Theorem can also be helpful while building the system). Eventually, desirable data might be obtained and applied to the real life.

References

[1] Avi Pfeffer. *Practical Probabilistic Programming*.
[2] Rick Durrett. *Probability Theory and Examples.* Fourth Edition.
[3] ZouYang, WuHecheng, Zhao Yingding, Jiang Yunzhi. Random walk recommendation algorithm based on multi-weight similarity [J]. Computer Application Research, 2020,37(11):3267-3270+3296.DOI:10.19734/j.issn. 1001-3695.2019.08.0275.
[4] Zou [1] Liao Yongxin. Research on the Joint Classification Algorithm of Heterogeneous Label Sets Based on Random Walk and Dynamic Label Propagation [D]. South China University of Technology, 2017.
[5] Lu Yuke. Research on Complex Code Recognition Technology [D]. University of Electronic Science and Technology of China, 2022. DOI: 10.27005/d.cnki.gdzku.2022.003580.
[6] Song Bin, Liu Lili, Zhang Lei, Wang Lei, Du Yuxin, Zhang Ning. Information Management of Coal Mine Equipment Based on Data Matrix Code [J]. Industrial and Mining Automation, 2020, 46(11): 83-86+94. DOI: 10.13272/ j.issn.1671-251x.2020050059.
[7] Yuan Tailing, Xu Kun. Length minimization algorithm of two-dimensional code bit stream [J]. Chinese Journal of Image and Graphics, 2022, 27(08): 2356-2367.
[8] GB/T 41208-2021, Data Matrix Code [S].
[9] Deng Huipeng, Wang Yi, Dong Xiaowen. Comparison of Hanxin code with PDF417, data matrix code (DM code), and QR code [J]. China Automatic Identification Technology, 2021(05):50-53.
[10] Guo Lijie. *Random walk and its application in leaf image segmentation [D]*. Lanzhou University, 2018.
[11] [Karatzas, I. and S. E. Shreve (1991). *Brownian Motion and Stochastic Calculus*. Springer-Verlag, New York.
[12] Lukacs, E. (1970). *Characteristic Functions*, 2nd Ed. Griffin.
[13] Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability*. Springer-Verlag, New York.
[14] Petrov, V. V. (1995). *Limit Theorems of Probability Theory*. Oxford University Press, Oxford.
[15] Samorodnitsky, G. and M. S. Taqqu. (1994). *Stable Non-Gaussian Random Processes*. Chapman-Hall/CRC,
[16] Shorack, G. R. (2000). *Probability for Statisticians*. Springer-Verlag, New York.
[17] Durrett, R. (2019). *Probability: theory and examples (Vol. 49).* Cambridge university press.
[18] Chung, K. L. (1967). *Markov Chains with Stationary Transition Probabilities*, 2nd Ed. Springer Verlag,