Data Process Language Midterm: RandomWalk Analysis

HyukSang Kwon

Department of Computer Engineering

제 I 절 INTRODUCTION

Random Walk란 확률론적 또는 무작위적 과정으로 알려진 수학적 대상으로, 임의로 움직이는 물체가 시작한 곳에서 멀리 떠돌아 다니는 과정입니다. Random Walk의 기본 예로는 0에서 시작하여 각 단계마다 똑같은 확률로 +1 또는 -1 이동하는 경우를 들 수 있습니다. 다른 예로 액체 또는 기체에서 확산하는 분자 추적 경로, 먹이를 동물 검색 경로, 주식 가격 변동 등이 있으며, 이들은 모두 Random Walk 모델에 의해 근사될 수 있습니다. Random Walk의 기본 사례인 0에서부터 시작하고 각 단계에서 동일한 확률로 +1 또는 -1 이동하는 정수 $\mathbb{U}$의 임의 행보인 1-Dimensional Random Walk의 평균, 표준 편차 그리고 diffusion equation은 다음과 같은 관계를 갖습니다.

1-Dimensional Walk에서 t번 시도된 모든 반복에 따른 위치를 L_t이라 할 때, 이에 대한 수식은

$$L_t = \sum_{i=1}^{t} U_i \qquad t = 1, 2, ... \tag{1}$$

where

$$p_U[k] = \begin{cases} \frac{1}{2} & k = -1 \\ \frac{1}{2} & k = 1 \end{cases} \tag{2}$$

입니다. U_i는 모두 독립이고 동일 분포를 가집니다. 수직선 상에 물체가 있다고 가정했을 때, U_i는 (i+1)번째에서의 이동이고 L_t은 물체의 변위라고 할 수 있습니다. 이 수식으로 1-Dimensional Walk의 평균과 분산을 다음과 같이 나타낼 수 있습니다.

$$E[L_t] = \sum_{i=0}^{t} E[U_i] = (t+1)E[U_0] = 0 \tag{3}$$

$$var(L_t) = \sum_{i=0}^{t} var[U_i] = (t+1)var(U_0) = t+1 \tag{4}$$

즉, 1-Dimensional Walk의 평균 반복 횟수에 관계없이 0인데 비하여, 분산은 선형적으로 증가합니다.

1-Dimensional Walk에서 물체가 t 시간에서 x 위치 있을 확률을 $p(x,t)$라고 할 때, 식 2를 만족하는 $p(x,t)$는 다음 확률 미분 방적식을 만족합니다.

$$p(x, t+\vec{\Delta}t) = \frac{1}{2}p(x-\vec{\Delta}x) + \frac{1}{2}p(x+\vec{\Delta}x) \tag{5}$$

여기서 양변에 $p(x,t)$를 빼면

$$\begin{aligned} &p(x, t+\vec{\Delta}t) - p(x,t) \\ &\quad = \frac{1}{2}((p(x-\vec{\Delta}x) - p(x,t)) - (p(x,t) - p(x+\vec{\Delta}x))) \end{aligned}$$

와 같습니다. 이 식에서 만약 반복 횟수가 무한대로 늘어나 $\vec{\Delta}x$와 $\vec{\Delta}t$가 한없이 0에 가까워 진다면, 좌변은 p의 시간(t)에 대한 미분 값에 근사하고, 우변은 p의 위치(x)에 대한 2차 미분 값에 수렴한다는 것을 알 수 있습니다.

$$p(x, t+\vec{\Delta}t) - p(x,t) \cong \vec{\Delta}t\frac{\partial p}{\partial t}$$
$$\frac{1}{2}(p(x-\vec{\Delta}x) + p(x+\vec{\Delta}x) - 2p(x,t)) \cong \vec{\Delta}x^2\frac{\partial^2 p}{\partial x^2}$$
$$where \quad \Delta t \to 0, \Delta x \to 0$$

해당 수식은

$$\frac{\partial p}{\partial t} = D\frac{\partial^2 p}{\partial x^2} \qquad (D = \frac{\vec{\Delta}x^2}{2\vec{\Delta}t}) \tag{6}$$

과 같은 Diffusion equation으로 정리할 수 있습니다. 결과적으로, 두 값은 상수인 $D(diffusionconstant)$의 곱 만큼 차이가 나는 비례 관계입니다.

해당 보고서는 위에서 살펴본 1-Dimensional Walk의 평균, 분산, Diffusion equation을 python으로 확인합니다. 이에 대한 자세한 접근법은 Section 2에 묘사됩니다. Section 3은 수집한 Random Walk 데이터를 평가하고 해석합니다. Section 4는 해당 보고서에 대한 결론을 기술합니다.

제 II 절 APPROACH

이 보고서는 1-Dimensional Random walk를 분석하며, 이 과정은 다음과 같은 총 세 가지 단계로 진행됩니다.

Algorithm 1 1-Dimensional Random Walk

```
procedure UpdateParticles(iterCnt, particles)
    Input: random process iteration count 'iterCnt'
    Input: location array 'particles'

    for i ← 0 to iterCnt do
        for j ← 0 to len(particles) do
            r ← random.rand()                    ▷ r = [0, 1)
            if r > 0.5 then particles[j] += 1
            if r <= 0.5 then particles[j] -= 1

```

1) **Step1**: 위치 정보(L_t)를 저장할 수 있는 0으로 초기화된 Particle을 총 1,000개 생성합니다. 그리고 한 번의 순환마다 모든 particle의 위치 값을 같은 확률로 +1 또는 -1 이동시키는 Random Walk를 1,000번 반복하도록 구현합니다. 이에 대한 자세한 내용은 Algorithm 1에 있는 pseudocode를 통하여 표현하였습니다.
2) **Step2**: Step 1에서 각 순환이 진행되는 동안 particle의 평균과 표준 편차를 계산한 후 결과를 분석합니다. 추가로 표준 편찻값, 이의 log 값, 제곱 값, 제곱근 값 등에 대하여 simple regression model을 이용하여 선형성을 확인하고, 표준 편차를 예측하는 수학 모델을 제시합니다.
3) **Step3**: Random Walk의 도달 공간[-500, 500]을 40개의 같은 크기의 구간으로 구분합니다. Step 1을 기반으로 10,000번 순환하며, 순환마다 구간별로 포함되는 particle 수를 카운팅($\mathbb{P}$)합니다. 그리고 $\mathbb{P}$에 대하여 시간의 미분 값($\frac{\partial P}{\partial t}$)과 위치의 이차 미분 값($D\frac{\partial^2 P}{\partial x^2}$)을 계산한 후 그래프로 나타냄으로써 관계를 알아보고자 합니다. 그리고 particle 수를 10,000, 100,000으로 변경하여 위와 같은 시뮬레이션을 반복한 후 변화를 포함한 전체적인 경향을 설명합니다.

Function	Correlation Coefficient
Standard Diviation	0.980
Log(Std)	0.872
Square(Std)	**0.999**
Sqrt(Std)	0.944

TABLE I

실제 값과 Regression 결과 사이의 correlation coefficient

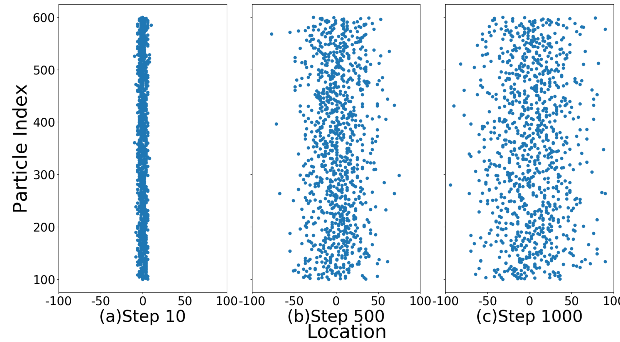

Fig. 1. 각각 10, 100, 1000번을 반복한 후 출력한 particle 값

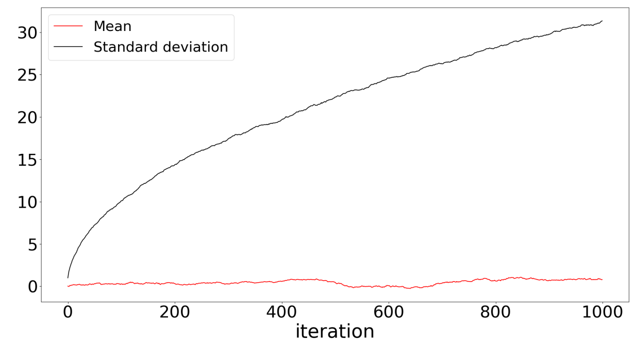

Fig. 2. Random Process를 1000번 실행한 1000개의 particle의 평균 및 분산

제 III 절 Analysis

Random Walk 평가를 위한 시뮬레이션은 파이썬 3.6.4 버전에서 실행되며, 각 Step에서 계산되는 particle들은 연계되지 않고 모두 독립적으로 수행됩니다. 또한 particle의 위치 이동을 위한 확률은 numpy 패키지에 존재하는 random.rand()함수를 사용했습니다.

그림 1은 Step 1을 실행하는 과정에서 특정 횟수에서 각 particle의 값을 표현한 그래프입니다. 그래프들은 각각 10회, 100회, 1000회의 반복 과정을 거친 후의 결과를 표시합

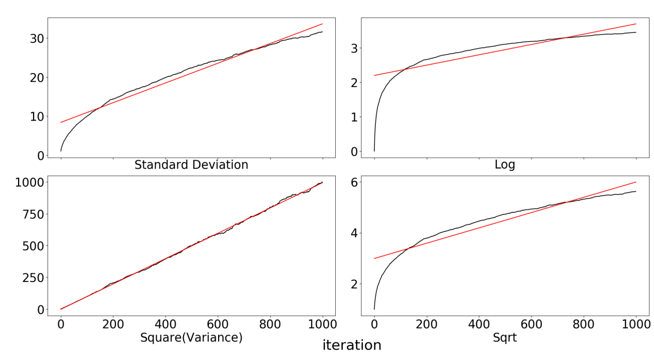

Fig. 3. Step 2의 결과로 나온 표준 편차를 이용해 다양한 형태의 Regression을 실행한 결과

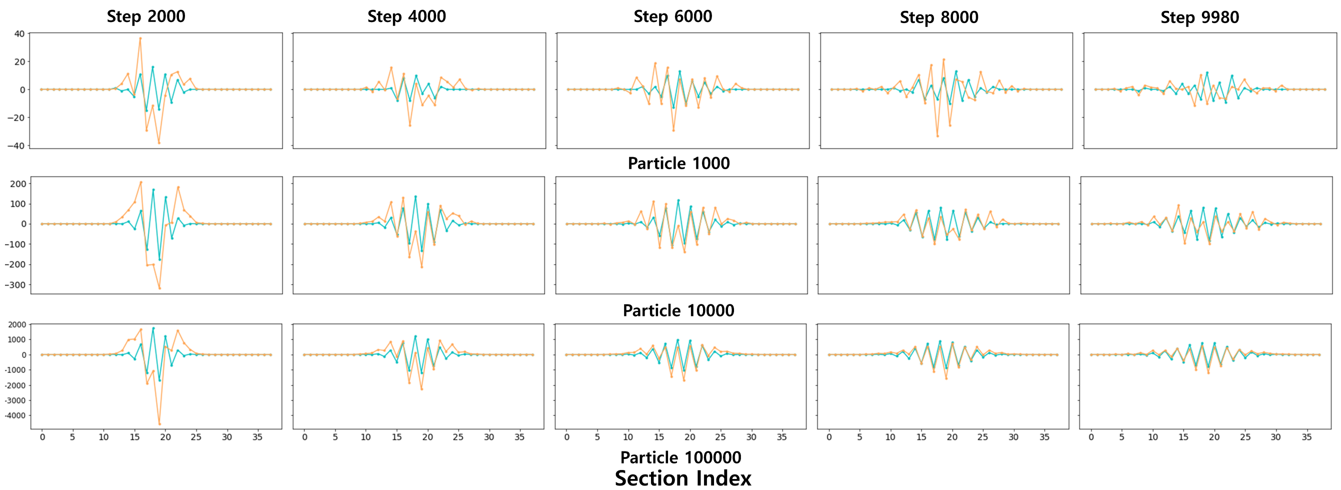

Fig. 4. $\frac{\partial P}{\partial t}$와$D\frac{\partial^2 P}{\partial x^2}$ 사이에 관계를 나타낸 그래프로 **하늘색**은 p_t, **주황색**은 $D * p_{xx}$를 표현합니다. 각 열은 그래프로 표현된 시점이며 행은 particle의 개수입니다.

니다. 10회 반복은 대부분의 값들이 0 주변에 몰려있는데 비해 그 횟수가 늘어날수록 particle의 분산이 커집니다.

Step 2의 방법으로 알아낸 실제 평균과 표준 편차를 그림 2에 나타납니다. 실제 평균과 표준 편차는 식 3와 식 4에서 증명한 바와 같이, 평균은 반복 횟수에 상관없이 0으로 수렴하고 표준 편차는 점점 증가합니다. 여기서 표준 편차, 이의 log, 제곱, 제곱근 등에 대한 선형성을 확인한 결과가 그림 3와 표 I입니다. 반복 횟수와 선형적 관계를 갖는 함수는 표준 편차의 제곱인 분산이 유일합니다. correlation coefficient를 계산한 표 I에서 가장 큰 연관 관계를 나타내는 값도 분산입니다. 이는 수식 4에서도 확인할 수 있으며 실제 regression 값의 기울기는 **0.996**으로 1에 굉장히 근사합니다. 결과적으로 표준 편차 예측 모델은

$$Standard_Deviation(iteration) = \sqrt{(iteration+1)} \quad (7)$$

으로 나타낼 수 있습니다.

마지막 Step 3에 관한 결과는 그림 4로 표현하였습니다. 해당 그래프는 D를 315로 설정한 후 p_x와 $p_x x$를 계산한 값들로, 반복 횟수와 particle 수에 따른 관계를 나타냅니다. 이미 수식 6에서 예상하였듯이 두 값들을 나타내는 선들의 모양이 유사합니다. 또한 두 값은 step과 particle의 커질 때마다 유사도가 늘어나는 것을 확인할 수 있습니다. step의 증가는 $\vec{\Delta} t$가 0으로 수렴시키므로 정확도가 높아지는 것이며, particle의 증가는 확률 분포가 정규 분포에 더욱 근사해지기 때문입니다.

제 IV 절 CONCLUSION

저는 수학적으로 증명된 1-Dimensional Random Walk의 평균, 표준 편차, Diffusion equation이 실제 적용되는지를 python을 통해 알아보았습니다. 1-Dimensional Random Walk 0으로 초기화된 값들을 같은 확률로 +1 또는 -1로 이동시켜 구현하였습니다. 그 결과, 평균과 표준 편차의 경우 수식과 일치하는 데이터들을 얻었습니다. Diffusion equation에서는 반복 횟수와 particle의 부족할 때 p_t와 p_{xx} 사이의 관계성을 발견하지 못했지만 그 수가 충분히 커질수록 유사도가 높아진다는 것을 확인하였습니다.