Random walks with hungry bacteria

David Wakeham

May 3, 2018

Prerequisites: basic probability, calculus, summation notation. Asterisks denote optional questions requiring more work or background knowledge.

Escherichia coli (E. coli) is a tiny, rod-shaped bacterium which can build a propeller and motor around in search of nutrients. We will follow a single E. coli, Colin, on his

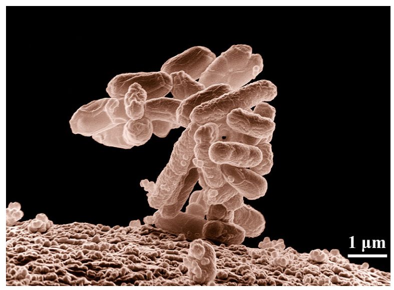

Figure 1: E. coli, magnified 10000 times under an electron microscope. Eric Erbe, USDA.

random foray around a one-dimensional environment. Let's model the environment as a one-dimensional lattice, with sites labelled by integers

$$\mathbb{Z} = \{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}.$$

Since space is discrete, it makes sense to represent time using discrete steps $t \in \mathbb{N} = \{0, 1, 2, 3, \ldots\}$. Colin and the lattice are pictured below.

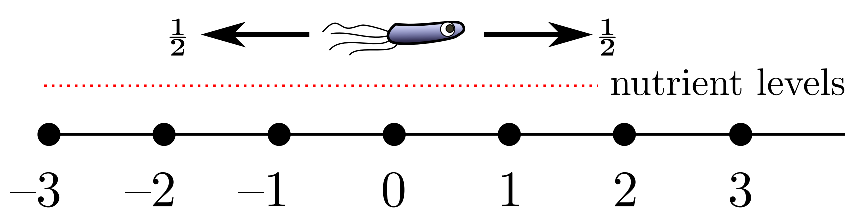

Figure 2: Colin randomly searching a lattice for nutrients. Nutrients are uniformly distributed.

Let $x(t)$ denote Colin's position at time t. If Colin wanders randomly, then at each time step t, the lattice position $x(t)$ is a *random variable*. We'll assume that Colin moves left or right with equal probability, a process called a *symmetric random walk*. This corresponds to a uniform nutrient density, where there is no preferred direction to explore. We'll also start Colin off at the origin with $x(0) = 0$.[1]

[1] To be more precise, the probability of starting at the origin is unity, $\Pr(x(0) = 0) = 1$.

It's useful to break the walk down into independent increments in position, or *jumps*, $X(t) = \pm 1$, where $X(t)$ is the jump in position between times $t-1$ and t. For a symmetric walk, $X(t) = X$ is a random variable taking values ± 1 with equal probability:

$$\Pr\big(X = \pm 1\big) = \frac{1}{2}. \tag{1}$$

In particular, for different times $t \neq t'$, $X(t)$ and $X(t')$ are independent random variables (IRVs) with the same probability distribution. We can express the jump in position as the difference between the current and previous position, $X(T) = x(T) - x(T-1)$. Alternatively, we can express Colin's position at time T as a sum of T jumps:

$$x(T) = X(1) + X(2) + \cdots + X(T) = \sum_{t=1}^{T} X(t).$$

Readers at home can simulate Colin's walk by flipping a coin at each time step! Heads is a step to the right, and tails a step to the left.

We'll now explore Colin's motion using three different approaches: central tendencies like mean and spread (Part A); the exact probability distribution (Part B); and large time behaviour (Part C).

Part A. We start with central tendencies. For any random variable Y (e.g. the jump X or the position x), the *expectation* of a function $f(Y)$ is the average value the function f takes over many realisations of Y. We can calculate this as a sum, weighted by probability:

$$\langle f(Y) \rangle := \sum_{y} \Pr(Y = y) f(y), \tag{2}$$

where y ranges over the values that Y can take. Simple examples are the *mean* $\langle Y \rangle$, and *variance* defined by

$$\mathrm{var}(Y) := \langle (Y - \langle Y \rangle)^2 \rangle = \langle Y^2 \rangle - \langle Y \rangle^2. \tag{3}$$

The square root of the variance is the *standard deviation* $\sigma(Y) := \sqrt{\mathrm{var}(Y)}$. Although we sometimes call the variance the "spread" of Y, the standard deviation is a better measure of spread, since it has the same units as Y, while the variance is measured in those units *squared*. For instance, if Y represents the number of bagels cooked at Barb's Bagelry each day, the mean and standard deviation are measured in bagels, but the variance in bagels squared.

We can view the angle brackets $\langle \cdot \rangle$ as an *operator*, or *function of functions*, which eats a function of Y and spits out the weighted sum:

$$\langle \cdot \rangle : \{\text{functions } f \text{ of } Y\} \to \{\text{weighted sums of } f\}.$$

It has the useful property of being *linear*:

$$\begin{aligned}\langle \alpha_1 f_1(Y) + \alpha_2 f_2(Y)\rangle &= \sum_y \Pr(Y=y)\Big[\alpha_1 f_1(Y) + \alpha_2 f_2(Y)\Big] \\ &= \alpha_1 \sum_y \Pr(Y=y) f_1(y) + \alpha_2 \sum_y \Pr(Y=y) f_2(y) \\ &= \alpha_1 \langle f_1(Y)\rangle + \alpha_2 \langle f_2(Y)\rangle.\end{aligned}$$

We can extend this to arbitrary linear combinations:

$$\left\langle \sum_{i=1}^n \alpha_i f_i(Y) \right\rangle = \sum_{i=1}^n \alpha_i \left\langle f_i(Y)\right\rangle. \tag{4}$$

In the next few exercises, you will apply these tools to Colin's walk.

(a) *Vanishing mean.* Show that $\langle X\rangle = 0$. By linearity, conclude that $\langle x(t)\rangle = 0$ for any t. On average, Colin doesn't move anywhere!

(b) *Expectations of IRVs factorises.* Random variables Y and Z are *independent* if the joint probability factorises:

$$\Pr(Y=y, Z=z) = \Pr(Y=y) \cdot \Pr(Z=z).$$

Argue that, for independent variables, the expectation also factorises, i.e. for any functions $f(Y), g(Z)$,

$$\langle f(Y) g(Z)\rangle = \langle f(Y)\rangle \langle g(Z)\rangle.$$

(c) *Variance of IRVs is additive.* Assuming that $Y_1, \ldots, Y_n$ are independent, use (b) to prove *Bienaymé's formula*,

$$\operatorname{var}\left(\sum_{i=1}^n Y_i\right) = \sum_{i=1}^n \operatorname{var}(Y_i). \tag{5}$$

In words, variance is additive for independent random variables. This is the advantage of talking about variance, rather than standard deviation. But don't forget Barb's Bagelry!

(d) *Variance equals time.* Show that the variance of the jump is 1:

$$\operatorname{var}(X) = \langle X^2\rangle = 1.$$

Deduce as a result that

$$\operatorname{var}[x(t)] = t.$$

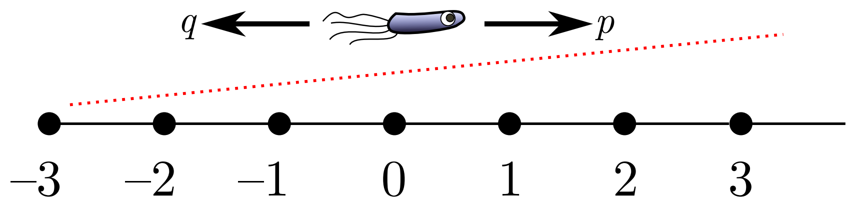

Figure 3: A uniform *gradient* in nutrients leads to a biased walk, since Colin tends to move towards high nutrient concentration.

(e) *Biased 1D walk.* Repeat the analysis for a *biased walk*, where

$$\Pr(X = +1) = p, \quad \Pr(X = -1) = 1 - p = q. \tag{6}$$

This corresponds to a uniform *gradient* in the nutrient density (see Figure 3). You should find that

$$\langle x(t) \rangle = (p - q)t, \quad \mathrm{var}[x(t)] = t[1 - (p - q)^2]. \tag{7}$$

It's almost like the mean of a biased walk moves with velocity $p - q$. Bias also shrinks the jump variance by $(p - q)^2$.

Part B. So far, we only have only looked at coarse measures of tendency. A more fine-grained measure is the *probability distribution*

$$f(x, t) := \Pr(x(t) = x). \tag{8}$$

This is the exact probability that Colin occupies a specific site x at time t.[2]

(a) *Left-right steps.* Suppose that after t steps, Colin makes ℓ steps to the left and r steps to the right. Argue that

$$r = \frac{1}{2}(t + x), \quad \ell = \frac{1}{2}(t - x). \tag{9}$$

(b) *Counting paths.* Show that, if Colin makes r right steps and ℓ left steps in time t, the number of ways $C(x, t)$ for him to arrive at his site is

$$C(x, t) = \binom{\ell + r}{\ell} = \binom{\ell + r}{r} = \frac{t!}{\left(\frac{1}{2}(t + x)!\right)\left(\frac{1}{2}(t - x)!\right)}. \tag{10}$$

Hint: Recall that $\binom{n}{k} = n!/k!(n - k)!$ is the number of ways of choosing k objects from a set of n, without regard to order.

[2]Note that we are using x without any time-dependence to denote a fixed lattice position, in contrast to the random variable $x(t)$.

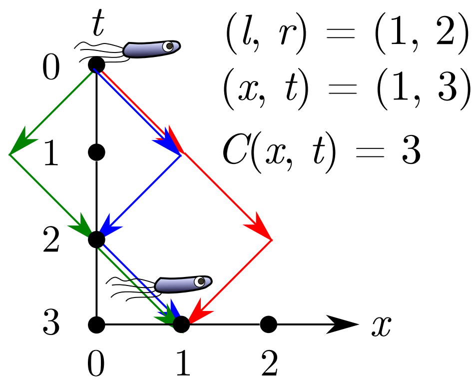

Figure 4: Three different paths (red, blue, green) from $x = 0$ to $x = 1$, consisting of one step to the left and two to the right.

(c) *Probability distribution.* Argue that the probability Colin occupies position x at time t is

$$f(x,t) = 2^{-t}C(x,t). \tag{11}$$

By convention, we set $C(x,t) = 0$ if $x + t$ is odd or $t > x$.

(d) *Biased walk distribution.* Show that for the biased walk of Part A, the probability distribution becomes

$$f_p(x,t) = p^r q^\ell C(x,t). \tag{12}$$

Here, r and ℓ can be expressed in terms of x and t using the dictionary in (a).

(e) *Book-keeping.** Check that for fixed t, the probabilities $f_p(x,t)$ for $x \in \mathbb{Z}$ add up to 1. *Hint*: Use the binomial expansion,

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k. \tag{13}$$

Part C. What happens when the number of steps t becomes very large? The position $x(t)$ is a sum of many copies of the jump variable X, and we can invoke the *Central Limit Theorem* (CLT). The CLT is a fundamental result in probability and statistics. It states that a sum of independent, identically distributed random variables (IIDs) can be approximated by a *Gaussian distribution*. We won't prove it; our more modest goal will be to state it precisely enough to use for the random walk.

Suppose Y is a random variable with mean μ and variance σ^2. Let S_N denote the average of N independent copies of Y, and $f_N(y)$ its probability distribution:

$$S_N := \frac{1}{N}\sum_{i=1}^{N} Y_i, \quad f_N(y) := \Pr(S_N = y). \tag{14}$$

The normal distribution with mean μ and variance σ^2 has distribution

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]. \tag{15}$$

Then the CLT states that, as $N \to \infty$, S_N converges to a normal distribution with mean μ and variance σ^2/N:

$$\lim_{N\to\infty} f_N(x) = f(x|\mu, \sigma^2 N^{-1}). \tag{16}$$

This explains why so many things in the real world are normally distributed: they are the result of adding together many independent, similarly distributed random variables.

(a) *Large random walks are Gaussian (CLT).* Use the CLT to show that, for a large number of steps $t \gg 1$, the probability distribution for the symmetric walk takes the Gaussian form

$$f(x, t) \approx \frac{1}{\sqrt{2\pi t}} \exp\left[-\frac{x^2}{2t}\right]. \tag{17}$$

In other words, for very large t, we can approximate Colin's position using a normal distribution centred at the origin. *Hint*: Consider $tS_t = x(t)$ in the CLT.

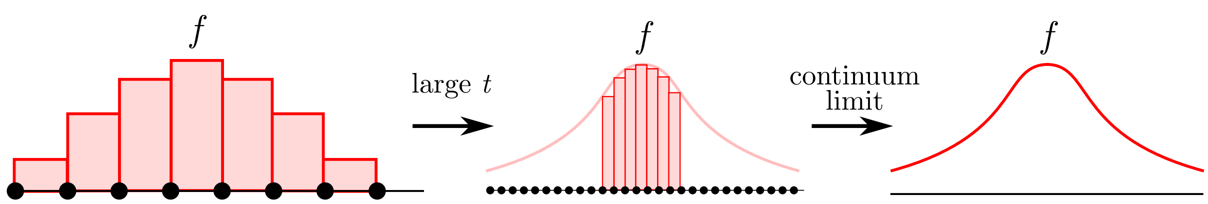

Figure 5: In the limit of a large number of steps, the probability distribution (at fixed t) becomes approximately Gaussian.

(b) *Large random walks are Gaussian (Stirling).** You can also derive the normal distribution directly from the exact probability distribution (11). This requires *Stirling's formula*, another large N result for approximating factorials:

$$N! \approx N^N e^{-N}\sqrt{2\pi N}. \tag{18}$$

Assuming that both ℓ and r are large, use Stirling's formula to rederive (17).

(c) *Higher dimensional random walk.** Imagine that Colin's counterpart, Colleen, randomly moves about on a d-dimensional "hypercubic" lattice

$$\mathbb{Z}^d = \{\mathbf{k} = (k_1, k_2, \ldots, k_d) \,|\, k_1, k_2, \ldots, k_d \in \mathbb{Z}\}.$$

At each time step, Colleen randomly chooses one of the d axes to move along, then jumps back or forth on unit in that direction with equal probability. Argue that, for large t, the probability distribution takes the form

$$f^{(D)}(\mathbf{k}, t) \approx \frac{1}{(2\pi t d^{-1})^{d/2}} \exp\left[-\frac{d|\mathbf{k}|^2}{2t}\right]. \tag{19}$$

Hint: Think of Colleen's d-dimensional walk as d one-dimensional walks.

(d) *Gaussian spread in higher dimensions.** Argue that at time t, Colleen is on average a distance $\sqrt{t}$ from the origin, whatever the dimension d! The moral is that the natural measure of Gaussian variance in d dimensions is σ^2/d rather than σ^2. *Hint*: Consider the average variance in each dimension and add them together.

(e) *Scaling the variance.* We can add many steps of unit size to get a normal distribution. But imagine zooming out so that the size of steps x_{step} and time between steps t_{step} both become small. If t continues to denote total elapsed time, the number of steps in that time will be t/t_{step}. Show that the variance of X and $x(t)$ become

$$\text{var}(X) = x_{\text{step}}^2, \quad \text{var}[x(t)] = t\left(\frac{x_{\text{step}}^2}{t_{\text{step}}}\right). \tag{20}$$

(f) *Scaling the distribution.* Let's apply our rescaling argument to the large random walk. Fix the ratio

$$\sigma^2 = \frac{x_{\text{step}}^2}{t_{\text{step}}}, \tag{21}$$

so that the variance is $t\sigma^2$. In this limit, derive the approximate distribution

$$f(x, t) \approx \frac{1}{\sqrt{2\pi\sigma^2 t}} \exp\left[-\frac{x^2}{2\sigma^2 t}\right]. \tag{22}$$

As long as we keep the ratio $\sigma^2 = x_{\text{step}}^2/t_{\text{step}}$ fixed, the distribution (22) is well-defined, even as we make x_{step} and t_{step} much smaller than our original step size. In particular, we could make them small enough to regard the probability distribution as *continuous* in time and space, rather than being defined on a lattice. This is sometimes called the *continuum limit*. In this limit, something cool happens: the function $f(x, t)$ obeys the *diffusion equation*.[3] You can check this in the following (optional) questions.

(g) *Initial condition.** Show that as t gets small, the probability distribution (22) gets infinitely high and thin. As we head towards $t \to 0$, Colin is increasingly likely to be found at the *origin*. The continuous distribution is trying valiantly to reproduce our initial condition $x(0) = 0$!

[3] In fact, since it applies to a probability distribution rather than a physical quantity, this is really the *Fokker-Planck equation*. But let's not split hairs.

(h) *Random walks diffuse.** Show that the continuous probability distribution (22) obeys the partial differential equation

$$\frac{\partial f(x,t)}{\partial t} = \sigma \frac{\partial^2 f(x,t)}{\partial x^2}. \tag{23}$$

This is called the *diffusion equation*. It describes a sharp spike (the initial condition from question (a)) smoothed into a spreading Gaussian as time evolves.

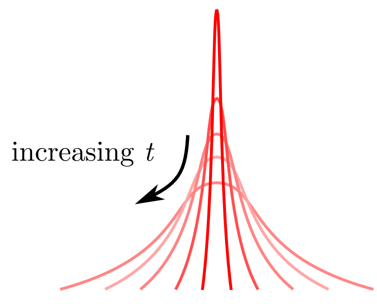

Figure 6: An initial spike smoothed by diffusion in the continuum limit of a random walk.

Many physical phenomena obey the diffusion equation (23), e.g. heat in a metal bar, loose atoms jiggling around a solid, or a drop of dye spreading in a glass of water. This means that if we have a whole swarm of (independently foraging) bacteria, they will spread like the drop of dye as they search for food. We can modify the diffusion equation to incorporate bias, and even *chemotaxis*, the tendency for bacteria to respond to local changes in nutrient concentration. Sadly, this is beyond our scope! But hopefully you see that, even in a simple one-dimensional model, there is a ballet in the random motion of hungry bacteria.