

*Introduction to Probability I: Expectations, Bayes  
Theorem, Gaussians, and the Poisson Distribution.*<sup>1</sup>

*Pankaj Mehta*

*April 9, 2020*

<sup>1</sup> Read: This will introduce some elementary ideas in probability theory that we will make use of repeatedly.

Stochasticity plays a major role in biology. In these next few lectures, we will develop the mathematical tools to treat stochasticity in biological systems. In this lecture, we will review some basic concepts in probability theory. We will start with some basic ideas about probability and define various moments and cumulants. We will then discuss how moments and cumulants can be calculated using generating functions and cumulant functions. We will then focus on two widely occurring “universal distributions”: the Gaussian and the Poisson distribution.

*Probability distributions of one variable*

Consider a probability distribution  $p(x)$  of a single variable where  $X$  can take discrete values in some set  $x \in \mathcal{A}$  or continuous values over the real numbers<sup>2</sup>. We can define the expectation value of a function of this random variable  $f(X)$  as

$$\langle f(X) \rangle = \sum_{x \in \mathcal{A}} p(x) \tag{1}$$

<sup>2</sup> We will try to denote random variables by capital letters and the values they can take by the corresponding lower case letter.

for the discrete case and

$$\langle f(X) \rangle = \int dx f(x) p(x), \tag{2}$$

in the continuous case. For simplicity, we will write things assuming  $x$  is discrete but it is straight forward to generalize to the continuous case by replacing by integrals.

We can define the  $n$ -th *moment* of  $x$  as

$$\langle X^n \rangle = \sum_x p(x) x^n. \tag{3}$$

The first moment is often called the mean and we will denote it as  $\langle x \rangle = \mu_x$ . We can also define the  $n$ -th *central* of  $x$  as

$$\langle (X - \langle X \rangle)^n \rangle = \sum_x p(x) (x - \langle x \rangle)^n. \tag{4}$$

The second central moments is often called the *variance* of the distribution and in many cases “characterizes” the typical width of the distribution. We will often denote the variance of a distribution by  $\sigma_x^2 = \langle (X - \langle X \rangle)^2 \rangle$ .

- **Exercise:** Show that the second central moment can also be written as  $\langle X^2 \rangle - \langle X \rangle^2$ .

*Example: Binomial Distribution*

Consider the Binomial distribution which describes the probability of getting  $n$  heads if one tosses a coin  $N$  times. Denote this random variable by  $X$ . Note that  $X$  can take on  $N + 1$  values  $n = 0, 1, \dots, N$ . If probability of getting a head on a coin toss is  $p$ , then it is clear that

$$p(n) = \binom{N}{n} p^n (1 - p)^{N-n} \tag{5}$$

Let us calculate the mean of this distribution. To do so, it is useful to define  $q = (1 - p)$ .

$$\begin{aligned}\langle X \rangle &= \sum_{n=0}^N np(n) \\ &= \sum_n n \binom{N}{n} p^n q^{N-n}\end{aligned}$$

Now we will use a very clever trick. We can formally treat the above expression as function of two independent variables  $p$  and  $q$ . Let us define

$$G(p, q) = \sum_n \binom{N}{n} p^n q^{N-n} = (p + q)^N \quad (6)$$

and take a partial derivative with respect to  $p$ . Formally, we know the above is the same as

$$\begin{aligned}\langle X \rangle &= \sum_n n \binom{N}{n} p^n q^{N-n} \\ &= p \partial_p G(p, q) \\ &= p \partial_p (p + q)^N \\ &= pN(p + q)^{N-1} = pN,\end{aligned} \quad (7)$$

where in the last line we have substituted  $q = 1 - p$ . This is exactly what we expect <sup>3</sup>. The mean number of heads is exactly equal to the number of times we toss the coin times the number probability we get a heads.

**Exercise:** Show that the variance of the Binomial distribution is just  $\sigma_n^2 = Np(1 - p)$ .

<sup>3</sup> When one first encounters this kind of formal trick, it seems like magic and cheating. However, as long as the probability distributions converge so that we can interchange sums/integrals with derivatives there is nothing wrong with this as strange as that seems.

### *Probability distribution of multiple variables*

In general, we will also consider probability distributions of  $M$  variables  $p(X_1, \dots, X_M)$ . We can get the *marginal distributions* of a single variable by integrating (summing over) all other variables <sup>4</sup> We have that

$$p(X_j) = \int dx_1 \dots dx_{j-1} dx_{j+1} \dots dx_M p(x_1, x_2, \dots, x_M). \quad (8)$$

Similarly, we can define the marginal distribution over any subset of variables by integrating over the remaining variables.

<sup>4</sup> In this section, we will use continuous notation as it is easier and more compact. Discrete variables can be treated by replacing integrals with sums.

We can also define the *conditional probability* of a variable  $x_i$  which encodes the probability that variable  $X_i$  takes on a give value given the values of all the other random variables. We denote this probability by  $p(X_i|X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_M)$ .

One important formula that we will make use of repeatedly is Bayes formula which relates marginals and conditionals to the full probability distribution

$$p(X_1, \dots, X_M) = p(X_i|x_1, \dots, X_{i-1}, X_{i+1}, \dots, X_M)p(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_M). \tag{9}$$

For the special case of two variables, this reduces to the formula

$$p(X, Y) = p(X|Y)p(Y) = p(Y|X)p(X). \tag{10}$$

This is one of the fundamental results in probability and we will make use of it extensively when discussing inference in our discussion of early vision.<sup>5</sup>

**Exercise: Understanding test results.**<sup>6</sup> Consider a mammogram to diagnose breast cancer. Consider the following facts (the numbers aren't exact but reasonable).

- 1% of women have breast cancer.
- Mammograms miss 20% of cancers when that are there.
- 10% of mammograms detect cancer even when it is *not* there.

Suppose you get a positive test result, what are the chances you have cancer?

**Exercise:** Consider a distribution of two variables  $p(x, y)$ . Define a variable  $Z = X + Y$ <sup>7</sup>

- Show that  $\langle Z \rangle = \langle X \rangle + \langle Y \rangle$
- Show that  $\sigma_z^2 = \langle Z^2 \rangle - \langle Z \rangle^2 = \sigma_x^2 + \sigma_y^2 + 2Cov(X, Y)$  where we have defined the covariance of  $X$  and  $Y$  as  $Cov(x, y) = \langle (X - \langle X \rangle)(Y - \langle Y \rangle) \rangle$ .

<sup>5</sup> Bayes formula is crucial to understanding probability and I hope you have encountered it before. If you have not, please do some exercises that use Bayes formula that are everywhere on the web.

<sup>6</sup> This is from the nice website <https://betterexplained.com/articles/an-intuitive-and-short-explanation-of-bayes-theorem/>

<sup>7</sup> We will make use of these results repeatedly. Please make sure you can derive them and understand them thoroughly.

- Show that if  $x$  and  $y$  are independent variables than  $Cov(x, y) = 0$  and  $\sigma_z^2 = \sigma_x^2 + \sigma_y^2$ .

*Law of Total Variance*

An important result we will make use of trying to understand experiments is the law of total variance. Consider two random variables  $X$  and  $Y$ . We would like to know how much the variance of  $Y$  is explained by the value of  $X$ . In other words, we would like to decompose the variance of  $Y$  into two terms: one term that captures how much of the variance of  $Y$  is “explained” by  $X$  and another term that captures the “unexplained variance”.

It will be helpful to define some notation to make the expectation value clearer

$$E_X[f(X)] = \langle X \rangle_x = \int dx p(x) f(x) \tag{11}$$

Notice that we can define the *conditional expectation* of a function  $Y$  given that  $X = x$  by

$$E_Y[f(Y)|X = x] = \int dy p(y|X = x) f(y). \tag{12}$$

For the special case when  $f(Y) = Y^2$  we can define the *conditional variance* which measures how much does  $Y$  vary is we fix the value of  $X$

$$Var_Y[Y|X = x] = E_Y[Y^2|X = x] - E_Y[Y|X = x]^2 \tag{13}$$

We can also define a different variance that measures the variance of the  $E_Y[Y|X]$  viewed as a random function of  $X$ :

$$Var_X[E_Y[Y|X]] = E_X[E_Y[Y|X]^2] - E_X[E_Y[Y|X]]^2. \tag{14}$$

This is a well defined probability distribution since  $p(y|x)$  is the marginal probability distribution over  $y$ . Notice now, that using Bayes rule we can write

$$\begin{aligned} E_Y[f(y)] &= \int dy p(y) f(y) = \int dx dy p(y|x) p(x) f(y) \\ &= \int dx p(x) \int dy p(y|x) f(y) = E_X[E_Y[f(Y)|X]] \end{aligned} \tag{15}$$

So now notice that we can write

$$\begin{aligned}
 \sigma_Y^2 &= E_Y[Y^2] - E_Y[Y]^2 \\
 &= E_X[E_Y[Y^2|X]] - E_X[E_Y[Y|X]]^2 \\
 &= E_X[E_Y[Y^2|X]] - E_X[E_Y[Y|X]^2] + E_X[E_Y[Y|X]^2] - E_X[E_Y[Y|X]]^2 \\
 &= E_X[E_Y[Y^2|X] - E_Y[Y|X]^2] + \text{Var}_X[E_Y[Y|X]] \\
 &= E_X[\text{Var}_Y[Y|X]] + \text{Var}_X[E_Y[Y|X]]
 \end{aligned} \tag{16}$$

The first term is often called the unexplained variance and the second term is the explained variance.

To see why, consider the case where the random variable  $Y$  are related by a linear function plus noise  $Y = aX + \epsilon$ , where  $\epsilon$  is a third random variable with mean  $\langle \epsilon \rangle = b$  and  $\langle \epsilon^2 \rangle = \sigma_\epsilon^2 + \mu^2$ . Furthermore, let  $E_X[X] = \mu_x$ . Then notice, that in this case all the randomness of  $Y$  is contained in  $\epsilon$ . Notice that

$$E[Y] = a\mu_x + b \tag{17}$$

and

$$E_\epsilon[Y|X] = aX + E_\epsilon[\epsilon] = aX + b \tag{18}$$

We can also write

$$\begin{aligned}
 \text{Var}_X[E_Y[Y|X]] &= E_X[E_Y[Y|X]^2] - E_X[E_Y[Y|X]]^2 \\
 &= E_X[(aX + b)^2] - (a\mu_x + b)^2 \\
 &= a^2 E_X[X^2] + 2ab\mu_x + b^2 - (a\mu_x + b)^2 \\
 &= a^2 \sigma_X^2
 \end{aligned} \tag{19}$$

This is clearly the variance in  $Y$  “explained” by the variance in  $X$ . Using the law of total variance, we see that in this case

$$E_X[\text{Var}_Y[Y|X]] = \sigma_Y^2 - a^2 \sigma_X^2 \tag{20}$$

is the unexplained variance.

### *Generating Functions and Cumulant Generating Functions*

We are now in the position to introduce some more machinery. These are the moment generating functions and cumulant generating functions. Consider a probability distribution  $p(x)$ . We would like an easy and efficient way to calculate the moments of this distribution. It turns out that there is an easy way to do this using ideas that will be familiar from Statistical Mechanics.

### Generating Functions

Given a discrete probability distribution,  $p(X)$  we can define an moment-generating function  $G(t)$  for the probability distribution as

$$G(t) = \langle e^{tX} \rangle = \int dx p(x) e^{tx} \quad (21)$$

Alternatively, we can also define an (ordinary) generating function for the distributions

$$G(z) = \langle Z^X \rangle = \sum_x p(x) z^x \quad (22)$$

If the probability distribution is continuous, we define the moment-generating function as <sup>8</sup>

$$G(t) = \langle e^{-tX} \rangle = \int dx p(x) e^{tx}. \quad (23)$$

<sup>8</sup> If  $x$  is defined over the reals, this is just the Laplace transform.

Notice that  $G(t = 0) = G(z = 1) = 1$  since probability distributions are normalized.

These functions have some really nice properties. Notice that the  $n - th$  moment of  $X$  can be calculated by taking the  $n$ -th partial derivative of the moment generating function evaluated at  $t = 0$ :

$$\langle X^N \rangle = \left. \frac{\partial^n G(t)}{\partial t^n} \right|_{t=0} \quad (24)$$

Alternatively, we can write in terms of the ordinary generating function

$$\langle X^N \rangle = (z\partial_z)^n G(z)|_{z=0} \quad (25)$$

where  $(z\partial_z)^n$  means apply this operator  $n$  times.

### Example: Generating Function of the Binomial Distribution

Let us return to the Binomial distribution above. The ordinary generating function is given by

$$G(z) = \sum_n \binom{N}{n} p^n q^{N-n} z^n = (pz + q)^N = (pz + (1 - p))^N \quad (26)$$

Let us calculate the mean using this

$$\mu_X = z\partial_z G(z) = [zNp(pz + (1 - p))^{N-1}]|_{z=1} = Np \quad (27)$$

We can also calculate the second moment

$$\begin{aligned} \langle X^2 \rangle &= z\partial_z(z\partial_z G(z)) \\ &= [zNp(pz + (1 - p))^{N-1}]|_{z=1} + \partial_z [Np(pz + (1 - p))^{N-1}]|_{z=1} \\ &= Np + N(N - 1)p^2 = Np(1 - p) + N^2p^2 \end{aligned} \quad (28)$$

and the variance

$$\sigma_X^2 = \langle X^2 \rangle - \langle X \rangle^2 = Np(1-p). \quad (29)$$

**Example: Normal Distribution**

We now consider a Normal Random Variable  $X$  whose probability density is given by

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \quad (30)$$

Let us calculate the moment generating function

$$G(t) = \int dx \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} e^{tx}. \quad (31)$$

It is useful to define a new variable  $u = (x - \mu)/\sigma$ <sup>9</sup>. In terms of this variable we have that

<sup>9</sup> Notice that  $u$  is just the “z-score” in statistics.

$$\begin{aligned} G(t) &= \frac{1}{\sqrt{2\pi}} \int du e^{-u^2/2+t(\sigma u+\mu)} \\ &= e^{\mu t+t^2\sigma^2/2} \frac{1}{\sqrt{2\pi}} \int du e^{-\frac{(u-\sigma t)^2}{2}} \\ &= e^{\mu t+t^2\sigma^2/2} \end{aligned} \quad (32)$$

**Exercise:** Show that the  $n - th$  moment of a Gaussian with mean zero ( $\mu = 0$ ) can be written as

$$\langle X^p \rangle = \begin{cases} 0 & \text{if } p \text{ is odd} \\ (p-1)!\sigma^p & \text{if } p \text{ is even} \end{cases} \quad (33)$$

This is one of the most important properties of the Gaussian/Normal distribution. All higher order moments can be expressed solely in terms of the mean and variance! We will make use of this again.

*The Cumulant Generating Function*

Another function we can define is the Cumulant generating function of a probability distribution

$$K(t) = \log \langle e^{tX} \rangle = \log G(t). \quad (34)$$

The  $n - th$  derivative of  $K(t)$  evaluated at  $t = 0$  is called the  $n$ -th cumulant:

$$\kappa_n = K^{(n)}(0). \quad (35)$$

Let us not look the first few cumulants. Notice that

$$\kappa_1 = \frac{G'(0)}{G(0)} = \frac{\mu_X}{1} = \mu_X. \quad (36)$$

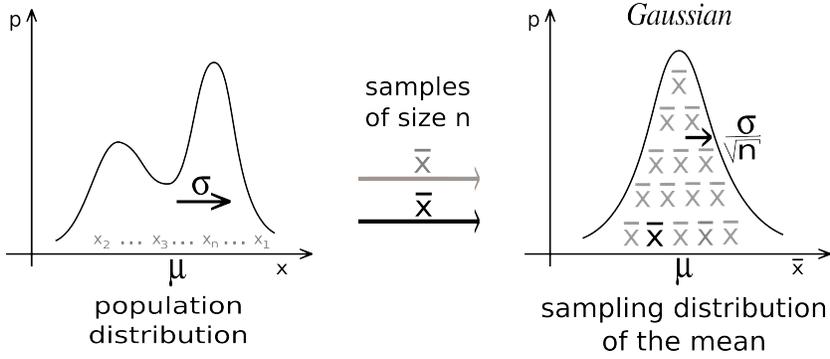


Figure 1: "Whatever the form of the population distribution, the sampling distribution tends to a Gaussian, and its dispersion is given by the Central Limit Theorem" (figure and caption from Wikipedia)

Similarly, it is easy to show that the second cumulant is just the variance:

$$\kappa_2 = \frac{G''(0)}{G(0)} - \frac{[G'(0)]^2}{G(0)^2} = \langle X^2 \rangle - \langle X \rangle^2 = \sigma_X^2 \quad (37)$$

Notice that the cumulant generating function is just the Free Energy in statistical mechanics and the moment-generating function is the partition function.

### Central Limit Theorem and the Normal Distribution

One of the most ubiquitous distributions that we see in all of physics, statistics, and biology is the Normal distribution. This is because of the Central Limit theorem. Suppose that we draw some random variables  $X_1, X_2, \dots, X_N$  identically and independently from a distribution with mean  $\mu$  and variance  $\sigma^2$ . Let us define a variable

$$S_N = \frac{X_1 + X_2 + \dots + X_N}{N}. \quad (38)$$

Then as  $N \rightarrow \infty$ , the distribution of  $S_N$  is well described by a Gaussian/Normal distribution with mean  $\mu$  and variance  $\sigma^2/N$ . We will denote such a normal distribution by  $\mathcal{N}(\mu, \sigma^2/N)$ . This is true regardless of the original distribution (see Figure 1). The main thing we should take away is that the variance decreases as  $1/N$  with the number of samples  $N$  that we take!

**Example: The Binomial Distribution** We can ask how the sample mean changes with the number of samples for a Bernoulli variable with probability  $p = 1/2$  for various  $N$ . This Figure 2 from Wikipedia.

### The Poisson Distribution

There is a second "universal" distribution that occurs often in Biology. This is the distribution that describes the number of rare events

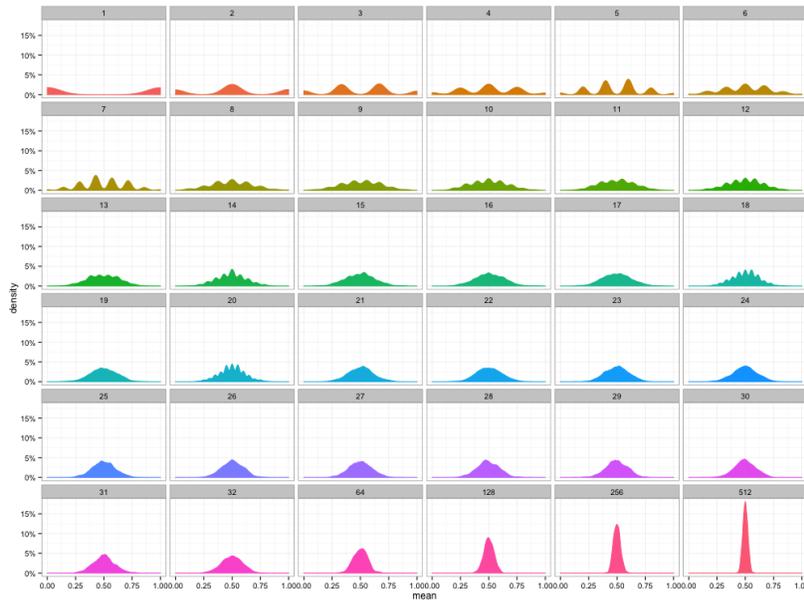


Figure 2: Means from various  $N$  samples drawn from Bernoulli distribution with  $p = 1/2$ . Notice it converges more and more to a Gaussian. (figure and caption from Wikipedia)

we expect in some time interval  $T$ . The Poisson distribution is applicable if the following assumptions hold:

- The number of events is discrete  $k = 0, 1, 2, \dots$
- The events are independent, cannot occur simultaneously, occur at a constant rate per unit time  $r$ .
- Generally the number of trials is large and probability of success is small.

Examples of things that can be described by the Poisson distribution include:

- Photons arriving in a microscope (especially at low intensity)
- The number of mutations per unit length of DNA
- The number of nuclear decays in a time interval
- "The number of soldiers killed by horse-kicks each year in each corps in the Prussian cavalry. This example was made famous by a book of Ladislaus Josephovich Bortkiewicz (1868D1931)" (from Wikipedia)
- "The targeting of V-1 flying bombs on London during World War II investigated by R. D. Clarke in 1946".

Let us denote the mean number of events that occur in a time  $t$  by  $\lambda = rT$ . Then, the probability that  $k$  events occurs is given by

$$p(k) = e^{-\lambda} \frac{\lambda^k}{k!}. \quad (39)$$

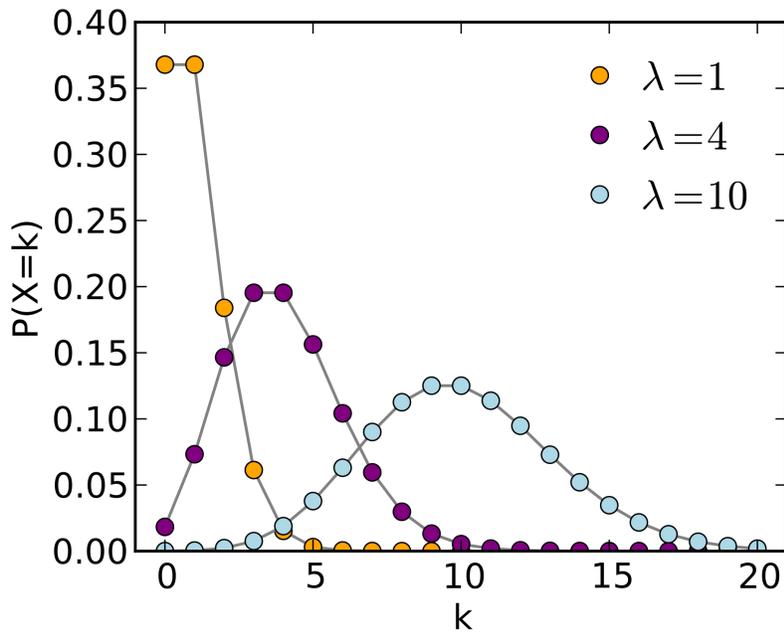


Figure 3: Examples of Poisson distribution for different means.. (figure from Wikipedia)

It is easy to check that  $\sum_k p(k) = 1$  using the Taylor series of the exponential. We can also calculate the mean and variance. Notice we can define the generating function of the Poisson as

$$G(t) = \sum_z e^{-\lambda} \frac{e^{tk} (\lambda)^k}{k!} = e^{\lambda(1+e^t)}. \tag{40}$$

The cumulant-generating function is just

$$K(t) = \log G(t) = \lambda(1 + e^t) \tag{41}$$

From this we can easily get the all cumulants since this is just differentiating the expression above  $n$  times

$$\kappa_n = K^{(n)}(0) = \lambda. \tag{42}$$

In other words, all the higher order cumulants of the Poisson distribution are the same and equal to the mean.

Another defining feature of the Poisson distribution is that it has “no memory”. Since things happen at a constant rate, there is no memory. We will see that in general to create non-Poisson distributions (with memory), we will have to burn energy! More on this cryptic statement later.

### Graph Theory

In this section, we will use some of the techniques above to think about graphs theory. To do so, we will first do some simple prob-

lems. We will start with some more properties about generating functions.

**Exercise:** Consider I toss two biased coins 10 times each (with probability of heads given by  $p_1, p_2$ ), for a total of 20 tosses. What is the probability that I will see 4 total heads out of the twenty tosses? Now, let us do this with generating functions. Consider the case

where the distribution of a property  $k$  of an object is characterized by a probability distribution  $p(k)$  and generating function

$$G_0(x) = \sum_{k=0}^{\infty} p_k x^k. \quad (43)$$

Argue that the distribution of the total of  $k$  summed over  $m$  independent realizations of the object is generated by the  $m^{\text{th}}$  power. Consider first the case where  $m = 2$ . Re-derive your result using the

generating functions.

**Exercise:** Consider that distribution  $q_k$  of the degree of vertices that we arrive at by following a randomly chosen edge at the graph starting at a random initial node. Argue that that  $q_k \propto kp_k$ . Show that the correctly normalized distribution is generated by

$$\frac{\sum x^k kp_k}{\sum_k kp_k} = \frac{\sum_k x^k kp_k}{\langle k \rangle} = x \frac{G'_0(x)}{G'_0(1)} \quad (44)$$

Show that if we start at a randomly chosen vertex and follow each of the edges at that vertex to reach the  $k$  nearest neighbors, then the vertices arrived at each have the distribution of remaining outgoing edges generated by this function, less one power of  $x$ , so that the distribution of outgoing edges is generated by the function above

$$G_1(x) = \frac{G'_0(x)}{G'_0(1)} = \frac{G'_0(x)}{\langle k \rangle} \quad (45)$$

**Exercise:** Show that for Binomial graph where each edge occurs with probability  $p$  that

$$G_0(x) = (1 - p + px)^N \quad (46)$$

Show that in the limit  $N \rightarrow \infty$  that this is just the generating function for the Poisson distribution

$$G_0(x) = e^{\langle k \rangle (x-1)} \quad (47)$$

**Exercise:** We can now calculate some really non-trivial properties. Consider the distribution of size of the connected components in the graphs. Let  $H_1(x)$  be the generating function for the distribution of size of components that are reached by choosing a random *edge* and following it to one of its ends. We will explicitly exclude the giant component if it exists (so below the percolation phase transition).

Argue that the chances of a component containing a closed loop of edges goes as  $N^{-1}$  (where  $N$  is the number of vertices), so that we can approximate each component as tree like.

Argue that this means that the distribution of components generated by  $H_1(x)$  can be represented graphically as in Figure 4; each component is tree-like in structure. If we denote by  $q_k$  the probability that

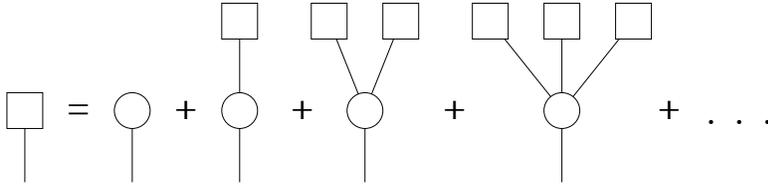


FIG. 3. Schematic representation of the sum rule for the connected component of vertices reached by following a randomly chosen edge. The probability of each such component (left-hand side) can be represented as the sum of the probabilities (right-hand side) of having only a single vertex, having a single vertex connected to one other component, or two other components, and so forth. The entire sum can be expressed in closed form as Eq. (26).

the initial state has  $k$  edges coming out of it other than the edge it came in along, argue that using the properties of generating functions discussed above that

$$H_1(x) = xq_0 + xq_1H_1(x) + xq_2[H_1(x)]^2 + \dots \quad (48)$$

Show that this gives the self consistency equation

$$H_1(x) = xG_1(H_1(x)) \quad (49)$$

Figure 4: Tree-like structure of components.

Argue that that if we start at a randomly chosen vertex, then we have one such component at the end of each edge leaving that vertex, and hence the generating function for the size of the whole component is

$$H_0(x) = xG_0(H_1(x)) \quad (50)$$

**Exercise:** Use these expression to examine the Poisson graph.