3

CHEMICAL KINETICS II: STOCHASTIC ASPECTS

In this [c1]chapter we delve into stochastic aspects of biomolecular reactions. [c2]Stochasticity is often important in biological systems because of the small number of molecules involved in fundamental biological processes. For example, when a protein is expressed in bacteria there are often only of the order of one to ten molecules of the corresponding mRNA. Since stochastic fluctuations typically scale as the square root of the number of molecules, stochastic effects can be extremely significant.

We start by discussing the relationship between diffusion and random walks in one dimension. These sections [c3]serve as a general introduction to the theory of stochastic processes. Diffusion of molecules would also be of importance when we take up formation of spatial pattern in the last chapter. [c4][c5]We then show how the techniques developed to study diffusion in space can be easily generalized to study stochastic fluctuations in molecule number.

3.1 Molecules in solution and diffusion

[c6] Molecules in a liquid solution are constantly bombarded by other molecules. If we track the trajectory of any single molecule over small time scales like miliiseconds, the trajectory would appear to be a random walk. Thus, it is fruitful to talk about the probability distribution of positions of individual molecules. We will derive the evolution equation for this probability distribution in this chapter. For completeness, some of the background material from probability theory is reviewed in the Appendix.

[c7]The evolution of the probability distribution of particles in a liquid is governed by the diffusion equation. When there are many particles of the same species, the density of particles satisfies the same evolution equation as the probability distribution of individual particle. The diffusion equation has the form

$$\frac{\partial}{\partial t}\rho(\vec{x},t) = -\vec{\nabla}\cdot\vec{j}(\vec{x},t) = D\nabla^2\rho(\vec{x},t), \tag{3.1}$$

[c1] *Pankaj*: ~~section~~

[c2] *Pankaj*: *Text added.*

[c3] *Pankaj*: ~~teach us about stochastic processes in general~~

[c4] *Pankaj*: ~~The preparation will be useful even when we study fluctuation in numbers (and not position of molecules)~~

[c5] *Pankaj*: *Text added.*

[c6] *Pankaj*: Fixed grammar and other small changes

[c7] *Pankaj*: *Text added.*

with $\rho(\vec{x},t)$ the density of particles at position$\vec{x}$ and time t and D the diffusion constant. The diffusion equation is valid for macroscopic length scales much larger than the mean-free path of the molecule under consideration. [c8] The diffusion equation is a consequence of Fick's law for the local current of particles

$$\vec{j}(\vec{x},t) = -D\vec{\nabla}\rho(\vec{x},t). \tag{3.2}$$

[c1] Fick's law and the diffusion equation are the mathematical statement that collisions between molecules tends to homogenize the particle density Typically, diffusion constants of molecules in solution depends upon its size. Bigger molecules like proteins have diffusion constants smaller than $1\mu\mathrm{m}^2\mathrm{s}^{-1}$, whereas for ions or small molecules, it could be of the order of $10^2\mu\mathrm{m}^2\mathrm{s}^{-1}$. We will explore the consequences of this equation and leave its derivation to the next section.

Armed with the diffusion equation, we return to the discussion in the last chapter on perfect enzymes and diffusion limited kinetics. It is shown in the appendix that a reaction, $\mathrm{A}+\mathrm{B}\rightarrow\mathrm{C}$ where the rate limiting step is the collision of the two types spherical molecules A and B (with radii R_A, R_B and diffusion constants D_A, D_B), the reaction rate is given by $4\pi(R_A+R_B)(D_A+D_B)[\mathrm{A}][\mathrm{B}]$. Note that in this formula the concentration is being expressed as number per meter cube. To use Molar (=mole per liter), we need a conversion factor of $1000N_{\mathrm{Av}}$ m^3 M^{-1}, N_{Av} being the Avogadro number ($\approx 6.023\times10^{23}$). In the context of enzymatic reactions where an enzyme E binds a substrate S, we can assume $a = (R_E + R_S)$ to be roughly the size of the reactive pocket of the enzyme. In cases where the substrate is a small molecule, its diffusion constant of the substrate will be much larger than that of the enyzme protein and therefore $D_{\mathrm{E}} + D_{\mathrm{S}} \approx D_{\mathrm{S}} \equiv D$.

To see why diffusion limited rates are in the same range for a wide variety of kinetically perfect enzymes, we need to develop a few more concepts. First, we need the Einstein relation between diffusion and temperatiure. Consider the generalization of equation 3.1 to the case where there is an external potential $U(\vec{x})$ exerting a force $\vec{F}(\vec{x}) = -\vec{\nabla}U(\vec{x})$ on each particle. Fick's law generalizes to

$$\vec{j}(\vec{x},t) = -D\vec{\nabla}\rho(\vec{x},t) - \mu\vec{F}(\vec{x})\rho(\vec{x},t), \tag{3.3}$$

where μ is the mobility and characterizes the average velocity of a particle in the solution in response to an applied force. Consequently, in the presence of an external potential the diffusion equation becomes

$$\frac{\partial}{\partial t}\rho(\vec{x},t) = -\vec{\nabla}\cdot\vec{j}(\vec{x},t) = \vec{\nabla}(D\vec{\nabla}\rho(\vec{x},t) + \mu\vec{F}(\vec{x})\rho(\vec{x},t)). \tag{3.4}$$

A steady-state solution to the diffusion equation (i.e. $\frac{\partial}{\partial t}\rho(\vec{x},t) = 0$) is obtained by having $\vec{j}(\vec{x},t) = 0$. This implies that

[c8] ~~T satisfies this equation if Fick's law for the current~~

[c1] *Pankaj: Text added.*

$$\rho(\vec{x},t) \sim \exp(-\frac{\mu}{D}U(\vec{x})). \tag{3.5}$$

. Comparing this to what is expected in thermal equilibrium, namely a Boltzmann distribution with $\rho(\vec{x},t) \sim \exp(-\frac{1}{k_B T}U(\vec{x}))$, where $k_B = 1.38 \times 10^{-23}$N m K^{-1} is the Boltzmann constant and T the temperature), yields the Einstein relation between the diffusion constant and the temperature,

$$D = \mu k_B T. \tag{3.6}$$

Second, we will need Stokes law. [c1]Stokes law relates the drag force experienced by a particle to the viscosity of the medium it is moving in. In particular, a particle of radius R moving in a medium of viscosity η with velocity $\vec{v}$ experiences a drag force of $\vec{F}_{\rm drag} = -6\pi\eta R\vec{v}$. Thus, in the presence of an applied force $\vec{F}$ particles settle down to a terminal velocity $\vec{v}$ determined by the no force condition

$$\vec{F} + \vec{F}_{\rm drag} = \mu^{-1}\vec{v} - 6\pi\eta R\vec{v} = 0. \tag{3.7}$$

This yields a relationship between the motility and the viscosity $\mu = (6\pi\eta R)^{-1}$, and through the Einstein relation the diffusion constant and viscosity, $D = k_B T/(6\pi\eta R)$.

Returning to enzyme substrate reaction, the relationships derived above imply that for kinetically perfect enzymes,

$$\frac{k_{\rm cat}}{K_M} \approx 4\pi a D = \frac{2k_B T a}{3\eta R} \approx 1.7 \times 10^9 \frac{a}{R}{\rm M}^{-1}{\rm s}^{-1}, \tag{3.8}$$

where we have used the fact that the viscosity of water is about 1centiPoise ($= 10^{-3}$Nsm^{-2}), $T \approx 300$K at room temperature, the conversion factor of $1000 N_{\rm Av}$m^{-3} M^{-1}). Now, if we make the reasonable assumption that a and R are of the same order of magnitude, we see why enzymes with $k_{\rm cat}/K_M$ in the ranges above are considered kinetically perfect.

3.2 Random walk

[c2] Let us now focus on describing random walks of single molecules. For mathematical simplicity, we start with a particle moving in one dimension on a lattice ("Be wise, discretize!" as M. Kac is supposed to have advised). The lattice sites are are specified by integers. Assume that at every discrete time step a particle moves either to the left or to the right with equal probability. If the particle starts, say at position 0, what is the probability distribution of its positions after N such time steps?

A random walk can be described as an N letter word of the form "LRLRRRL...R" where a L or R at the i-th position indicates that the i-th move is

[c1] *Pankaj: Text added.*

[c2] *Pankaj*: Mostly minor stylistic/grammar tweaks here

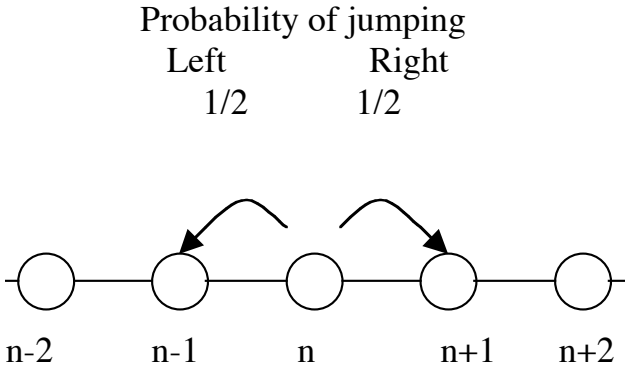

FIG. 3.1. Random walk on a lattice

to the left or to the right, respectively. If after N steps, we have l left moves and $r = N - l$ right moves, our total displacement is $q = r - l = N - 2l$. The number of moves with a fixed l (and, therefore, a fixed displacement q) is the combinatorial factor ${}^{N}C_{l} = N!/(l!(N-l)!)$. Since there are a total of 2^N possible walks, the probability of having a displacement q after N steps is just

$$P_{q,N} = \frac{{}^{N}C_{l}}{2^N} = \frac{N!}{l!r!}\frac{1}{2^N} = \frac{N!}{(\frac{N+q}{2})!(\frac{N-q}{2})!}\frac{1}{2^N} \tag{3.9}$$

Note that q only takes odd or even values depending upon whether number of steps N is odd or even.

To simplify this equation, we make use of the Stirling approximation to the factorial for large M,

$$\ln M! = M\ln M - M + \frac{1}{2}\ln(2\pi M) + O(\frac{1}{M}) \tag{3.10}$$

and $1 \gg m \gg M$,

$$\begin{aligned}\ln(M+m)! &= (M+m)\ln(M+m) - (M+m) + \frac{1}{2}\ln(2\pi(M+m)) + O(\frac{1}{M}) \\ &= M\ln M - M + \frac{1}{2}\ln(2\pi M) + m\ln M + \frac{1}{2}\frac{m^2}{M} + O(\frac{m}{M}).\end{aligned} \tag{3.11}$$

Using the equations 3.10 and 3.11 to expand the factorial in Eq. **??** and noting $1 \ll q \ll N$, we have

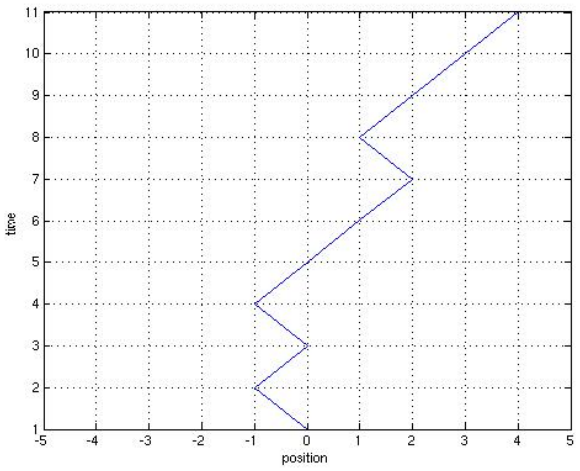

FIG. 3.2. An example of a random walk

$$P_{q,N} = \frac{N!}{(\frac{N+q}{2})!(\frac{N+q}{2})!}\frac{1}{2^N} \approx \frac{1}{\sqrt{2\pi N}} e^{-\frac{q^2}{2N}} \times 2 \tag{3.12}$$

This is just the gaussian approximation to a binomial distribution. Thus,

$$\begin{aligned} \text{Prob}[a \le q \le b] &= \sum_{q\in\{a,a+2,\ldots,b-2,b\}} P_{q,N} \\ &\approx \sum_{q\in\{a,a+2,\ldots,b-2,b\}} \frac{1}{\sqrt{2\pi N}} e^{-\frac{q^2}{2N}} \times 2 \\ &\approx \int_a^b \frac{dq}{\sqrt{2\pi N}} e^{-\frac{q^2}{2N}}, \end{aligned} \tag{3.13}$$

where we have used the fact that as the lattice size goes to zero sums can be replaced by integrals and changing l by 1 changes q by a factor of 2. Coming back to continuous space and time, let us have the lattice spacing to be Δx and time steps to be Δt. Then $x = q\Delta x$ and $t = N\Delta t$. The approximate probability distribution of x is then written as

$$p(x,t) \approx \frac{1}{\sqrt{2\pi N}} e^{-\frac{q^2}{2N}} \frac{dq}{dx} = \frac{1}{\sqrt{4\pi Dt}} e^{-\frac{x^2}{4Dt}}, \tag{3.14}$$

where we have defined the diffusion constant $D = \Delta x^2/(2\Delta t)$. To see that this definition is consistent with the usual definition of D, we can directly derive the evolution equation for $p(x,t)$ starting with the recursion relation

$$P_{q,N+1} = \frac{1}{2}[P_{q-1,N} + P_{q+1,N}]. \tag{3.15}$$

Since $P_{q,N}$ is proportional to $p(x,t)$ for $x = q\Delta x$ and $t = N\Delta t$

$$p(x, t + \Delta t) = \frac{1}{2}[p(x - \Delta x, t) + p(x + \Delta x, t)] \tag{3.16}$$

or

$$\begin{aligned} p(x,t) + \Delta t \partial_t p(x,t) + O(\Delta t^2) &= \frac{1}{2}[p(x - \Delta x, t + \Delta t) + p(x + \Delta x, t + \Delta t)] \\ &= \frac{1}{2}[p(x,t) - \Delta x \partial_x p(x,t) + \frac{1}{2}\Delta x^2 \partial_x^2 p(x,t) \\ &\quad + p(x,t) + \Delta x \partial_x p(x,t) + \frac{1}{2}\Delta x^2 \partial_x^2 p(x,t)] + O(\Delta x^4) \\ &= p(x,t) + \frac{1}{2}\Delta x^2 \partial_x^2 p(x,t) + O(\Delta x^4) \end{aligned} \tag{3.17}$$

implying

$$\partial_t p(x,t) \approx \frac{\Delta x^2}{2\Delta t}\partial_x^2 p(x,t) = D\partial_x^2 p(x,t). \tag{3.18}$$

This is known as the Fokker-Planck equation.

If we have many particles with positions x_i, the average density $\rho(x,t) = \sum_i \langle \delta(x - x_i(t)) \rangle = p(x,t) \times$ Number of particles (see the Appendix for a discussion of the delta function). Thus $\rho(x,t)$ also satisfies equation 3.18. Thus we derive the diffusion equation, namely equation 3.1, in one dimension, with D identified as the diffusion constant. We leave the generalization to higher dimensions as an exercise.

Excercise: Check that $p(x,t) = \exp(-x^2/(4\pi Dt))/\sqrt{4\pi Dt}$ is the solution of equation 3.18 with the initial condition $p(x,0) = \delta(x)$.

Excercise: Consider that case of a bias one-dimensional random walker where a particle can hop to the left with probability $1/2+b$ and to the right with $1/2-b$, with $0 < b < 1/2$.
a) Write down the evolution equation for $P_{q,N+1}$ and $p(x,t)$.
b) Derive the appropriate diffusion equation.

3.3 Langevin equation

Since diffusion of a particle is such an important process, it worth solving the problem in more than one way. [c1] In this section, we resolve the diffusion equation using Langevin equations. The main purpose for introducing the Langevin equation is to familiarize you with the mathematics needed for analyzing fluctuations in chemical kinetics.

[c1] *Pankaj*: ~~Solving the Langevin equation, which will be introduced in this section, is one of them.~~

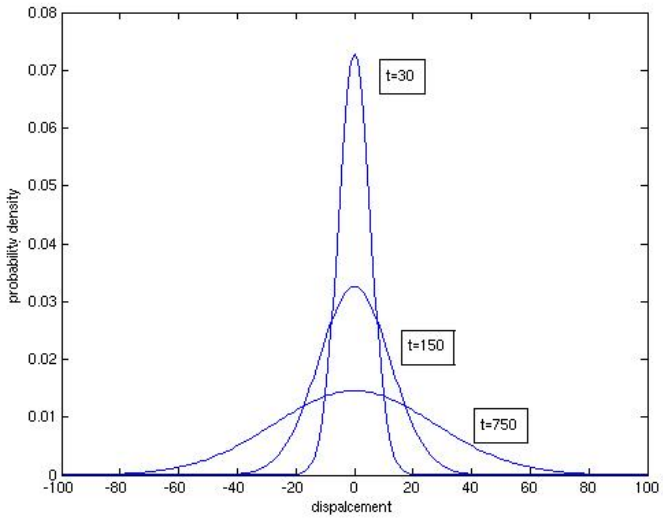

FIG. 3.3. Gaussian distribution for different values of time.

[c2]To derive the Langevin equation, it is useful to start once again with a diffusing particle on a lattice. Like before, at each time step the particle can hop to the left or right with equal probability. We start by Introducing the variables $\sigma_i = \pm 1$, with $\sigma_i = -1$ if the i-th move was to the left and $\sigma_i = 1$ if it was to the right. The σ_i can be viewed as independent random variables with equal probability of being ± 1. Consequently, the expectation value of σ_i is zero, $\langle \sigma_i \rangle = 0$, and [c3] $\langle \sigma_i \sigma_j \rangle = \delta_{i,j}$, where the Kronecker delta is defined as follows:

$$\begin{aligned} \delta_{i,j} &= 1 \text{ when } i = j, \\ &= 0 \text{ otherwise.} \end{aligned} \tag{3.19}$$

It is easy to calculate the first two moments of q_N using these relations.

$$\langle q_N \rangle = \sum_i \langle \sigma_i \rangle = 0 \text{ and } \langle q_N^2 \rangle = \sum_{i,j} \langle \sigma_i \sigma_j \rangle = \sum_{i,j} \delta_{i,j} = N. \tag{3.20}$$

In the continuum limit, N is large and the central limit theorem tells us that the distribution of q_N is going to be approximately gaussian. Since the mean and the variance of a gaussian distribution is enough to specify it, knowing first two

[c2] *Pankaj: Text added.*

[c3] *Pankaj:* ~~Going back to the discrete representation, one could write the displacement after N steps to be $q_N = \sum_{i=1}^{N} \sigma_i$, where each $\sigma_i = \pm 1$, is an independent variable. If the i-th move was to the left, $\sigma_i = -1$ and if it was to the right $\sigma_i = 1$. For unbiased random walks (equal probability of going to the left or to the right), $< \sigma_i >= 0$. The independence leads to~~

moments of q_N is good enough for many applications. For example it tells us that the probability density of q_N is $\exp(-q_N^2/(2N))/\sqrt{2\pi N}$, which in terms of continuum variables x lead to the desired result $p(x,t) = \exp(-x^2/(4\pi Dt))/\sqrt{4\pi Dt}$.

Can we do this directly in the continuum representation? Yes. We strat with the recursion formula

$$q_N - q_{N-1} = \sigma_N. \tag{3.21}$$

Multiplying by $\Delta x/\Delta t$ and noting that in the continuum limist $x(t) = \Delta x q_N$ and $t = N\Delta t$ yields

$$\begin{aligned} \text{implies} \quad & \frac{x(t) - x(t-\Delta t)}{\Delta t} = \eta(t) = \frac{\Delta x}{\Delta t}\sigma_N \\ \text{or} \quad & \frac{dx(t)}{dt} = \eta(t) \end{aligned} \tag{3.22}$$

with the conditions $\langle \eta(t) \rangle = 0$ and

$$\langle \eta(t)\eta(t') \rangle = \left(\frac{\Delta x}{\Delta t}\right)^2 \langle \sigma_N \sigma_{N'} \rangle = \frac{2D}{\Delta t}\delta_{N,N'} = 2D\delta(t-t'), \tag{3.23}$$

If you haven't seen this kind of jugglery before, you may be understandably uncomfortable. Just to see how to apply (3.22), let's calculate the variance of $x(t)$:

$$\begin{aligned} \langle x(t)^2 \rangle &= \langle \int_0^t ds\eta(s) \int_0^t ds'\eta(s') \rangle \\ &= \int_0^t ds \int_0^t ds' \langle \eta(s)\eta(s') \rangle \\ &= 2D \int_0^t ds \int_0^t ds' \delta(s-s') \\ &= 2D \int_0^t ds = 2Dt. \end{aligned} \tag{3.24}$$

This is indeed the right answer, with the square of distance travelled being proportional to time elapsed times the diffusion equation. Note that these manipulations are just the continuum version of the discrete calculation in the equation 3.20.

[c1]

Generalization of this equation to higher dimension in isotropic medium is obvious: $\dot{\vec{x}}(t) = \vec{\eta(t)}$ with $\langle \vec{\eta}(t) \rangle = 0$ and $\langle \eta_i(t)\eta_j(t') \rangle = 2D\delta_{i,j}\delta(t-t')$. Each coordinate behaves as an independent variable. If we start at the origin, that

[c1] *Pankaj*: Made some minor tweaks below and checked Fourier transform conventions.

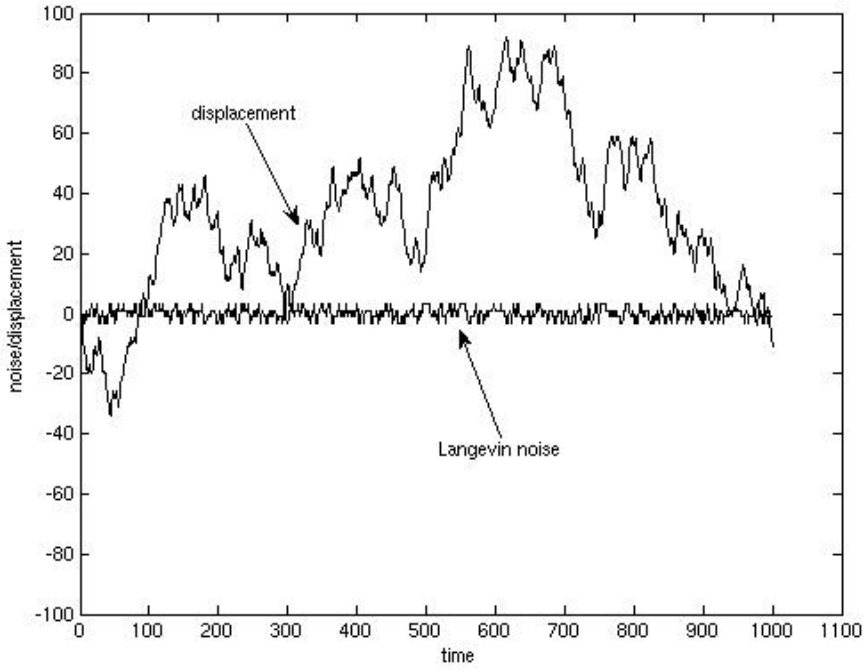

FIG. 3.4. Displacement and the Langevin noise in random walk.

is $p(\vec{x},0) = \delta^{(d)}(\vec{x}) = \prod_i \delta(x_i)$, then after time t the probability distribution is given by

$$p(\vec{x},t) = \prod_i \frac{1}{\sqrt{4\pi Dt}} e^{-\frac{x_i^2}{4Dt}} = \frac{1}{(4\pi Dt)^{d/2}} e^{-\frac{\vec{x}^2}{4Dt}} \tag{3.25}$$

What happens to the Langevin equation in presence of an external force? You might guess that now we have a biased random walk where the probability of going left or right are not equal. Such biased walk can be represented by $dx(t)/dt$ having an average part proportional to the bias. Remembering that in a viscous medium, the average velocity of a particle is proportional to applied force, and the proportional constant is mobility μ, we have

$$\frac{d\vec{x}(t)}{dt} = \mu\vec{F}(\vec{x}) + \vec{\eta}(t) = -\mu\vec{\nabla}U(\vec{x}) + \vec{\eta}(t) \tag{3.26}$$

With the statistics of $\vec{\eta}(t)$ as before.

It is possible to show that equation 3.28 is equivalent to the Fokker-Planck equation for the probability density (equation 3.4 with $\rho(\vec{x},t)$ replaced by $p(\vec{x},t)$). Why, then, do we bother to write down the alternate Langevin formulation? It

turns out that some questions that relate to kinetics are often easier to answer starting from the Langevin formulation.

For example, in a one dimensional harmonic potential $U(x) = Kx^2/2$, we know that the equilibrium probability distribution $p_0(x)$ is given by

$$p_0(x) = \frac{\exp(-\frac{\beta K x^2}{2})}{\sqrt{\pi \beta K}} \text{ where } \beta = 1/(k_B T). \tag{3.27}$$

This result is derived easily from the Fokker-Planck equation by setting $\vec{j}(x,t) = 0$ in Equation 3.1. However if we want to calculate equilibrium time lagged correlation of the positions of the particle, the Langevin equation is a better starting point,

$$\frac{dx(t)}{dt} = -\mu K x(t) + \eta(t). \tag{3.28}$$

Understanding stochastic differential equations like this is important. The $-\mu K x$ term is the restoring force trying to bring the particle to the potential minimum, at $x = 0$, while the noise term $\eta(t)$ kicks it in random directions. The time constant for deterministic dynamics is given by $\tau = (\mu K)^{-1}$. This is just a linear equation with an inhomogeneous term. To solve this equation we rewrite it as

$$\frac{d}{dt}(e^{\frac{t}{\tau}} x(t)) = e^{\frac{t}{\tau}} \eta(t) \tag{3.29}$$

and integrate to get

$$x(t) = e^{-\frac{t}{\tau}} x(0) + \int_0^t dt' e^{-\frac{t-t'}{\tau}} \eta(t') \tag{3.30}$$

For times t much larger than τ the effect of the initial value if forgotten and the position is determined mostly by noise in the immediate past. Thus, in practice we make the approximation

$$x(t) \approx \int_{-\infty}^t dt' e^{-\frac{t-t'}{\tau}} \eta(t'). \tag{3.31}$$

The lower limit of the integral can be changed because the contribution from $|t - t'| >> \tau$ is negligible due to the exponential surpression.

We are interested in the correlation between $x(s)$ and $x(s + t)$, where s is arbitrary and time lag t is taken to be positive without loss of generality. The correlation function $C(t)$ then calculated as

$$\begin{aligned}\langle x(s)x(s+t)\rangle &\approx \langle \int_{-\infty}^{s} dt' e^{-\frac{s-t'}{\tau}}\eta(t') \int_{-\infty}^{s+t} dt'' e^{-\frac{t+s-t''}{\tau}}\eta(t'')\rangle \\ &= \int_{-\infty}^{s} dt' \int_{-\infty}^{s+t} dt'' e^{-\frac{s-t'}{\tau}} e^{-\frac{t+s-t''}{\tau}} \langle \eta(t')\eta(t'')\rangle \\ &= 2D \int_{-\infty}^{s} dt' \int_{-\infty}^{s+t} dt'' e^{-\frac{s-t'}{\tau}} e^{-\frac{t+s-t''}{\tau}} \delta(t'-t'') \\ &= 2De^{-\frac{t}{\tau}} \int_{-\infty}^{s} dt' e^{-\frac{2(s-t')}{\tau}} = D\tau e^{-\frac{t}{\tau}} = \frac{k_B T}{K} e^{-\frac{t}{\tau}} \end{aligned} \quad (3.32)$$

Note that the correlation dies exponentially with a time constant set by τ.

Excercise: Show that $C(0) = \langle x(s)^2\rangle = \frac{k_B T}{K} = (\beta K)^{-1}$ is consistent with the equilibrium distribution given by Equation (3.27).

A more sleek calculation can be performed via a related quantity: the power spectrum[c0]. If we take Fourier transforms (see Appendix **CHECK** $2\pi\delta$ **as convention**) we get

$$-i\omega\hat{x}(\omega) = -\mu K\hat{x}(\omega) + \hat{\eta}(\omega) \Rightarrow \hat{x}(\omega) = \frac{\hat{\eta}(\omega)}{-i\omega + \mu K}. \quad (3.33)$$

Excercise: Show that that $\langle \hat{\eta}(\omega)\hat{\eta}(\omega')^*\rangle = 4\pi D\delta(\omega - \omega')$.

Thus,

$$\langle \hat{x}(\omega)\hat{x}(\omega')^*\rangle = \frac{4\pi D}{\omega^2 + (\mu K)^2}\delta(\omega - \omega') \quad (3.34)$$

The power spectrum, $N(\omega)$ is defined as the coefficient of the delta function above, and is

$$N(\omega) = \frac{2D}{\omega^2 + (\mu K)^2} = \frac{2k_B T}{K}\frac{\tau}{1 + (\omega\tau)^2}. \quad (3.35)$$

Excercise: Show that the correlation function $C(t)$ is the Fourier transform of the power spectrum $N(\omega)$. Exploit the fact that $x(t)$ is real.

Since $C(t)$ is just the Fourier transform of power spectrum $N(\omega)$, one has

$$C(t) = \int \frac{d\omega}{2\pi} N(\omega) e^{i\omega t} = \frac{2k_B T}{K} \int \frac{d\omega}{2\pi} \frac{\tau e^{i\omega t}}{1 + (\omega\tau)^2} = \frac{k_B T}{K} e^{-\frac{t}{\tau}} \quad (3.36)$$

3.4 Poisson arrivals

Having our first taste of random processes in the context of diffusion, let us move on and discuss how we analyze stochastic aspects of chemical reaction. Let us

[c0] The importance of power spectrum becomes clear as we go on. See Appendix for definition

start with the simplest example. Imagine a molecule is synthesized at a certain rate R. It could be a small molecule made by an enzyme or could be RNA being made from a gene. Let us consider the case where the time required to make the molecule is small compared to the typical time lag between the synthesis of two consecutive molecules. In this limit we can consider synthesis as a point process: random events that happen at a well-defined time, [c1]at some fixed rate, R, per unit time.[c2] [c3]Since the probability of synthesizing a molecule per unit time is small, the production can be modeled as a Poisson process.

What is the distribution of the lag time t between the two consecutive events? If $P(t)$ is the probability of nothing happening in the time interval $[0, t]$, then $P(t + \Delta t)$ is the probability of nothing happening in the time interval $[0, t]$ *and* in the time interval $[t, t + \Delta t]$. Independence of the last two events imply that

$$P(t + \Delta t) = P(t)(1 - \text{Prob}[\text{ Something happens in the interval } [t, t + \Delta t]]) \tag{3.37}$$

Probability of something happening in the small interval $[t, t + \Delta t]$ is approximately $R\Delta t$. Expanding to first order in Δt and noting that [c1] $P(0) = 1$, yields an expression for the cumulative distribution function (cdf) of a Poisson process,

$$P(t) = e^{-Rt}. \tag{3.38}$$

[c2]A fundamental quantity in the theory of Poisson processes is the waiting time distribution. This is also sometimes referred to as the time lag distribution. The waiting time distribution, $p(t)$, is the probability density of the lengths of inter-arrival times in a Poisson process. In particular, the probability that the time between synthesis events falls in the the interval $[t, t + dt]$ is given by $p(t)dt$. Since $p(t)$ is just the probability of having no events in the interval $[0, t]$ followed by an event in the interval $[t, t + dt]$, one has

$$p(t) = Re^{-Rt}. \tag{3.39}$$

Equivalently, we can define the waiting time distribution as the negative rate of change of $P(t)$, $p(t) = -dP(t)/dt$.

Exercise: Show that for a Poisson process the mean waiting time is $\langle t \rangle = R^{-1}$ and that the standard deviation is $\sqrt{\text{Var}(t)} = \sqrt{\langle t^2 \rangle - \langle t \rangle^2} = R^{-1}$.

[c1] *Pankaj*: *Text added.*

[c2] *Pankaj*: ~~If the synthesis apparatus gets ready to produce the molecules at the same rate very quickly (compared to typical lag, once more) then~~

[c3] *Pankaj*: *Text added.*

[c1] *Pankaj*: ~~This fact along with the observation~~

[c2] *Pankaj*: *Text added.*

[c3] How is the number of events in an interval distributed? We can borrow tricks from the random walk calculation (remember "Be wise, discretize!"?). Let us divide up the interval $[0, t]$ into N subintervals, each of length Δt (so that $t = N\delta t$). Choose N large enough so that each subinterval is much smaller than R^{-1}. In this case, it is unlikely that there will be two or more events happening in the same subinterval[c3]. Now we can ask what is the probability of n events happening? We need to choose n or the N subintervals, providing a factor ${}^{N}C_{n}$, and multiply by the probability of the n intervals being occupied and the remaining $N - n$ intervals being unoccupied, $(R\Delta t)^n * (1 - R\Delta t)^{N-n}$. This yields

$$\text{Prob}[n \text{ events}] =^{N} C_n (R\Delta t)^n (1 - R\Delta t)^{N-n}. \tag{3.40}$$

Hence, in the limit $N \to \infty$ with n fixed,

$$\begin{aligned} \text{Prob}[n \text{ events}] &= \frac{N(N-1)\cdots(N-n+1)}{n!} N^{-n} (Rt)^n (1 - \frac{Rt}{N})^N (1 - \frac{Rt}{N})^{-n} \\ &\approx \frac{(Rt)^n e^{-Rt}}{n!}. \end{aligned} \tag{3.41}$$

This is just the Poisson distribution with the mean $\nu = Rt$.

Excercise: Show that $\langle n \rangle = \nu$ and the $\text{Var}(n) = \nu$. Notice that the standard deviation of the number is $\sqrt{\nu}$ and the relative fluctuation, the ratio of standard deviation to mean, scale as $\nu^{-1/2}$, which becomes small as ν becomes large.

[c1] This is a good place to discuss the connections between the deterministic description of the previous chapter and the stochastic description in this one. As time grows, the number of molecules made becomes large. To give an example not too far from reality, if the time period is such that, on the average, a thousand molecules are made , we expect the fluctuation around the mean to be about thirty(which is about 3% of the mean). In this case, the deterministic approximation is a pretty good one. If, on the other hand, if we are dealing with systems with an average of ten molecules, the relative fluctuation are about 30% and the stochastic effects can play an important role. To account for these fluctuations, one possible approach is to use a continuum Langevin description as an approximation to the Poisson process

$$\frac{dn(t)}{dt} = R + \eta(t) \tag{3.42}$$

[c3] *Pankaj*: Tweaked presentation slightly

[c3] The probability of that happening goes as $(R\Delta t)^2$ and therefore the number of subintervals with double events is expected to be $N(R\Delta t)^2 = (Rt)^2/N$, something that tends to zero as N goes to infinity.

[c1] *Pankaj*: Added exercise and tweaked discussion slightly

with $\langle \eta(t) \rangle = 0$ and $\langle \eta(t)\eta(t') \rangle = R\delta(t-t') \rangle$.

Exercise: Show that the Langevin approximation reproduces the right mean and variance for the Poisson process.

When time t and $\nu = Rt$ become large, the poission distribution is well approximated by a gaussian with mean ν and variance ν. In this limit, the Langevin description becomes a good description of the Poisson process.

Excercise: In this problem, we consider a protein production from a single mRNA molecule. Assume that proteins are produced by a Poisson process with rate α_p and that mRNA degradation is also a Poisson process with rate τ_m^1. Show that the probability of producing b proteins from an mRNA molecule is given by the Geometric distribution with mean $\bar{b} = \alpha_p \tau_m$,

$$G(b) = \frac{1}{1+\bar{b}} \left(\frac{\bar{b}}{1+\bar{b}} \right)^b . \tag{3.43}$$

3.5 Birth-death processes and the Gillespie Algorithm

So far we have only described a process in which a molecule is only made not destroyed. In biological systems many molecules have dedicated enzymes for destroying them. RNA and proteins are degraded by RNases and proteases, respectively, and both play important roles in gene expression regulation. Proteins like phosphodiesterase convert cyclic nucleotide monophosphate to nucleotide monophosphate, and affect signaling. For any posttranslational modification of proteins, like phosporylation etc., there are enzymes like posphatases, that undo the change.

In general, for each molecule the birth rate and the death rate can depend on the number of other molecules present. We specify the state of the cell by a vector of numbers for the different species of molecules, $\vec{n} = (n_1, n_2, ..., n_k)$, the rate of synthesis of species i, $(n_1, n_2, ..., n_i, ..., n_k) \to (n_1, n_2, ..., n_i + 1, ..., n_k)$, by $f_i(\vec{n})$, and the rate of degradation of species i,$(n_1, n_2, ..., n_i, ...n_k) \to (n_1, n_2, ..., n_i - 1, ..., n_k)$, by $g_i(\vec{n})$. For such a system, the deterministic equation describing the way the average changes when the numbers are large is

$$\frac{d\vec{n}}{dt} = \vec{f}(\vec{n}) - \vec{g}(\vec{n}). \tag{3.44}$$

How does one deal with the stochastic aspects of this model? Is there a limit in which one can apply the Langevin approximation to the system?
[c1]

Before proceeding to analytic approximations, it is useful to discuss how to numerically simulate the chemical reactions like birth-death processes. One

[c1] *Pankaj*: Added Gillespie into text. This is standard and must be here

approach one can imagine to simulating these reactions is to choose a small time step $\Delta t \ll 1$, draw a uniform random number for each reaction, check if a synthesis or degradation event occurs during the time step by determining if the corresponding random number is smaller than $f_i(\vec{n})\Delta t \ll 1$ or $g_i(\vec{n})\Delta t \ll 1$, updating the state of the cell, and then repeating the process. The problem with such a naive approach is that since the probability of an event occurring in any time step is extremely small. In fact, during most time steps nothing will happen. Consequently, such simulations are extremely inefficient and slow. One can imagine speeding up the simulation by increasing Δt. However, for larger Δt one quickly runs into the problem that there is a non-zero probability of having multiple events during each time step.

An alternative approach, often termed the "Gillespie Algorithm", circumnavigates the problems discussed above and has quickly become the standard technique for simulating stochastic chemical reactions in systems biology. We now discuss how to use the Gillespie algorithm to simulate an arbitrary set of chemical reactions. As before, denote the number of molecules present of all species by $\vec{n}$. Furthermore, index the possible reactions by P, with the rate of reaction P, $\vec{n} \to \vec{n}+\vec{e}_P$, given by $r_P(\vec{n})$. For example for the the birth-death processes discussed above, we can consider the reaction for the creation of a molecule of species i. For this case, $r_P = f_i(\vec{n})$ we have that $e_p = (0, 0, \ldots, 1, \ldots, 0)$, the vector with 1 at the i-th position and zero everywhere else. The key observation behind the Gillespie algorithm is that each reaction is an independent Poisson process so we can explicitly calculate the waiting time distribution between events. In particular, the probability that *any* event occurs is a Poisson process with rate $R = \sum_P r_P$. This suggests the following algorithm for simulating chemical reactions:

- Initialize the simulation at some $\vec{n} = \vec{n}_0$ and time $t = 0$.
- Draw a random number x_1 uniformly distributed between 0 and 1.
- Explicitly calculate the waiting time τ between events using Equation 3.39: $\tau = -\log(x_1)/r$. This step uses the usual Monte-Carlo sampling procedure based on the cumulative distribution function (cdf) of a Poisson process, $P(t)$, derived in Eq. 3.38.
- Draw a second random number x_2 uniformly distributed between 0 and 1 to choose which of the reactions occurs. Reaction P occurs if

$$\frac{\sum_{j=1}^{p-1} r_j}{R} \leq x_2 < \frac{\sum_{j=1}^{p} r_j}{R} \tag{3.45}$$

  .
- Update the time, $t \to t + \tau$, and $\vec{n}$ using appropriate reaction.

We end this section by emphasizing that the Gillespie algorithm outlined above is exact. No approximations of any kind were utilized. For this reason, the Gillespie algorithm has become one of the workhorses of simulating biochemical

reactions. We encourage the reader to simulate the various processes that occur through out the remainder of the book.

Excercise: Use the Gillespie algorithm to simulate a simple birth-death process for a single species where molecules are synthesized at a rate $f(n) = R$ and are degraded at a rate $g(n) = \tau^{-1}n$. Compare your results with those in Figure 3.6.

3.6 Noise in chemical reactions

[c1]

[c2] We start our discussion with a detailed analysis in the context of a simple birth-death process for a single species where molecules are synthesized at a rate $f(n) = R$ and are degraded at a rate $g(n) = \tau^{-1}n$. We will start with a deterministic description and then consider the effects of stochastic fluctuations due small molecule numbers. Comparison of the deterministic and stochastic modeling of this particular process will further our understanding of when and how noise affects biological systems as well as help explain thee analogy between biased random walks and noise in chemical kinetics.

The deterministic description of the birth-death process can be written in terms of a simple Ordinary Differential Equation of the form

$$\frac{dn(t)}{dt} = R - \frac{n(t)}{\tau}. \tag{3.46}$$

[c3]The first term just says that the change in n during a time dt is the difference between the production rate, R, and the degradation rate $\frac{n(t)}{\tau}$. It is easy to obtain a closed-form, analytic solution to this equation. However, for pedagogic reasons, it is useful to solve this equation in a different way. In particular, we will solve the problem using methods that can also be used to solve more complicated mon-linear system that frequently occur when modeling biological systems. [c4]

[c5] We start by asking about the fixed points of this dynamical system. Setting the left hand side of equation 3.47 to zero, we see that there is a single fixed point at $\bar{n} = R\tau$. Let us call this quantity ν. Next we ask whether this fixed point is stable. To do so, we analyze small perturbations around $\bar{n} = \nu$ and see if the they die out in time. Departures from ν, $\delta n = n - \nu$, satisfy the linearized equation,

$$\frac{d}{dt}\delta n(t) = -\frac{\delta n(t)}{\tau} \tag{3.47}$$

In this case, the linearized equation is exact since the original problem is also linear. However, for a general a non linear system, this equation is approximate

[c1] *Pankaj*: STILL HAVE TO FIX SECTION
[c2] *Pankaj*: Moved this from last section
[c3] *Pankaj*: *Text added.*
[c4] *Pankaj*: modified a little
[c5] *Pankaj*: Minor tweaks of paragraph below

since we have ignored all terms order δn^2 and above. In any case, the last equation implies that the perturbations die out exponentially with a time constant set by τ and the fixed point is stable.

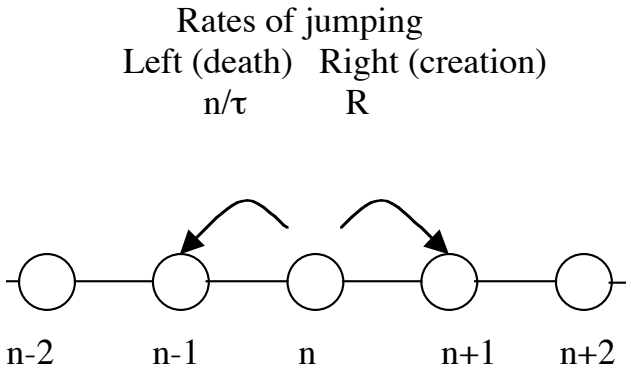

FIG. 3.5. Random walk on the space of numbers.

What about the effects of stochasticity? The state space of our system is the number of molecules present which takes on values in the set of nonnegative integers, . The dynamics of the system can be thought of as a random walk on the space of nonnegative numbers with the bias of hoping left and right related to the birth and death rates. The probability of hopping up in a small time interval $[t, t + \Delta t]$ (i.e. from n to $n + 1$) is $R\Delta t$. The probability of hopping down during the same interval (from n to $n - 1$) is $n\Delta t/\tau$. Notice this later rate depends on the present state the system. If $n < \nu$ the random walk the is biased towards going up, and for $n < \nu$, it is biased downward. In FIG. 3.6, we compare the deterministic dynamics with a simulation of the random dynamics for a system starting with no particles climbing up to roughly ν in number. After a time order τ the system settles down to its stationary state.

This analogy can be made more concrete by considering the Master equation for this process. Let $P(n, t)$ be the probability distribution of n at time t. This is related to the probability distribution at a time $t + \Delta t$, $P(n, t + \Delta t)$, through the equation

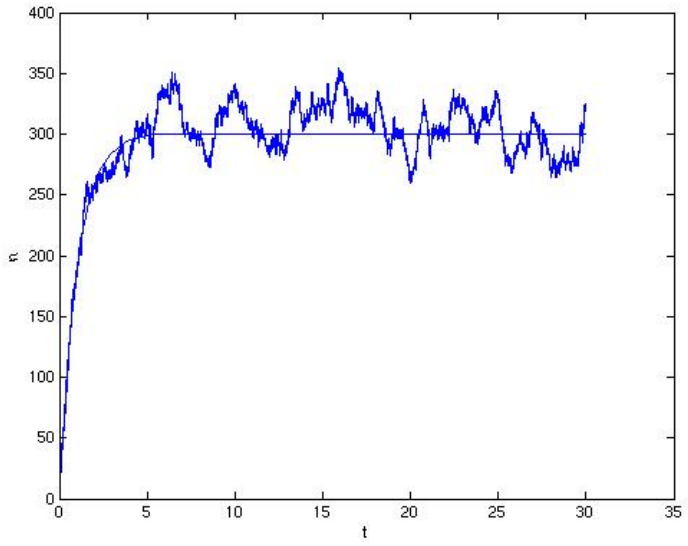

FIG. 3.6. Comparing deterministic and stochastic descriptions of reaction kinetics for a simple birth-death process.

$$\begin{aligned} P(n,t+\Delta t) - P(n,t) &= \text{gain from jumps from } n-1 \\ &\quad + \text{gain from jumps from } n+1 \\ &\quad - \text{losses from jumps away from n} \\ &= R\Delta t P(n-1,t) + \frac{n+1}{\tau}\Delta t P(n+1,t) - (R\Delta t + \frac{n}{\tau}\Delta t)P(n,t) \\ &= \Delta t[RP(n-1,t) - (R+\frac{n}{\tau})P(n,t) + \frac{n+1}{\tau}P(n+1,t)]. \end{aligned} \tag{3.48}$$

Taking the limit $\Delta t \to 0$ and using a Taylor expansion yields the Master equation

$$\frac{d}{dt}P(n,t) = RP(n-1,t) - (R+\frac{n}{\tau})P(n,t) + \frac{n+1}{\tau}P(n+1,t) = J(n,t) - J(n+1,t) \tag{3.49}$$

where

$$J(n,t) = RP(n-1,t) - \frac{n}{\tau}P(n,t) \tag{3.50}$$

is the overall probability current flowing between n and $n-1$. When we come to $n=0$, we have

$$\frac{d}{dt}P(0,t) = -RP(0,t) + \frac{1}{\tau}P(1,t) = J(0,t). \tag{3.51}$$

Just as we asked about the fixed points of the deterministic dynamics, we can ask about stationary distributions that describe the long time dynamics. To do

so, we have to set dP/dt to zero in equation 3.49 and solve for time independent $P(n,t) = P_0(n)$.

[c1] One way to do this is to realize that stationarity implies that the current $J(n,t)$ is constant throughout the system and independent of n and t (i.e. $J(n,t) = J(n+1,t) = J_c$ with J_c a constant). Furthermore, $dP(0,t)/dt = 0$ implies $J_c = 0$. Taken together, this is simply the statement that a net current to the left or right is inconsistent with a stationary distribution where there is no flux at $n = 0$.

Excercise: Show that $J_c = 0$ implies that the stationary distribution is a Poisson distribution of the form

$$P_0(n) = e^{-\nu}\frac{\nu^n}{n!}. \tag{3.52}$$

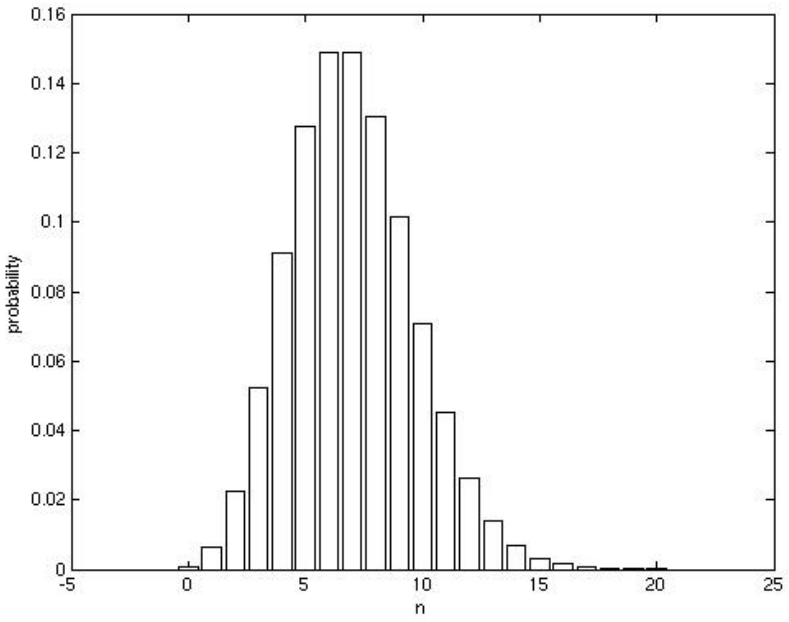

FIG. 3.7. Poisson distribution for $\nu = 7$.

[c2] For $\nu >> 1$, the stationary Poisson distribution is strongly peaked around ν and one can gainfully apply the Langevin description to understand the stochas-

[c1] *Pankaj*: Reworked next paragraph

[c2] *Pankaj*: I have removed a lot of the technical detail below. It is confusing even to me, let alone beginners

tic dynamics. The ra deterministic equation (3.47) is supplemented by a noise term $\eta(t)$ with $n(t)$-dependent noise strength,

$$\frac{dn(t)}{dt} = R - \frac{n(t)}{\tau} + \eta(t) \tag{3.53}$$

with $\langle \eta(t) \rangle = 0$ and $\langle \eta(t)\eta(t') \rangle = (R + \frac{n(\frac{t+t'}{2})}{\tau})\delta(t-t')$. To see when the Langevin description applies, note that in this system typical time scale for change of $n(t)$ is set by τ. Let us choose a time interval $[t, t']$ so that $(t' - t) << \tau$ but $R(t' - t) >> 1$ and $\int_t^{t'} ds n(s)/\tau \approx n(\bar{t})/\tau >> 1$, with $\bar{t} = (t + t')/2$. This is possible when $\nu, n(t) >> 1$. Since the time interval $[t, t']$ is much shorter than τ, we can assume $n(t)$ is nearly constant over the interval. Within this approximation, the birth and death are independent poisson processes with rate R and rate $n(\bar{t})/\tau$, respectively. [c1] Since the variance of the sum of two independent Poisson processes is the sum of the variances of each process, one has $\langle \eta(t)\eta(t') \rangle = (R + n(\bar{t})/\tau)\delta(t - t')$. The choice of the midpoint $\bar{t}$ as an argument of n has to do with technicalities of interest to experts (see comments in the Appendix). For most calculations that we will do, the choice of the midpoint is not crucial. A special case of interest are stochastic fluctuations around the fixed point $\bar{n} = \nu$. In this case, the noise takes on the simpler form

$$\langle \eta(t)\eta(t') \rangle = (R + \nu/\tau)\delta(t - t') = 2R\delta(t - t') = 2\nu/\tau\delta(t - t') \tag{3.54}$$

The discussion above can also be generalized to more complicated systems. The one subtlety is that one needs to be careful about which processes are independent. Label the independent processes by an index p. Let process p happen at rate $f_p(\vec{n})$ with $\vec{n}$ to $\vec{n} + \vec{e}_p$. Then, the corresponding Langevin equation takes the form

$$\frac{d\vec{n}(t)}{dt} = \sum_p f_p(\vec{n}(t))\vec{e}_p + \vec{\eta}(t) \tag{3.55}$$

with $\langle \vec{\eta}(t) \rangle = 0$ and $\langle \eta_i(t)\eta_j(t') \rangle = \sum_p e_{pi}e_{pj}f_p(\vec{n}(\frac{t+t'}{2}))\delta(t-t')\delta_{ij}$ We will apply this method in the next section to analyze fluctuations in gene expression.

3.7 Application to intrinsic fluctuation of gene expression

It would be good at this point to comeback to biology and jump in the middle of current discussion on noise in gene expression. Gene expression consists of two steps: transcription and translation. The first step make mRNA and the second step makes a protein using the information from mRNA. A biologically important quantity to understand are fluctuations in protein numbers to due to

[c1] *Pankaj*: ~~The fluctuation of the gain during the interval is the difference between the fluctuation in birth and the fluctuation in death, and the last two are independent variables. Therefore the variance of the fluctuation in net gain is some of the the variances from birth and that from death.~~

stochastic effects. Our discussion will closely follow the treatment by Thattai and Oudenaarden.

Our system is specified at any time t by the total number of mRNA molecules r and protein molecules p present. We [c2]restrict ourselves to the case where mRNA molecules are synthesized constitutively off the template DNA strand and are translated at some constant rate (FIG. 1). We treat transcription and translation [c3]as independent processes that occur instantaneously. [c4]The kinetics of gene expression can be represented as follows:

$$\begin{array}{ccccc} & \overset{k_R}{} & & \overset{k_P}{} & \\ \mathrm{DNA} & \rightarrow & \mathrm{mRNA} & \rightarrow & \mathrm{protein}, \\ & & \downarrow 1/\tau_R & & \downarrow 1/\tau_P \\ & & \emptyset & & \emptyset \end{array} \tag{3.56}$$

with k_R the transcription rate, k_P the translation rate per mRNA molecule, τ_R the mRNA lifetime, and τ_P the protein lifetime.

The probability that the system is in a given state (r, p) is specified by the time dependent joint probability distribution $P(r, p; t)$. [c1] Instead of using Master Equations to analyze noise, we will start directly with the Langevin approximation and analyze fluctuation around the stationary state.

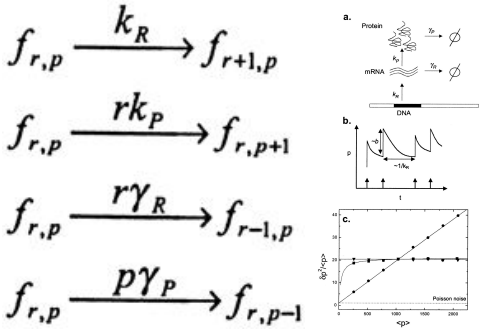

FIG. 3.8. Stochastic descriptions of gene expression kinetics. **FIX CONSTANTS**

[c2] *Pankaj*: ~~assume that for the gene involved,~~

[c3] *Pankaj*: ~~events as point processes and as independent of each other~~

[c4] *Pankaj*: *Text added.*

[c1] *Pankaj*: ~~We could have started by writing down the Master equation. Instead,~~

$$\frac{dr}{dt} = k_R - \frac{r}{\tau_R} + \eta_R$$
$$\frac{dp}{dt} = k_P r - \frac{p}{\tau_P} + \eta_P \tag{3.57}$$

The stationary solution [c1] for the mean mRNA number, $\bar{r}$, and the mean protein number, $\bar{p}$, can be found from the equations above by setting the time derivatives to zero and ignoring the noise terms. This yields $\bar{r} = k_R\tau_P$ and $\bar{p} = k_P\bar{r}\tau_P$. Around this fixed point the noise spectrum is given by $\langle\eta_{R,P}\rangle = 0$ and

$$\langle\eta_R(t)\eta_R(t')\rangle = (k_R + \frac{\bar{r}}{\tau_R})\delta(t-t') = 2k_R\delta(t-t') \tag{3.58}$$
$$\langle\eta_P(t)\eta_P(t')\rangle = (k_P\bar{r} + \frac{\bar{p}}{\tau_P})\delta(t-t') = 2k_Rk_P\tau_R\delta(t-t') \tag{3.59}$$
$$\langle\eta_R(t)\eta_P(t')\rangle = 0. \tag{3.60}$$

[c2] In writing these expressions, we have used the fact that the birth and death of mRNAs and proteins are independent Poisson processes with variance equal to the their mean.

The departures for the stationary values, $\delta r(t) = r(t) - \bar{r}$ and $\delta p(t) = p(t) - \bar{p}$, satisfy the matrix equation

$$\begin{pmatrix} \frac{d}{dt} + \frac{1}{\tau_R} & 0 \\ -k_P & \frac{d}{dt} + \frac{1}{\tau_P} \end{pmatrix} \begin{pmatrix} \delta r \\ \delta p \end{pmatrix} = \begin{pmatrix} \eta_R \\ \eta_P \end{pmatrix}. \tag{3.61}$$

[c3] These equations can be derived by linearizing Eq. 3.57 [c4] around the fixed point. Taking Fourier transformation

$$\begin{pmatrix} -i\omega + \frac{1}{\tau_R} & 0 \\ -k_P & -i\omega + \frac{1}{\tau_P} \end{pmatrix} \begin{pmatrix} \delta\hat{r}(\omega) \\ \delta\hat{p}(\omega) \end{pmatrix} = \begin{pmatrix} \hat{\eta}_R(\omega) \\ \hat{\eta}_P(\omega) \end{pmatrix} \tag{3.62}$$

and inverting the matrix, we have

$$\hat{p}(\omega) = \frac{k_P\hat{\eta}_R(\omega)}{(-i\omega + \frac{1}{\tau_R})(-i\omega + \frac{1}{\tau_P})} + \frac{\hat{\eta}_P(\omega)}{-i\omega + \frac{1}{\tau_P}}. \tag{3.63}$$

The two terms in the right hand side of the equation 3.63 can be interpreted as follows: the first term is the effect of fluctuations in transcription and the second is the effect of fluctuation in translation.

[c1] *Pankaj: Text added.*
[c2] *Pankaj: Text added.*
[c3] *Pankaj: Text added.*
[c4] *Pankaj: Text added.*

The Fourier space version of the noise correlations,

$$\langle \hat{\eta}_R(\omega)\hat{\eta}_R(\omega')\rangle = 4\pi k_R \delta(\omega-\omega') \tag{3.64}$$
$$\langle \hat{\eta}_P(\omega)\hat{\eta}_P(\omega')\rangle = 4\pi k_R k_P \tau_R \delta(\omega-\omega') \tag{3.65}$$
$$\langle \hat{\eta}_R(\omega)\hat{\eta}_P(\omega')\rangle = 0 \tag{3.66}$$

imply

$$\langle \hat{p}(\omega)\hat{p}(\omega')^*\rangle = \left[\frac{2k_P k_R}{(\omega^2+\frac{1}{\tau_R^2})(\omega^2+\frac{1}{\tau_P^2})} + \frac{2k_R k_P \tau_P}{\omega^2+\frac{1}{\tau_P^2}}\right]\delta(\omega-\omega'). \tag{3.67}$$

Instantaneous fluctuation in $\delta p(t)$, [c1]

$$\langle(\delta p(t))^2\rangle = \langle \int \frac{d\omega}{2\pi} e^{-i\omega t}\hat{p}(\omega) \int \frac{d\omega'}{2\pi} e^{-i\omega t'}\hat{p}(\omega')\rangle \tag{3.68}$$
$$= \int \frac{d\omega}{2\pi}\int \frac{d\omega'}{2\pi}\langle \hat{p}(\omega)\hat{p}(\omega')^*\rangle e^{-i(\omega-\omega')t}$$
$$= \int \frac{d\omega}{2\pi}\left[\frac{2k_P^2 k_R}{(\omega^2+\frac{1}{\tau_R^2})(\omega^2+\frac{1}{\tau_P^2})} + \frac{2k_R k_P\tau_P}{\omega^2+\frac{1}{\tau_P^2}}\right]$$
$$= \frac{k_R k_P^2 \tau_R^2 \tau_P^2}{\tau_P^2-\tau_R^2}\int \frac{d\omega}{\pi}\left\{\frac{\tau_P^2}{1+\omega^2\tau_P^2} - \frac{\tau_R^2}{1+\omega^2\tau_R^2}\right\} + k_R k_P \tau_R \int \frac{d\omega}{\pi}\frac{\tau_P^2}{1+\omega^2\tau_P^2}$$
$$= \frac{k_R k_P^2 \tau_R^2 \tau_P^2(\tau_P-\tau_R)}{\tau_P^2-\tau_R^2} + k_R k_P \tau_P \tau_R$$
$$= \bar{p}\left(1+\frac{k_P\tau_R\tau_P}{\tau_R+\tau_P}\right). \tag{3.69}$$

[c2] It is interesting to consider the results in the limit where protein lifetimes are much longer than mRNA lifetimes, $\tau_P >> \tau_R$. In bacteria, τ_P is often hours whereas τ_R is minutes. In this limit

$$\langle \delta p^2\rangle \approx \bar{p}(1+k_P\tau_R) \equiv \bar{p}(1+\bar{b}) \tag{3.70}$$

The quantity $\bar{b} = k_P\tau_R$ is the average number of proteins per mRNA, is called the "burst size". If $\bar{b}$ is large (often tens of protein molecules are made from a single mRNA molecule) the transcriptional noise dominates. [c3] A commonly used measure of the noise is the Fano factor, $\nu =< \delta p^2 > /\bar{p}$. For Poisson distributions, the Fano factor is one. The fact that in our process the Fano factor is $1+\bar{b}$ indicates the non Poissonian nature of the protein distribution.

[c1] *Pankaj*: added first equation, fixed pi's
[c2] *Pankaj*: Changed this
[c3] *Pankaj*: ~~One could calculate~~

What is the experimental signature of [c4]bursty protein synthesis? Since the Fano factor ν depends upon the burst size, it should be unaffected by transcription rate, but should be affected by translation rate. This was tested in experiments done in *Bacillus subtillis* by Ozbudak et al. A reporter expressing a Green Fluorescent Protein provided a way to measure relative levels of gene expression. This provides protein numbers up to a scale factor. Ozbudak et al played around with the strength of the sigma factor binding site which affects transcription rate as well as with the Shine Dalgarno sequence, i. e. the ribosome binding site, which affects the translation rate. Mutations of the latter kind affected the Fano factor ν strongly, whereas the mutations of the first kind did not affect ν seriously. [c5]This is consistent with the idea that the dominant source of noise is the amplification of mRNA fluctuations through bursty protein synthesis.

[c1] There has been some discussion on which is the better measure of noisiness: the Fano factor $\sqrt{\langle \delta p^2 \rangle}/\bar{p}$ or $\sqrt{\langle \delta p^2 \rangle}/\bar{p}$? The first measure explicitly scales out overall number/size of the system, the second has the advantage that it can be estimated even if one knows protein abundances only up to a scale. However, this measure explicitly depend upon the overall number. Defining a measure of noisiness without any regard to what the purpose of having low noise is pointless. For example, in particular contexts, like signaling, we know that the latter quantity is related to the signal-to-noise ration and hence more meaningful whereas the Fano factor may be more useful for interpreting certain experiments.

A more interesting question is whether gene expression noise is dominated by intrinsic noise inherent to protein production or noise extrinsic to the process such as fluctuations in cell size, ribosome number, RNA polymerase number, etc. One might surmise that intrinsic noise is uncorrelated between two genes but the extrinsic noise is correlated for two genes within the same cell. Using this tactics, Elowitz *et al.* analyzed gene expression of two reporter Fluorescent proteins in E. coli and showed that extrinsic noise is as large or larger than the intrinsic noise.

[c2] **Exercise**: Use the Gillespie Algorithm to simulate the kinetics of gene expression depicted in Eq. 3.56. Choose reasonable parameters for the four rates. Compare your results with results from the Langevin approximation in Eq. 3.69.

3.8 Fluctuation of gene expression in eukaryotes

Eukaryotic DNA is neatly spooled around proteins called histones. The combination of DNA and DNA bound histones is referred to as chromatin. Some times positioning of the nucleosomes and/or the covalent modification of components of it are such that initiation of transcription becomes unlikely. In those cases,

[c4] *Pankaj*: ~~this phenomenon~~
[c5] *Pankaj*: *Text added.*
[c1] *Pankaj*: made some minor changes
[c2] *Pankaj*: added excercise

the control of transcription in often have an additional layer, namely chromatin remodeling.

$$\begin{array}{ccccccc} & & k_a & & k_R & & k_P & \\ \text{inactive DNA} & \rightleftharpoons & \text{active DNA} & \rightarrow & \text{mRNA} & \rightarrow & \text{protein} \\ & & 1/\tau_a & & \downarrow 1/\tau_R & & \downarrow 1/\tau_P & \\ & & & & \emptyset & & \emptyset & \end{array} \tag{3.71}$$

[c1] If the fluctuation between active and inactive DNA is much faster than the rest of the gene expression process, the net effect is to lower the 'effective' k_P by a factor proportional to the probability of the DNA being in the open state. The rest of the analysis is very similar to what we did in the previous section. However, if either the opening up inactive DNA or shutting of active DNA is a slow process, qualitative differences arise from the simple scenario considered in the last section. When both processes are slow, the cells divide into two sub populations, a productive population with active DNA, and unproductive population with inactive DNA. In this limit, the distribution of gene expression can become bimodal, with a peak coming from each of the two sub-populations. Such distributions show a large variance of protein from cell to cell. If by some means, say by over-expressing and activator, we could tilt the balance in favor of active DNA, then the average protein number would go up while the variance would decrease since the distribution would no longer be bimodal. Note that this is qualitatively different from the simpler model in the last section where increasing the mean protein number always increased the variance. We will return to bimodal gene expression profiles in the context of genetic switches in later chapters.

Understanding effects of noise on the function of biochemical networks as well as the various sources of noise has drawn much attention, currently. We will discuss some these topics in the chapters to come. Ability to record fluctuating single cell measurements in real time is making the field productive and exciting. There are many reviews on the subject and the reader is urged to consult them for a deeper understanding of stochasticity in biological phenomena.

3.9 Post-transcriptional regulation by small RNAs

[c2] **Excercise:** Small, non-coding RNAs (sRNAs) play important roles as genetic regulators in prokaryotes. sRNAs act post-transcriptionally via complementary pairing with target mRNAs to regulate protein expression. One major class of prokaryotic sRNAs (antisense sRNAs) negatively regulate proteins by destabilizing the target protein's mRNA (Fig. 3.9). These $\sim$ 100 bp antisense sRNAs

[c1] *Pankaj*: Minor editorial changes
[c2] *Pankaj*: added excercise/section

prevent translation by binding to the target mRNAs in a process mediated by the RNA chaperone Hfq. Upon binding, both the mRNAs and sRNAs are degraded.

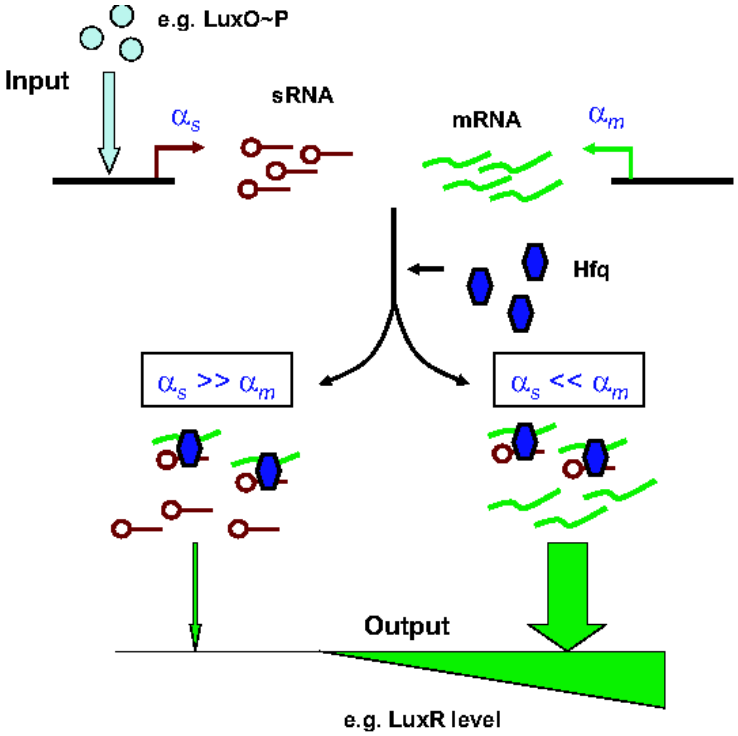

FIG. 3.9. Genetic Regulation via sRNAs. Left: Small non-coding RNAs (sRNAs) regulate protein expression as part of a larger genetic network with a specific biological task (e.g. quorum sensing in *Vibrio* bacteria). The sRNAs (stem loops) regulate target proteins by destabilizing target-protein mRNAs (wavy lines), a stoichiometric process mediated by the RNA chaperone Hfq (hexagons). When the rate of sRNA transcription α_s greatly exceeds the rate of mRNA transcription α_m, i.e. when $\alpha_s \gg \alpha_m$, nearly all the mRNAs are bound by sRNAs and cannot be translated. By contrast, when $\alpha_m \gg \alpha_s$, there are many more mRNAs than sRNAs, and protein is highly produced.

In this problem, post-transcriptional regulation via sRNAs is modeled using mass-action equations with three molecular species: the number of sRNA molecules s, the number of target mRNA molecules m, and the number of regulated protein molecules p. The effect of intrinsic noise is modeled by Langevin terms, $\hat{\eta}_j$, that describe the statistical fluctuations in the underlying biochemical reactions. The kinetics of the various species are described by the differential

equations

$$\begin{aligned}\frac{ds}{dt} &= \alpha_s - \tau_s^{-1} s - \mu m s + \hat{\eta}_s + \hat{\eta}_\mu \\ \frac{dm}{dt} &= \alpha_m - \tau_m^{-1} m - \mu m s + \hat{\eta}_m + \hat{\eta}_\mu \\ \frac{dp}{dt} &= \alpha_p m - \tau_p^{-1} p + \hat{\eta}_p \,. \end{aligned} \tag{3.72}$$

The terms can be interpreted as follows. sRNAs (mRNAs) are transcribed at a rate α_s (α_m), and are degraded at a rate τ_s^{-1} (τ_m^{-1}). Additionally, sRNAs and mRNAs are *both* stoichiometrically degraded by pairing via Hfq at a rate that depends on the sRNA-mRNA interaction strength μ . Proteins are translated from mRNAs at a rate α_p and are degraded at a rate τ_p^{-1}. $\hat{\eta}_s$, $\hat{\eta}_m$, and $\hat{\eta}_p$ model the noise in the creation and degradation of individual sRNAs, mRNAs, and the regulated protein, respectively. $\hat{\eta}_\mu$ models sRNA-mRNA mutual-degradation noise.

a) Recall, that the Langevin terms are characterized within the linear-noise approximation by two-point time-correlation functions ($j = s, m, p, \mu$) , which for steady states take the form

$$\langle \hat{\eta}_j(t) \hat{\eta}_j(t') \rangle = \sigma_j^2 \delta(t - t') \qquad . \tag{3.73}$$

What are the σ_j for the three processes?
b) Calculate the steady-state protein number as a function of the parameters? Plot the steady-state as function of α_m ? Show that in the limit $\mu \to \infty$ limit it has a "threshold-linear" form as a function of α_m, with no expression when $\alpha_m < \alpha_s$, and linear dependence on α_m for $\alpha_m > \alpha_s$.
c) Use the Gillespie algorithm to simulate this process for various values of α_m? Plot the noise $\sigma_p^2/\bar{p}^2$ as function of α_m. What happens around the threshold $\alpha_m = \alpha_s$? Use reasonable parameters for $\alpha_s, \tau_m, \tau_p, \tau_s, \alpha_p$ and $\mu = 2$.
d) Solve for the noise using the Langevin approximation and compare with your simulation results.