Stochastic Gene Expression Project: Homework 4
Due February 8th at 2:15 PM

Problem 1: 'Deriving' kernel density estimation.

Given N independent samples $x_1, ..., x_N$ from a probability distribution with density $P(x)$, we can estimate $P(x)$ by

$$P(x) \approx \frac{1}{Nh} \sum_{k=1}^{N} K\left(\frac{x - x_k}{h}\right) \tag{1}$$

where K is the kernel function and h is the *bandwidth*. Increasing h is like using a histogram with larger bins, while decreasing h is like using a histogram with smaller bins.

(a) Consider the *characteristic function* $\phi(t)$ defined as

$$\phi(t) := \mathbb{E}(e^{itx}) = \int_{-\infty}^{\infty} e^{itx} P(x) dx \ , \tag{2}$$

where $t \in \mathbb{R}$. It turns out that it is often mathematically convenient to study $P(x)$ through its characteristic function $\phi(t)$; one reason is because knowing $\phi(t)$ is equivalent to knowing $P(x)$. Show that

$$P(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \phi(t) dt \tag{3}$$

using the identities

$$\delta(x - y) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{\pm it(x-y)} dt \qquad \text{and} \qquad f(y) = \int \delta(x - y) f(x) dx \ ,$$

where δ is the Dirac delta function and f is an arbitrary function.

(b) Given N independent samples $x_1, ..., x_N$, it makes sense to estimate $\phi(t)$ as

$$\phi(t) \approx \frac{1}{N} \sum_{k=1}^{N} e^{itx_k}$$

since this is how we normally take averages given some data. But if we put this approximation into Eq. 3, we don't get a great approximation for P. What is the problem?

(c) Consider *regulating* (i.e. making more mathematically tame) our estimate by writing

$$\phi(t) \approx \frac{1}{N}\sum_{k=1}^{N} e^{itx_k}\psi(ht)$$

for some function ψ which satisfies $\psi(0) = 1$ and $\psi(\pm\infty) = 0$. For example, we could choose ψ to be a Gaussian, and changing h would correspond to changing its standard deviation. Use Eq. 3 to show that

$$P(x) \approx \frac{1}{Nh}\sum_{k=1}^{N}\left[\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-iu\left(\frac{x-x_k}{h}\right)}\psi(u)du\right] \ .$$

If we define

$$K(y) := \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-iuy}\psi(u)du \ ,$$

then we recover Eq. 1. Hence, we have 'proven' that kernel density estimation yields a good approximation for $P(x)$.

(d) Show that if $\psi(t) = \exp(-t^2/2)$, then

$$K(y) = \frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}$$

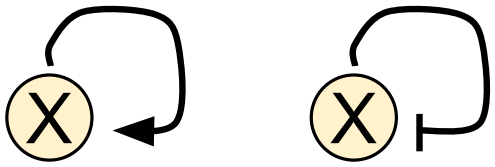
i.e. that using a Gaussian regulator corresponds to using a Gaussian kernel function (the kind we are using). You may want to use the identity

$$\int_{\infty}^{\infty} e^{-ax^2+bx} = \sqrt{\frac{\pi}{a}}\exp\left(\frac{b^2}{4a}\right) \ ,$$

where $b > 0$ and $a \in \mathbb{C}$.

Problem 2: One gene topologies.

In this question we will take a look at some basic gene regulation phenomenology. In the one gene case, only two interactions are possible:



(a) self-activation **(b)** self-inhibition

i.e. the more protein there is, protein production can either increase or decrease. The pointy arrow corresponds to activation (higher p leads to higher production), while the flat arrow corresponds to inhibition (higher p leads to *lower* production).

Most quantitative biologists start by writing models down like this (genes connected to other genes, i.e. directed graphs with two types of arrows) rather than writing down reactions and the CME. Without the CME, we have to associate arrows in these graphs with functions. These functions are *phenomenological*, because they are really just guesses for what the proper dynamics should be; if you want a rigorously valid model, you should write down reactions and the CME first.

Let activation correspond to $A(x)$, and inhibition correspond to $I(x)$, where

$$A(x) := k_A \frac{x^n}{c^n + x^n} \qquad \text{and} \qquad I(x) := -k_I \frac{x^n}{c^n + x^n}$$

for some constants k_A, k_I, and c. These are called *Hill functions*, and are the most popular kind of functions to associate interactions with.

Let b be the basal (i.e. when all interactions are ignored) protein production rate, and d be the decay rate.

(a) Suppose that $k_A = 1$ and $c = 1/2$. Plot $A(x)$ for $n = 1$, $n = 3$, and $n = 5$. What do you notice as you increase n?

(b) Write down an ODE according to the rule

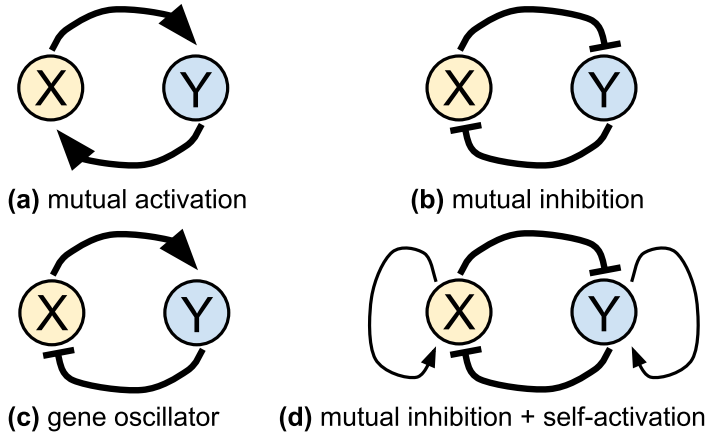
$$\dot{x} = (\text{basal production rate}) + (\text{all interactions hitting x}) - (\text{decay term})$$

for each of the one gene topologies pictured above.

(c) How do these equations compare to the SDEs we derived using the CME and some QSS approximations? How are they similar and how are they different?

Problem 3: Two gene topologies.

In this question we will take a look at more basic gene regulation phenomenology. In the one gene case, only two interactions were possible; however, in the two gene case, many more topologies are possible. Here are a few:



Let activation coming from gene z correspond to $A(z)$, and inhibition coming from gene z correspond to $I(z)$, where

$$A(z) := k_A \frac{z^n}{c^n + z^n} \qquad \text{and} \qquad I(z) := -k_I \frac{z^n}{c^n + z^n} \ .$$

Let b_x and b_y be the basal (i.e. when all interactions are ignored) protein production rates of X and Y respectively, and d_x and d_y be their decay rates.

(a) Let's do some guessing. Consider the mutual inhibition topology. If X and Y inhibit each other equally strongly, and there is initially a lot of x and not a lot of y, what state do you expect the system to settle into? What if there is initially a lot of y and not a lot of x? How many steady states do you think this system has?

(b) Consider the gene oscillator topology. This one is a little bit special compared to the others, for a reason we are about to see. Suppose that there is initially a lot of x and not a lot of y. What happens next? As the amount of y increases, how does the amount of x change? If there is not any x to activate the production of y, suppose that d_y is high enough that y tends to decay away. If there is not any y to inhibit x, suppose x has a high enough basal production rate b_x that x tends to increase. Can you see why this is called a gene oscillator?

(c) Write down ODEs (for $\dot{x}$ and $\dot{y}$) according to the rule

$$\dot{z} = (\text{basal production rate}) + (\text{all interactions hitting z}) - (\text{decay term})$$

for each of the two gene topologies pictured above.