BE 159 Spring 2016
Homework #1
Due at the start of lecture, January 20, 2016.

Problem 1.1 (Toy gap gene profiles and mutual information)**.**
This problem was inspired by some of the discussion in the Ph.D. thesis of Julien Dubuis, Princeton University, 2012. For this problem, as in the Dubuis, et al. paper, we will assume

$$P(g|x) \approx \frac{1}{\sqrt{2\pi\sigma_g^2}} \exp\left\{-\frac{(g-\bar{g}(x))^2}{2\sigma_g^2}\right\}. \tag{1.1}$$

We will assume σ_g is constant (not a function of x), but $\bar{g} = \bar{g}(x)$. Finally, as in the Dubuis, et al. paper, we will take $P_x(x)$ to be uniform, i.e., $P_x = 1$.

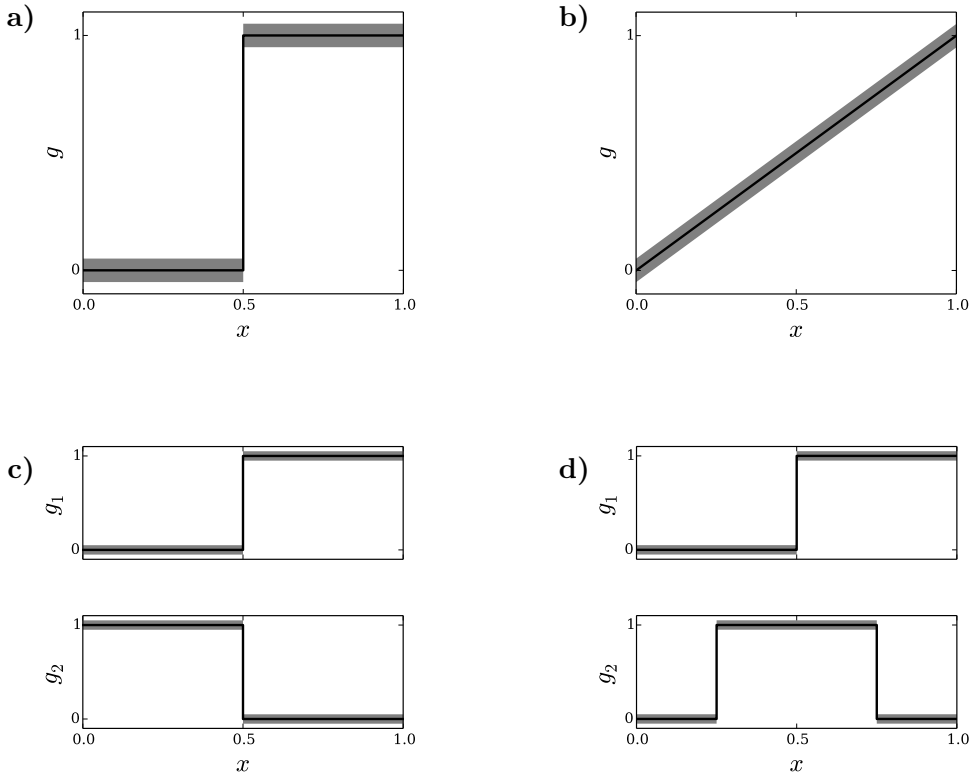

Figure 1: Figure adapted from the Ph.D. thesis of J. Dubuis, Princeton University, 2012. Each figure represents a gene expression profile. Lines represent $\bar{g}(x)$ and the shaded areas represent $\bar{g}(x) \pm \sigma_g$, where σ_g is a constant.

a) Compute the mutual information between normalized position x and normalized level of gene expression level g for the profile given in Fig. 1a in the limit of $\sigma_g = 0$. Is $I_{g\to x}$ for the case where $\sigma_g = 0$ an underestimate of overestimate of the mutual information when $\sigma_g > 0$? *Hint:* When $\sigma_g = 0$, g becomes a discrete variable.

b) We will now compute the approximate mutual information for the expression profile in Fig. 1b.

i) In our calculation, we will take $P_g(g) \approx \theta(g) - \theta(g-1)$, where θ denotes the Heaviside step function. Explain why this approximation is reasonable.

ii) Show that in the limit of low gene expression noise (small but nonzero σ_g), the mutual information for the expression profile in Fig. 1b is approximately

$$I_{g\to x} \approx -\frac{1}{2}\log_2\left(2\pi e\sigma_g^2\right). \tag{1.2}$$

Is this approximation an overestimate or an underestimate of the mutual information? *Hints:* The position x must strictly follow $0 \le x \le 1$, but g does not have to be between zero and one, since only $\bar{g}$ goes from zero to one. Let's say $g_{\min} \le g \le g_{\max}$. Because $P(g|x)$ is Gaussian, $P(g|x)$ is negligible for $|g| \gg \bar{g}$, so there is little error introduced by performing integrals with infinite bounds. In other words,

$$\begin{aligned} I_{g\to x} &= \int_0^1 \mathrm{d}x \int_{g_{\min}}^{g_{\max}} \mathrm{d}g\, P_x(x)P(g|x)\log_2\frac{P(g|x)}{P_g(g)} \\ &\approx \int_0^1 \mathrm{d}x \int_{-\infty}^{\infty} \mathrm{d}g\, P_x(x)P(g|x)\log_2\frac{P(g|x)}{P_g(g)}. \end{aligned} \tag{1.3}$$

Also, remember some identities for Gaussian integrals.

$$\int_{-\infty}^{\infty} \mathrm{d}u\, \mathrm{e}^{-u^2} = \sqrt{\pi}, \tag{1.4}$$

$$\int_{-\infty}^{\infty} \mathrm{d}u\, u^2\mathrm{e}^{-u^2} = \frac{\sqrt{\pi}}{2}. \tag{1.5}$$

c) Compute the mutual information between $\{g_1, g_2\}$ and x for the profiles shown in Figures 1c and d in the limit where $\sigma_{g_1} = \sigma_{g_2} = 0$.

d) In summary, what do the results of parts (a), (b), and (c) say about "design principles" for informative expression profiles?

Problem 1.2 (Mutual information from Hunchback profiles, 20 pts extra credit)**.**
This problem is inspired by problem 137 of Bialek, Biophysics: Searching for Principles, *Princeton University Press, 2012*. In this problem we will investigate how the mutual information $I_{g\to x}$ is calculated from real data. We will do this for the profile of a single gap gene, Hunchback. On the website of Bialek's book, he made measurements of the Hunchback profile from 20 embryos available. (Note that these are not the same measured profiles from the Dubuis paper.) These can be downloaded here: http://be159.caltech.edu/2016/handouts/hb.csv. Each column gives the normalized profile of a single embryo. The data are evenly spaced, going from normalized position $x = 0$ to $x = 1$.

Your task is to use these data to compute the mutual information content between the gene expression level of Hunchback, g, and the position along the anterior-posterior axis in the embryo, x. As in the paper, use only the middle 80% of the profile in your analysis. We will not carry out the more sophisticated analysis used in the Dubuis papers, but will make the following approximations.

i) We assume that at each position x, the gene expression level is Gaussian. I.e.,

$$P(g|x) = \frac{1}{\sqrt{2\pi\sigma_g^2}} \exp\left\{-\frac{(g-\bar{g}(x))^2}{2(\sigma_g(x))^2}\right\}, \tag{1.6}$$

where $\bar{g}(x)$ and $\sigma_g(x)$ are computed directly from the experimental data.

ii) We assume the the data are of sufficient quality that both $\bar{g}(x)$ and $\sigma_g(x)$ vary smoothly with x such that we can naively use numerical quadrature as if the data accurately represent continuous functions.

We will not do further analysis to get an error bar on our mutual information, as done in the paper.

So, with these approximations in hand, compute $I_{g\to x}$ from the data you downloaded.