COGS 202: HW 4

Attempt these problems on your own. Solutions will be presented by an assigned group next Monday.

1. (Math) Given the observations $x_1, \ldots, x_t$, iid samples from hidden random variable $s \in \{s_1, s_2\}$, and assuming the prior over $s$ is uniform ($P(s = s_1) = P(s = s_2) = .5$), show that the log posterior ratio is a sum of the individual log likelihood ratios:

$$\log \frac{p(s_1|x_1, \ldots, x_t)}{p(s_2|x_1, \ldots, x_t)} = \sum_{i=1}^{t} \log \frac{p(x_i|s_1)}{p(x_i|s_2)} \tag{0-1}$$

2. (Math) Assume that each $x_i$ is generated from $s$ independently via a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$, where $\mu = 2$ if $s = s_1$, and $\mu = -2$ if $s = s_2$:

$$p(x|s) = \mathcal{N}(\mu_s, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)²}{2\sigma²}}.$$ 

(a) (Matlab) Generate 500 samples of $x$ for $s = s_1$, and $\sigma \in \{5, 6, 8, 10\}$. For each $\sigma$, plot a histogram of the empirical distribution of $x$. Is the distribution widest for the largest $\sigma$?

(b) (Math) Show that the log likelihood ratio, $d_i := \log \frac{p(x_i|s_1)}{p(x_i|s_2)}$, is

$$d_i = \frac{-(x_i - \mu_1)^2 + (x_i - \mu_2)^2}{2\sigma²}.$$ 

(c) (Matlab) For each $\sigma$, plot a histogram of the log likelihood values $\{d_i\}$ corresponding to the 500 $\{x_i\}$ values generated in (a); indicate the sample mean and standard deviation for each. Why is the distribution widest for the smallest $\sigma$?

3. (a) (Matlab) Assume the same generative model as the last problem, generate 80 sequences of $x_i$ (100 samples in each sequence) for $s = s_1$, and $\sigma \in \{5, 6, 8, 10\}$. For each $\sigma$, plot the average cumulative log likelihood ratio as a function of samples within a sequence (from 1 to 100); use different colored lines in the same figure.

(b) (Math) Suppose that the cumulative log likelihood ratio after observing $x_1, \ldots, x_t$ is $r_t = \sum_{i=1}^{t} d_i$, show that the posterior probability is $P(s = s_1|x_1, \ldots, x_t) = \frac{\exp(r_t)}{1 + \exp(r_1)}$.

(c) (Matlab) Convert the average cumulative log likelihood ratio for each data point in (a) to posterior probability, and plot this quantity for the different $\sigma$ in the same figure.

(d) (Math) Formally derive the probability distribution over $d_i$ as a function of the prior $P(s_1)$, $\sigma$, and $\mu$, assuming a Gaussian generative model as before. Are these properties reflected in your simulation results?

4. (a) (Matlab) Use same generative model as the last two problems. For each $\sigma$, use Matlab to generate as many samples as necessary until the cumulative log likelihood ratio $r_i$ first crosses the threshold $z = \exp(\eta)/(1 + \exp(\eta))$ or $-z$; where $\eta = 0.95$. Decide $s = s_1$ if $z$ is first crossed, and $s = s_2$ if $-z$ is first crossed; record the number of samples required to exceed the threshold as the "response time" $\tau_i$ for that "trial." Simulate 50 such "trials", and plot the empirical histogram of the "accuracy" for each $\sigma$. Separately, plot the empirical histogram of the "response time" for each $\sigma$.

(b) (Math) You should have found that the "response time" increases for increasing $\sigma$, but the accuracy
is approximately the same. Prove why the accuracy should be $\eta$ or slightly bigger, and independent of $\sigma$.

(c) (Matlab) Assume the cost function to be

$$l(\hat{s}, s) = (1 - \delta_{s, \hat{s}}) + c\tau$$

Let $c = 0.005$. We will try to find the optimal stopping threshold $z^*$ that minimizes the cost function. Set $z$ to different values corresponding to probabilities between 0.5 and 0.95, in small increments, generate 50 trials of data, and evaluate the cost function in each trial; compute the average cost function for each $z$ and find the value of $z$ that minimizes the average loss function. Now set $c = 0.05$ and see how that changes the optimal $z^*$. You should find that $z^*$ decreased for larger $c$, why?