Boulder Theoretical Biophysics 2019

Neuroscience Mini-course: Exercise Set 4 **Solutions**

1. Squared error distortion: We consider a continuous random variable with mean zero and variance σ^2 of unknown distribution under squared-error distortion, which can be written as

$$D = E\left[d(x,\hat{x})\right] = E\left[(x-\hat{x})^2\right]$$

and which is, on average, the familiar concept of mean-squared error.

This time, instead of finding a single rate-distortion function, we are finding upper and lower bounds on the rate-distortion function over all possible sources.

We can begin by finding the lower bound, assuming that we have random variables X and $\hat{X}$ such that $E\left[(X-\hat{X})^2\right] \leq D$,

$$\begin{aligned}
I(X;\hat{X}) &= H(X) - H(X|\hat{X}) \\
&= H(X) - H(X-\hat{X}|\hat{X}) \\
&\geq H(X) - H(X-\hat{X}) \\
&\geq H(X) - H\left(\mathcal{N}\left(0, E\left[(X-\hat{X})^2\right]\right)\right) \\
&= H(X) - \frac{1}{2}\log\left[2\pi e\, E\left[(X-\hat{X})^2\right]\right] \\
I(X;\hat{X}) &\geq H(X) - \frac{1}{2}\log\left[2\pi e D\right]
\end{aligned}$$

For the upper bound, we consider the channel

$$\hat{X} = \frac{\sigma^2 - D}{\sigma^2}(X+Z)$$

where $Z \sim \mathcal{N}\left(0, \frac{D\sigma^2}{\sigma^2 - D}\right)$.

2. First, we verify that this channel operates with our desired distortion,

$$\begin{aligned}
E\left[(X-\hat{X})^2\right] &= E\left[\left(X - \frac{\sigma^2 - D}{\sigma^2}(X+Z)\right)^2\right] \\
&= E\left[\left(\frac{D}{\sigma^2}X - \frac{\sigma^2 - D}{\sigma^2}Z\right)^2\right] \\
&\quad \text{because } X \text{ and } Z \text{ are independent} \\
&= \left(\frac{D}{\sigma^2}\right)^2 E[X^2] + \left(\frac{\sigma^2 - D}{\sigma^2}\right)^2 E[Z^2] \\
&= \left(\frac{D}{\sigma^2}\right)^2 \sigma^2 + \left(\frac{\sigma^2 - D}{\sigma^2}\right)^2 \frac{D\sigma^2}{\sigma^2 - D} \\
&= D
\end{aligned}$$

Back to part 1. Next, we find the second moment of $\hat{X}$,

$$\begin{aligned}
E\left[\hat{X}^2\right] &= \left(\frac{\sigma^2 - D}{\sigma^2}\right)^2 E\left[(X+Z)^2\right] \\
&= \left(\frac{\sigma^2 - D}{\sigma^2}\right)^2 \left(E\left[X^2\right] + E\left[Z^2\right]\right) \\
&= \left(\frac{\sigma^2 - D}{\sigma^2}\right)^2 \left(E\left[X^2\right] + E\left[Z^2\right]\right) \\
&= \left(\frac{\sigma^2 - D}{\sigma^2}\right)^2 \left(\sigma^2 + \frac{D\sigma^2}{\sigma^2 - D}\right) \\
&= \sigma^2 - D
\end{aligned}$$

Now, using both of these and, again, leveraging the fact that the gaussian maximizes entropy for a constrained variance, we can calculate our upper bound on the mutual information,

$$\begin{aligned}
I(X;\hat{X}) &= H(\hat{X}) - H(\hat{X}|X) \\
&= H(\hat{X}) - H\left(\frac{\sigma^2 - D}{\sigma^2} Z\right) \\
&= H(\hat{X}) - H(Z) - \log \frac{\sigma^2 - D}{\sigma^2} \\
&\quad \text{see Theorem 8.6.4 in Cover \& Thomas} \\
&= H(\hat{X}) - \frac{1}{2}\log\left[2\pi e \frac{D\sigma^2}{\sigma^2 - D}\right] - \frac{1}{2}\log\left(\frac{\sigma^2 - D}{\sigma^2}\right)^2 \\
&= H(\hat{X}) - \frac{1}{2}\log\left[2\pi e \frac{D(\sigma^2 - D)}{\sigma^2}\right] \\
&\leq \frac{1}{2}\log\left[2\pi e(\sigma^2 - D)\right] - \frac{1}{2}\log\left[2\pi e \frac{D(\sigma^2 - D)}{\sigma^2}\right] \\
I(X;\hat{X}) &\leq \frac{1}{2}\log\frac{\sigma^2}{D}
\end{aligned}$$

3. Now that we have shown both our upper and lower bounds, we can observe (for the last part of the question) that, iff X is Gaussian, the lower bound is equal to the upper bound – otherwise, the lower bound is strictly less than the upper bound. Thus, the readout of a Gaussian X is "harder" than the readout of an X with any other distribution, in the sense that the $R(D)$ function is greater for a Gaussian than for any other distribution (to further clarify, more bits are required to describe a Gaussian at a particular level of distortion than are required for any other distribution at the same level of distortion).

4. Information bottleneck: Now, we want to explore an important product of rate-distortion theory known as the information bottleneck. In particular, we want to minimize the function

$$\min_{p(\hat{x}|x)} \mathcal{L} = I(X;\hat{X}) - \beta I(\hat{X};Y) - \sum_x \lambda(x) \left(\sum_{\hat{x}} p(\hat{x}|x) - 1 \right)$$

where our distortion is now a somewhat more abstract quantity,

$$d(x,\hat{x}) = -I(\hat{X};Y)$$

but has an easy interpretation: We want to minimize our information rate, as before, but now we want to maximize the amount of information $I(\hat{X};Y)$ that we transmit about Y, some quantity that is encoded in X. Note that,

$$I(X;Y) \geq I(\hat{X};Y)$$

To begin, we obtain

$$\begin{aligned}
\frac{\delta p(\hat{x})}{\delta p(\hat{x}|x)} &= \frac{\delta}{\delta p(\hat{x}|x)} \left[\sum_x p(\hat{x}|x) p(x) \right] \\
&= p(x)
\end{aligned}$$

and

$$\begin{aligned}
\frac{\delta p(\hat{x}|y)}{\delta p(\hat{x}|x)} &= \frac{\delta}{\delta p(\hat{x}|x)} \left[\sum_x p(\hat{x}|x) p(x|y) \right] \\
&= p(x|y)
\end{aligned}$$

5. Next, we obtain

$$\begin{aligned}
\frac{\delta \mathcal{L}}{\delta p(\hat{x}|x)} &= \frac{\delta}{\delta p(\hat{x}|x)} \left[I(X;\hat{X}) - \beta I(\hat{X};Y) - \sum_x \lambda(x) \left(\sum_{\hat{x}} p(\hat{x}|x) - 1 \right) \right] \\
&= \frac{\delta}{\delta p(\hat{x}|x)} I(X;\hat{X}) - \beta \frac{\delta}{\delta p(\hat{x}|x)} I(\hat{X};Y) \\
&\quad - \frac{\delta}{\delta p(\hat{x}|x)} \sum_x \lambda(x) \left(\sum_{\hat{x}} p(\hat{x}|x) - 1 \right)
\end{aligned}$$

and we will approach each of these terms in turn.

First,

$$
\begin{aligned}
\frac{\delta}{\delta p(\hat{x}|x)} I(X;\hat{X}) &= \frac{\delta}{\delta p(\hat{x}|x)} \left[H(\hat{X}) - H(\hat{X}|X)\right] \\
&= \frac{\delta}{\delta p(\hat{x}|x)} \left[-\sum_{\hat{x}} p(\hat{x}) \log p(\hat{x}) + \sum_{x} p(x) \sum_{\hat{x}} p(\hat{x}|x) \log p(\hat{x}|x)\right] \\
& \quad \text{due to fixed } x,\ \hat{x} \\
&= \frac{\delta}{\delta p(\hat{x}|x)} \left[-p(\hat{x}) \log p(\hat{x}) + p(x) p(\hat{x}|x) \log p(\hat{x}|x)\right] \\
& \quad \text{using the product rule twice} \\
&= -p(x) \log p(\hat{x}) - p(x) + p(x) \log p(\hat{x}|x) + p(x) \\
&= p(x) \log \frac{p(\hat{x}|x)}{p(\hat{x}x)}
\end{aligned}
$$

Then,

$$
\begin{aligned}
\frac{\delta}{\delta p(\hat{x}|x)} \beta I(\hat{X};Y) &= -\beta \frac{\delta}{\delta p(\hat{x}|x)} \sum_{\hat{x},y} p(\hat{x},y) \log \left[\frac{p(\hat{x},y)}{p(\hat{x})p(y)}\right] \\
&= -\beta \frac{\delta}{\delta p(\hat{x}|x)} \sum_{y} \left[p(\hat{x}|y)p(y) \log p(\hat{x}|y) - p(y|\hat{x})p(\hat{x}) \log p(\hat{x})\right] \\
&= -\beta \frac{\delta}{\delta p(\hat{x}|x)} \left[\sum_{y} p(\hat{x}|y)p(y) \log p(\hat{x}|y) - p(\hat{x}) \log p(\hat{x})\right] \\
&= -\beta \left[\sum_{y} p(x|y)p(y) \log p(\hat{x}|y) + p(x|y)p(y) - p(x) \log p(\hat{x}) - p(x)\right] \\
&= -\beta \sum_{y} p(x|y)p(y) \left[\log p(\hat{x}|y) + 1\right] + \beta p(x) \left[\log p(\hat{x}) + 1\right]
\end{aligned}
$$

Finally,

$$
\frac{\delta}{\delta p(\hat{x}|x)} \sum_{x} \lambda(x) \left(\sum_{\hat{x}} p(\hat{x}|x) - 1\right) = \lambda(x)
$$

Putting these all together,

$$
\begin{aligned}
\frac{\delta \mathcal{L}}{\delta p(\hat{x}|x)} &= p(x)\log\frac{p(\hat{x}|x)}{p(\hat{x})} - \beta\sum_y p(x|y)p(y)\left[\log p(\hat{x}|y)+1\right] \\
&\quad +\beta p(x)\left[\log p(\hat{x})+1\right] - \lambda(x) \\
&= p(x)\log\frac{p(\hat{x}|x)}{p(\hat{x})} + \beta p(x)\sum_y p(y|x)\log\frac{p(\hat{x}|y)}{p(\hat{x})} - \lambda(x) \\
&\quad \text{using Bayes rule} \\
&= p(x)\log\frac{p(\hat{x}|x)}{p(\hat{x})} + \beta p(x)\sum_y p(y|x)\log\frac{p(y|\hat{x})}{p(y)} - \lambda(x) \\
&= p(x)\left[\log\frac{p(\hat{x}|x)}{p(\hat{x})} + \beta\sum_y p(y|x)\log\frac{p(y|\hat{x})}{p(y)} - \frac{\lambda(x)}{p(x)}\right]
\end{aligned}
$$

Now, we observe that

$$
\log\frac{p(y|\hat{x})}{p(y)} = -\log\frac{p(y|x)}{p(y|\hat{x})} + \log\frac{p(y|x)}{p(y)}
$$

and so, with

$$
\tilde{\lambda}(x) = \frac{\lambda(x)}{p(x)} - \beta\sum_y p(y|x)\log\frac{p(y|x)}{p(y)}
$$

we can rewrite $\mathcal{L}$, set it to zero, and solve for $p(\hat{x}|x)$:

$$
\begin{aligned}
0 = \frac{\delta \mathcal{L}}{\delta p(\hat{x}|x)} &= p(x)\left[\log\frac{p(\hat{x}|x)}{p(\hat{x})} + \beta\sum_y p(y|x)\log\frac{p(y|x)}{p(y|\hat{x})} - \tilde{\lambda}(x)\right] \\
p(\hat{x}|x) &= p(\hat{x})\exp\left[-\beta\sum_y p(y|x)\log\frac{p(y|x)}{p(y|\hat{x})} - \tilde{\lambda}(x)\right] \\
p(\hat{x}|x) &= p(\hat{x})\exp\left[-\beta D_{KL}[p(y|x)||p(y|\hat{x})] - \tilde{\lambda}(x)\right] \\
p(\hat{x}|x) &= \frac{p(\hat{x})}{Z(x,\beta)}\exp\left[-\beta D_{KL}[p(y|x)||p(y|\hat{x})]\right] \\
p(\hat{x}|x) &\propto p(\hat{x})\exp\left[-\beta D_{KL}[p(y|x)||p(y|\hat{x})]\right]
\end{aligned}
$$

using

$$
D_{KL}[Q(i)||P(i)] = \sum_i Q(i)\log\frac{Q(i)}{P(i)}
$$

and

$$
Z(x,\beta) = \exp(\tilde{\lambda}(x)) = \sum_{\hat{x}} p(\hat{x})\exp(-\beta D_{KL}[p(y|x)||p(y|\hat{x})])
$$