These notes will take you through the two formulations of the spike train information as presented in Brenner *et al.* (2000). Our goal is to compute the mutual information between the spiking response from a neuron and the input stimulus, $I(\text{stimulus}; \text{response})$. (Here will use lowercase s to denote the stimulus and σ to denote response. Capital S denotes the entropy.) We will express this information in units of bits/spike. We have seen that there are several ways to write down the mutual information, and that the mutual information is symmetric in its arguments. Here, we will take two equivalent perspectives on computing the mutual information and will show that the information we obtain is indeed the same.

Preliminaries: The stimulus is a time-varying input, $s(t)$, that runs from time $t = 0$ to $t = T$, and is presented several times during an experiment. Let the label on our response variable be σ, signifying the number of spikes we observe in our cell in response to the stimulus. In our experiment, the response we observe is the time of spikes ($\sigma = 1$) in a neuron. This spike train can be represented by $\rho(t)$, a sum of Kronecker delta functions at those spike times, $\{t_j\}$,

$$\rho(t) = \sum_j \delta(t - t_j). \tag{1}$$

We may bin time during the experiment in small segments, Δt, so that the number of time bins, $N_t = T/\Delta t$. During this experiment we have recorded M repeats of the stimulus, which allows us to compute the time-dependent probability of observing a spike during the recording, $P(\text{spike}|t)$. Dividing $P(\text{spike}|t)$ by ΔT yields a probability per unit time, which is the usual definition of the instantaneous firing rate, $r(t) = P(\text{spike}|t)/\Delta t$. Later we will take the limit that $\Delta t \to 0$ to obtain results for the continuous time limit. In general, we assume that Δt is small enough that we only observe two response states within any given bin: either we see a single spike $\sigma = 1$, or no spike $\sigma = 0$.

One of the main conceptual difficulties one might have with the approach laid out in this paper is the 'ergodic' assumption the authors make about the stimulus sampling. We presume here that the stimulus $s(t)$ is long enough to sample the full dynamic range of s, so that we obtain the full entropy of the distribution $P(s)$ via the trajectory $s(t)$. This assumption tends to work if T is much longer than the longest correlation time in s. To put things plainly, if the stimulus prior gives a probability $P(s = s_i)$[1] for some particular stimulus value, s_i, then during $s(t)$ this particular stimulus value should occur at a set of n_i time points, $\{t^i_k\}$, such that $n_i/N_t = P(s = s_i)$. Since the mapping between s and t is injective, meaning that there are several t's which correspond to the same s, but only one s at any given t, the union of all $\{t^i_k\}$'s is equal to the set of all times. This allows us to write

$$\sum_{s_i} P(s_i) f(s_i) = \sum_{s_i} \frac{n_i}{N_t} f(s_i) = \frac{1}{N_t} \sum_{t_i} f(t_i). \tag{2}$$

[1] To be clear, I will write $P(s = s_i)$ to denote the probability of observing a particular stimulus value, s_i. I will sometimes take a short cut and just write this as $P(s_i)$, but it means the same thing.

This is the essence of *Monte Carlo* sampling, where the expected value of a function is estimated using a random sample average. Here, $s(t)$ generates our samples of the function we want to estimate, such as $P(\text{spike}|s)$.

You might find it difficult to imagine that you have sampled the full range of possible stimulus values when viewing, say, natural movies, which tend to have a range of correlation times stretching into long time correlations on the order of minutes and hours. This ergodic assumption should certainly be scrutinized under such stimulus conditions.

We will put that question aside for now and work in a regime where $s(t)$ samples $P(s)$ fully. This allows us to swap stimulus averages, $\sum_{s_i} P(s_i)f(s_i)$, for time averages, $\sum_{t_i} 1/N_t f(t_i)$.

Perspective A: $I(s;\sigma) = S(s) - S(s|\sigma)$. Here we take the perspective that the spikes we observe constrain the set of stimuli we have likely been shown. In other words, the spikes carry information about what stimulus was played.

First, let's compute the prior entropy on the stimulus. Let us take the stimulus to be comprised of a set of discrete values, $\{s_i\}$. The entropy of the full distribution of stimuli we observe (without any reference to the spikes) is

$$S(\text{stimulus}) = S(s) = -\sum_i P(s_i)\log_2(P(s_i)). \tag{3}$$

To compute the mutual information, we also need to compute the conditional entropy, and average over s. Writing out the conditional entropy, we have

$$S(s|\sigma) = \sum_\sigma P(\sigma)\left[-\sum_s P(s|\sigma)\log_2 P(s|\sigma)\right] = \left\langle -\sum_s P(s|\sigma)\log_2 P(s|\sigma)\right\rangle_\sigma \tag{4}$$

where $\langle\cdot\rangle_\sigma$ indicates an average over $P(\sigma)$. To compute this, we need $P(s|\sigma)$. Using Bayes' Rule, we rewrite this as

$$P(s|\sigma) = \frac{P(\sigma|s)P(s)}{P(\sigma)}. \tag{5}$$

Now the entropy can be written (in full gory detail) as

$$\begin{aligned} S(s|\sigma) &= \sum_{\sigma=1,\sigma=0} P(\sigma)\left[-\sum_{s_i}\frac{P(\sigma|s_i)P(s_i)}{P(\sigma)}\log_2\left(\frac{P(\sigma|s_i)P(s_i)}{P(\sigma)}\right)\right] \\ &= -\sum_{s_i}\left[P(\sigma=1)\frac{P(\sigma=1|s_i)P(s_i)}{P(\sigma=1)}\log_2\left(\frac{P(\sigma=1|s_i)P(s_i)}{P(\sigma=1)}\right)\right. \\ &\qquad \left. +\ P(\sigma=0)\frac{P(\sigma=0|s_i)P(s_i)}{P(\sigma=0)}\log_2\left(\frac{P(\sigma=0|s_i)P(s_i)}{P(\sigma=0)}\right)\right]. \end{aligned} \tag{6}$$

Subtracting the conditional entropy from the prior entropy, we obtain the mutual information between stimulus and response

$$\begin{aligned} I(s;\sigma) =& S(s) - S(s|\sigma) \\ =& -\sum_i P(s_i)\log_2(P(s_i)) \\ &+ \sum_{s_i}\left[P(\sigma=1)\frac{P(\sigma=1|s_i)P(s_i)}{P(\sigma=1)}\log_2\left(\frac{P(\sigma=1|s_i)P(s_i)}{P(\sigma=1)}\right)\right. \\ &\left.+ P(\sigma=0)\frac{P(\sigma=0|s_i)P(s_i)}{P(\sigma=0)}\log_2\left(\frac{P(\sigma=0|s_i)P(s_i)}{P(\sigma=0)}\right)\right]. \end{aligned} \tag{7}$$

Since we only have two response states, 0 (no spike) and 1 (single spike in a bin), the probability of no spike is just $P(\sigma = 0) = 1 - P(\sigma = 1)$. Substituting in these expressions for the probability and conditional probabilities of silences

$$\begin{aligned} I(s;\sigma) = -\sum_{s_i} P(s_i)\left[\log_2(P(s_i)) - P(\sigma=1|s_i)\log_2\left(\frac{P(s_i)P(\sigma=1|s_i)}{P(\sigma=1)}\right)\right. \\ \left. - (1-P(\sigma=1|s_i))\log_2\left(\frac{P(s_i)(1-P(\sigma=1|s_i))}{1-P(\sigma=1)}\right)\right] \end{aligned} \tag{8}$$

and gathering terms depending on $\log_2(P(s))$, we obtain

$$\begin{aligned} I(s;\sigma) = &\left[-\sum_{s_i} P(s_i)\log_2(P(s_i))\left[1 - P(\sigma=1|s_i) - (1-P(\sigma=1|s_i))\right]\right] \\ &+ \sum_{s_i} P(s_i)\left[P(\sigma=1|s_i)\log_2\left(\frac{P(\sigma=1|s_i)}{P(\sigma=1)}\right)\right. \\ &\left.+ (1-P(\sigma=1|s_i))\log_2\left(\frac{(1-P(\sigma=1|s_i))}{1-P(\sigma=1)}\right)\right]. \end{aligned} \tag{9}$$

The terms that depend on $\log_2(P(s))$ cancel and we are left with the second and third lines in the previous equation.

We assume that T is long enough so that we can substitute a time average for the stimulus average, $\sum_{s_i} P(s_i)f(s_i) = 1/N_t \sum_{t_i} f(t_i)$. We rewrite the probability of observing a spike given a particular stimulus as a function of the instantaneous firing rate. In particular, $P(\sigma = 1|s = s(t_i)) = r(t_i)\Delta t$, and $P(\sigma = 1) = \bar{r}\Delta t$. Substituting these expressions into the equation for the mutual information, we obtain

$$I(s;\sigma) = \frac{1}{N_t}\sum_{t_i}\left[r(t_i)\Delta t\log_2\left(\frac{r(t_i)}{\bar{r}}\right) + (1-r(t_i)\Delta t)\log_2\left(\frac{(1-r(t_i)\Delta t)}{1-\bar{r}\Delta t}\right)\right]. \tag{10}$$

We now take the limit of very small bin size, $\Delta t \to 0$. Recall that for small x, $\log(1 - x) \approx -x$. The mutual information between the stimulus and the response is approximately

$$\lim_{\Delta t\to 0} I(s;\sigma) \approx \frac{1}{N_t}\sum_{t_i}\left[r(t)\Delta t\log_2\left(\frac{r(t)}{\bar{r}}\right) + (1-r(t)\Delta t)[-r(t)\Delta t + \bar{r}\Delta t]\right]. \tag{11}$$

Keeping terms that have up to a linear dependence on Δt, we have

$$\lim_{\Delta t \to 0} I(s;\sigma) \approx \frac{1}{N_t} \sum_{t_i} \left[r(t_i)\Delta t \log_2 \left(\frac{r(t_i)}{\bar{r}} \right) + [-r(t_i)\Delta t + \bar{r}\Delta t] \right]. \tag{12}$$

We have that $1/N_t \sum_{t_i} r(t) = \bar{r}$, so that the second set of terms in this equation cancel. We are left with

$$\lim_{\Delta t \to 0} I(s;\sigma) \approx \frac{1}{N_t} \sum_{t_i} \left[r(t)\Delta t \log_2 \left(\frac{r(t)}{\bar{r}} \right) \right]. \tag{13}$$

Finally, this is an expression for the mutual information in bits *per bin* between the stimulus and the response. To express this as a bit rate, we can divide by the bin size, Δt, or to express it in terms of the number of bits *per spike*, we divide by the average number of spikes per bin $P(\sigma = 1) = \bar{r}\Delta t$. This gives

$$I(s;\sigma) = \frac{1}{N_t} \sum_{t_i} \left[\frac{r(t)}{\bar{r}} \log_2 \left(\frac{r(t)}{\bar{r}} \right) \right]. \tag{14}$$

We can rewrite this in a few ways, recalling that $N_t = T/\Delta t$, and that we have taken $\Delta t \to 0$ in a continuous time limit:

$$I(s;\sigma) = \frac{1}{T} \int_0^T dt \frac{r(t)}{\bar{r}} \log_2 \left(\frac{r(t)}{\bar{r}} \right) = \left\langle \frac{r(s)}{\bar{r}} \log_2 \left(\frac{r(s)}{\bar{r}} \right) \right\rangle_{P(s)}, \tag{15}$$

which emphasizes the ergodic assumption we have made in our sampling of the stimulus via our experimental trajectory of length T.

Perspective B: $I(\sigma; s) = S(\sigma) - S(\sigma|s)$. Now we take a slightly different perspective, from the point of view of how much we know about when a spike will occur during our experiment. Before we observe the stimulus, we have no reason to believe a spike is any more likely at one time or another (no idea when the stimulus starts or ends, what it is or how temporally modulated it it). This neglects any history dependence there may be in the firing of our neuron. For example, if we observe a spike at some time, we might know that we are unlikely to observe another spike for a few milliseconds. Here, we assume that the arrival times of spikes are independent events, so that our prior on the arrival time of a spike is flat (independent of time). The stimulus defines the set of likely times when a spike will occur. We compute the difference in the entropy of a flat prior distribution on the time of the spike and the observed PSTH, which is the conditional distribution of time of the spike given the stimulus at that time, $s(t)$.

The response here is the arrival time of a spike, which can take values between 0 and T. To compute the information the stimulus conveys about the arrival time of a spike, we first compute our prior entropy on the response.

$$S(P(\sigma = 1(t))) = -\sum_i P_{\sigma=1}(t) \log_2(P_{\sigma=1}(t)) \tag{16}$$

We will adopt a bit of shorthand here, specifically $P(\sigma = 1 \textit{ at time } t) = P(t)$, and similarly for the conditional distribution, $P(\sigma = 1 \textit{ at time } t|s(t)) = P(t|s)$. Next, we need to compute the entropy of the probability distribution of spike arrival times conditioned on the stimulus. (Again, we are making the assumption that the long sample T of the stimulus properly samples all stimulus states, s.) Writing down the conditional entropy, we obtain

$$S(P(t|s)) = \left\langle -\int_0^T dt P(t|s)\log_2(P(t|s))\right\rangle_{P(s)} \tag{17}$$

Once more, we assume that integrating over t effectively averages over $P(s)$, so we drop the $\langle\cdot\rangle$, presuming we've sufficiently averaged with the $1/T\int dt\,\cdot$.

The distribution, $P(t|s)$, is proportional to the PSTH, $P(\sigma = 1|s(t))$, but normalized so that $\int_0^T P(t|s)dt = 1$. Let's run through that normalization,

$$\int_0^T dt P(t|s) \propto \int_0^T dt\,\text{PSTH} = \int_0^T dt\, r(t)\Delta t = T\bar{r}\Delta t, \tag{18}$$

so that we obtain

$$P(t|s) = \frac{r(t)\Delta t}{T\bar{r}\Delta t} = \frac{r(t)}{T\bar{r}}. \tag{19}$$

Now that we have the conditional distribution, let's compute its entropy

$$S(P(t|s)) = -\int_0^T dt\frac{r(t)}{T\bar{r}}\log_2\left(\frac{r(t)}{T\bar{r}}\right). \tag{20}$$

The mutual information between the arrival time of a spike and the stimulus is just

$$\begin{aligned} I(\sigma;s) &= S(P(t)) - S(P(t|s)) \qquad (21)\\ &= \log_2 T + \frac{1}{T}\int_0^T dt\frac{r(t)}{\bar{r}}\log_2\left(\frac{r(t)}{T\bar{r}}\right)\\ &= \log_2 T + \frac{1}{T}\int_0^T dt\frac{r(t)}{\bar{r}}\left[\log_2\left(\frac{r(t)}{\bar{r}}\right) - \log_2(T)\right]. \end{aligned}$$

Note that

$$\frac{1}{T}\int_0^T dt\frac{r(t)}{\bar{r}} = \frac{1}{\bar{r}}\left[\frac{1}{T}\int_0^T dt\cdot r(t)\right] = \frac{1}{\bar{r}}\bar{r} = 1. \tag{22}$$

Let's multiply that first term in the information by this factor of 1 so that

$$\begin{aligned} I(\sigma;s) &= \log_2 T + \frac{1}{T}\int_0^T dt\frac{r(t)}{\bar{r}}\left[\log_2\left(\frac{r(t)}{\bar{r}}\right) - \log_2(T)\right] \qquad (23)\\ &= \frac{1}{T}\int_0^T dt\frac{r(t)}{\bar{r}}\left[\log_2 T + \log_2\left(\frac{r(t)}{\bar{r}}\right) - \log_2(T)\right]\\ &= \frac{1}{T}\int_0^T dt\frac{r(t)}{\bar{r}}\log_2\left(\frac{r(t)}{\bar{r}}\right). \end{aligned}$$

We can swap the time average for the stimulus average, assuming long T, such that

$$I(\sigma; s) = \frac{1}{T}\int_0^T dt \frac{r(t)}{\bar{r}} \log_2\left(\frac{r(t)}{\bar{r}}\right) = \left\langle \frac{r(s)}{\bar{r}} \log_2\left(\frac{r(s)}{\bar{r}}\right)\right\rangle_{P(s)}, \tag{24}$$

which is exactly what we obtained in Perspective A. We have explicitly shown that the information is symmetric, namely that $I(s; \sigma) = I(\sigma; s)$.

To be completely transparent, we have computed in Perspective B the information that the stimulus conveys about the arrival time, which can take values between 0 and T, of a single spike. We have assumed that spikes carry information about the stimulus, but could just as easily compute the information that the arrival times of silences carry about the stimulus. In the limit of small time bins, the silences will carry zero information. One should always be careful to consider information in silences when your average spike rate is high, meaning your bin size is not, in this sense, small.