Introduction

When it comes to correlations and neural coding, our premise is that the question

> "Do correlations among spike trains carry information above and beyond what can be obtained from the individual spike trains?"

is the same as

> "If the brain were to construct the conditional probability distribution of stimuli given responses *without knowledge of the correlational structure in the response distribution*, would that conditional distribution be close to the true one?"

The nice thing about this reformulation is that the second question can be phrased mathematically as

> "Is $P(s|r_1, r_2, ...)$ close to $P_{IND}(s|r_1, r_2, ...)$?"

where $P(s|r_1, r_2, ...)$ is the true conditional stimulus distribution and $P_{IND}(s|r_1, r_2, ...)$ is the one built without any knowledge of the correlational structure in the responses.

The only remaining issue is to define "close to". We used a pretty common measure, the conditional relative entropy between P and P_{IND}. We called this measure ΔI, and it's given by

$$\Delta I = \sum_r P(r) \sum_s P(s|r) \log \frac{P(s|r)}{P_{IND}(s|r)}$$

where r is shorthand for $(r_1, r_2, ...)$. See Nirenberg et al. (2001) [*Nature* **411**:698-701] for details.

When it comes down to it, cost functions are always a matter of taste. However, there are several good reasons for choosing ΔI: 1) ΔI provides a universal bound on optimal cost functions; 2) ΔI is an upper bound on information loss; and 3) when it comes to comparing codes, ΔI is completely standard, and it's what everybody's been using for years.

In the next few pages we make the above statements precise, and we prove them. Note: these are separate documents, so equation numbers are internal to each of them.

Optimal cost functions and ΔI

Let's say we want to build a deterministic decoder based on neuronal responses. In other words, we want to construct a mapping that takes the neuronal response, $\mathbf{r}$, to an estimate of the stimulus, $\hat{\mathbf{s}}(\mathbf{r})$, such that the difference between the true stimulus, $\mathbf{s}$, and the estimated stimulus, $\hat{\mathbf{s}}(\mathbf{r})$, is as small as possible. "As small as possible", of course, means with respect to some cost function, $C(\hat{\mathbf{s}}(\mathbf{r}), \mathbf{s})$. The total cost is some functional of $C(\hat{\mathbf{s}}(\mathbf{r}), \mathbf{s})$; here we'll use the average, denoted $\langle C(\hat{\mathbf{s}})\rangle_p$,

$$\langle C(\hat{\mathbf{s}})\rangle_p = \int d\mathbf{r}\, p(\mathbf{r}) \int d\mathbf{s}\, p(\mathbf{s}|\mathbf{r}) C(\hat{\mathbf{s}}(\mathbf{r}), \mathbf{s}) \,.$$

The estimator that minimizes the average cost, denoted $\hat{\mathbf{s}}_p(\mathbf{r})$, is

$$\hat{\mathbf{s}}_p(\mathbf{r}) = \arg\min_{\hat{\mathbf{s}}} \int d\mathbf{s}\, p(\mathbf{s}|\mathbf{r}) C(\hat{\mathbf{s}}, \mathbf{s}) \,.$$

Suppose we don't know the true distribution $p(\mathbf{s}|\mathbf{r})$; instead we know only an approximate distribution, $q(\mathbf{s}|\mathbf{r})$. If we minimized the average cost with respect to $q(\mathbf{s}|\mathbf{r})$, we would get a different estimator, $\hat{\mathbf{s}}_q$, which would be given by

$$\hat{\mathbf{s}}_q(\mathbf{r}) = \arg\min_{\hat{\mathbf{s}}} \int d\mathbf{s}\, q(\mathbf{s}|\mathbf{r}) C(\hat{\mathbf{s}}, \mathbf{s}) \,. \tag{1}$$

The difference between the two costs, denoted ΔC, is given by

$$\Delta C = \langle C(\hat{\mathbf{s}}_q)\rangle_p - \langle C(\hat{\mathbf{s}}_p)\rangle_p \,. \tag{2}$$

Note that, even though $\hat{\mathbf{s}}_q$ was constructed using $q(\mathbf{r}|\mathbf{s})$, the cost associated with $\hat{\mathbf{s}}_q$ is found by averaging with respect to the true distribution.

We want to compute ΔC in the limit that p is close to q, and then compare that to ΔI (defined in Eq. (9) below). We can find $\hat{\mathbf{s}}_q(\mathbf{r})$ by minimizing the right hand side of Eq. (1) with respect to $\hat{\mathbf{s}}$. In other words, $\hat{\mathbf{s}}_q(\mathbf{r})$ is a solution to the equation

$$\int d\mathbf{s}\, q(\mathbf{s}|\mathbf{r}) \nabla C(\hat{\mathbf{s}}_q(\mathbf{r}), \mathbf{s}) = 0 \tag{3}$$

where the gradient is with respect to $\hat{\mathbf{s}}$: $\nabla C(\hat{\mathbf{s}}, \mathbf{s}) \equiv \partial C(\hat{\mathbf{s}}, \mathbf{s})/\partial \hat{\mathbf{s}}$. Expanding $\hat{\mathbf{s}}_q$ around $\hat{\mathbf{s}}_p$ and q around p, and working to lowest order in $(p - q)$, Eq. (3) becomes

$$\int d\mathbf{s} \left[p(\mathbf{s}|\mathbf{r})(\hat{\mathbf{s}}_q - \hat{\mathbf{s}}_p) \cdot \nabla\nabla C(\hat{\mathbf{s}}_p, \mathbf{s}) + [q(\mathbf{s}|\mathbf{r}) - p(\mathbf{s}|\mathbf{r})] \nabla C(\hat{\mathbf{s}}_p, \mathbf{s}) \right] = 0 \tag{4}$$

where we used the condition $\int ds\, p(\mathbf{s}|\mathbf{r}) \nabla C(\hat{\mathbf{s}}_p, \mathbf{s}) = 0$. Solving Eq. (4) for $\hat{\mathbf{s}}_q - \hat{\mathbf{s}}_p$ yields

$$\hat{\mathbf{s}}_q - \hat{\mathbf{s}}_p = \langle \nabla\nabla C(\hat{\mathbf{s}}_p, \mathbf{s}) \rangle^{-1}_{p(\mathbf{s}|\mathbf{r})} \cdot \langle \nabla C(\hat{\mathbf{s}}_p, \mathbf{s}) [q(\mathbf{s}|\mathbf{r}) - p(\mathbf{s}|\mathbf{r})]/p(\mathbf{s}|\mathbf{r}) \rangle_{p(\mathbf{s}|\mathbf{r})} \,. \tag{5}$$

The notation $\langle ... \rangle_{p(\mathbf{s}|\mathbf{r})}$ means average over $\mathbf{s}$ with respect to the distribution $p(\mathbf{s}|\mathbf{r})$.

Now that we know $\hat{\mathbf{s}}_q$ in terms of $\hat{\mathbf{s}}_p$ we can compute ΔC. Taylor expanding the first term in Eq. (2) around $\hat{\mathbf{s}}_p$, we find, to second order in $\hat{\mathbf{s}}_p - \hat{\mathbf{s}}_q$, that

$$\Delta C = \langle C(\hat{\mathbf{s}}_p) \rangle_p + \langle (\hat{\mathbf{s}}_q - \hat{\mathbf{s}}_p) \cdot \nabla C(\hat{\mathbf{s}}_p, \mathbf{s}) \rangle_p + \langle (\hat{\mathbf{s}}_q - \hat{\mathbf{s}}_p) \cdot \nabla\nabla C(\hat{\mathbf{s}}_p, \mathbf{s}) \cdot (\hat{\mathbf{s}}_q - \hat{\mathbf{s}}_p) \rangle_p - \langle C(\hat{\mathbf{s}}_p, \mathbf{s}) \rangle_p \,. \tag{6}$$

Again using $\int ds\, p(\mathbf{s}|\mathbf{r}) \nabla C(\hat{\mathbf{s}}_p, \mathbf{s}) = 0$, Eq. (6) becomes

$$\Delta C = \langle (\hat{\mathbf{s}}_q - \hat{\mathbf{s}}_p) \cdot \nabla\nabla C(\hat{\mathbf{s}}_p) \cdot (\hat{\mathbf{s}}_q - \hat{\mathbf{s}}_p) \rangle_p \,. \tag{7}$$

Inserting Eq. (5) into (7) then yields

$$\Delta C = \int d\mathbf{r}\, p(\mathbf{r}) \langle (\delta p/p) \nabla C(\hat{\mathbf{s}}_p, \mathbf{s}) \rangle_{p(\mathbf{s}|\mathbf{r})} \cdot \langle \nabla\nabla C(\hat{\mathbf{s}}_p, \mathbf{s}) \rangle^{-1}_{p(\mathbf{s}|\mathbf{r})} \cdot \langle \nabla C(\hat{\mathbf{s}}_p, \mathbf{s}) (\delta p/p) \rangle_{p(\mathbf{s}|\mathbf{r})} \tag{8}$$

where $\delta p/p$ is shorthand for $[p(\mathbf{s}|\mathbf{r}) - q(\mathbf{s}|\mathbf{r})]/p(\mathbf{s}|\mathbf{r})$.

What we want to do now is compare this expression for ΔC to the one for ΔI. The latter is defined to be

$$\Delta I = \left\langle \log \frac{p(\mathbf{s}|\mathbf{r})}{q(\mathbf{s}|\mathbf{r})} \right\rangle_p \,. \tag{9}$$

Expanding this to lowest order in $(p-q)$ and using $\langle (p-q)/p \rangle_p = 0$, ΔI becomes, to lowest nonvanishing order in $(p-q)$,

$$\Delta I = \langle (\delta p/p)^2 \rangle_p \, . \tag{10}$$

To compare ΔI to ΔC, we need the following inequality. If $\mathbf{A}$ is symmetric and positive semi-definite, then, for any functions f and $\mathbf{g}$,

$$\begin{aligned} \langle f\mathbf{g} \rangle \cdot \mathbf{A} \cdot \langle \mathbf{g}\mathbf{f} \rangle &= \langle f\mathbf{g} \rangle \cdot \left(\sum_k \lambda_k \mathbf{v}_k \mathbf{v}_k \right) \cdot \langle \mathbf{g} f \rangle = \sum_k \lambda_k \langle f \mathbf{g} \cdot \mathbf{v}_k \rangle^2 \\ &\leq \sum_k \lambda_k \langle f^2 \rangle \langle \mathbf{g} \cdot \mathbf{v}_k^2 \rangle \\ &= \langle f^2 \rangle \left\langle \mathbf{g} \cdot \left(\sum_s \lambda_k \cdot \mathbf{v}_k \mathbf{v}_k \right) \cdot \mathbf{g} \right\rangle = \langle f^2 \rangle \langle \mathbf{g} \cdot \mathbf{A} \cdot \mathbf{g} \rangle \, . \end{aligned} \tag{11}$$

where the lone inequality in the above list of expressions follows from the Schwarz inequality, and λ_k and $\mathbf{v}_k$ are the eigenvalues and eigenvectors of $\mathbf{A}$.

We would like to use this inequality in Eq. (8), but we can do that only if $\langle \nabla\nabla C(\hat{\mathbf{s}}_p, \mathbf{s}) \rangle_{p(\mathbf{s}|\mathbf{r})}^{-1}$ is positive semi-definite. Fortunately, it is: $\hat{\mathbf{s}}_p$ was chosen to make $\langle C(\hat{\mathbf{s}}_p, \mathbf{s}) \rangle_{p(\mathbf{s}|\mathbf{r})}$ a minimum, which implies that $\langle \nabla\nabla C(\hat{\mathbf{s}}_p, \mathbf{s}) \rangle_{p(\mathbf{s}|\mathbf{r})}$ is positive semi-definite, so its inverse is also. Thus, using Eq. (11), Eq. (8) becomes

$$\Delta C \leq \int d\mathbf{r} \, p(\mathbf{r}) \langle (\delta p/p)^2 \rangle_{p(\mathbf{s}|\mathbf{r})} \Big[\langle \nabla C(\hat{\mathbf{s}}_p, \mathbf{s}) \cdot \langle \nabla\nabla C(\hat{\mathbf{s}}_p, \mathbf{s}) \rangle_{p(\mathbf{s}|\mathbf{r})}^{-1} \cdot \nabla C(\hat{\mathbf{s}}_p, \mathbf{s}) \rangle_{p(\mathbf{s}|\mathbf{r})} \Big] \, . \tag{12}$$

Comparing Eqs. (10) and (12), we see that, so long as $\langle \nabla\nabla C(\hat{\mathbf{s}}_p, \mathbf{s}) \rangle_{p(\mathbf{s}|\mathbf{r})}$ is invertible and ΔI is sufficiently small,

$$\frac{\Delta C}{\Delta I} \leq \int d\mathbf{r} \, \tilde{p}(\mathbf{r}) \Big[\langle \nabla C(\hat{\mathbf{s}}_p, \mathbf{s}) \cdot \langle \nabla\nabla C(\hat{\mathbf{s}}_p, \mathbf{s}) \rangle_{p(\mathbf{s}|\mathbf{r})}^{-1} \cdot \nabla C(\hat{\mathbf{s}}_p, \mathbf{s}) \rangle_{p(\mathbf{s}|\mathbf{r})} \Big]$$

where

$$\tilde{p}(\mathbf{r}) \equiv \frac{p(\mathbf{r}) \langle (\delta p/p)^2 \rangle_{p(\mathbf{s}|\mathbf{r})}}{\int d\mathbf{r} \, p(\mathbf{r}) \langle (\delta p/p)^2 \rangle_{p(\mathbf{s}|\mathbf{r})}} \, .$$

ΔI can act as an upper bound on information loss

Consider the following game: person A chooses, at random, one object out of a set of objects, and person B has to guess which one was chosen by asking yes/no questions. The i^{th} object is chosen with probability p_i. Person B, however, thinks the i^{th} object is chosen with probability q_i, and will base her question-asking strategy on that wrong distribution. Cover and Thomas (mainly Chap. 5) tells us that the average number of yes/no questions person B has to ask to guess the object, denoted $N(p, q)$, is

$$N(p, q) = -\sum_i p_i \log q_i = H(p) + \Delta I \tag{1}$$

where $H(p) \equiv -\sum_i p_i \log p_i$ is the entropy of the true distribution and $\Delta I \equiv \sum_i p_i \log p_i/q_i = D(p||q)$ is the Kullback-Leibler distance between p and q.

The quantity ΔI that appears in Eq. (1) has a natural interpretation: it is the penalty, in yes/no questions, that person B pays for using the wrong distribution to design her question-asking strategy. We can also interpret ΔI as an information loss, in the following sense: We can make up for the wrong distribution by supplying person B with ΔI bits. In other words, if we give ΔI bits to person B, on average she will do as well guessing what object is present as a person who knows the true distribution.

In fact, what we show below is stronger than that: if we give ΔI bits to person B, she will do *no worse* at guessing the object than a person who knows the true distribution, and she may do better. Alternatively, if we want person B to guess the object in $H(p)$ yes/no questions (the same number as a person who knows the true distribution), we could do that by supplying her with *at most* ΔI bits, and sometimes less than that. The actual number of bits she needs depends on the distributions p and q.

When we say "give bits to person B", we have in mind the following: Person A chooses an object, and then sends a string of symbols (0s and 1s, say) to person B through a noisy channel. Those strings provide information about which object was chosen using a pretty standard coding scheme: the objects are divided into groups, and each string tells which group the object is in. For example, a coding scheme for 6 objects might be: objects 1 and 2 are labeled with the string 1, objects 3 and 4 are labeled with the string 01, and objects 5 and 6 are labeled with the string 00. If, say, object 5 is chosen, then the string 00 would be sent. Since the channel is noisy, a string different than 00 might be received.

So here's the situation. Person A chooses object i, determines that it is labeled with string k, and then sends string k to person B. Because the channel through which the string is sent is noisy, person B receives string l. She then revises her estimate of the probability that object i is chosen. Letting $P(k|l)$ be the probability that string k was sent given that string l was received, her new estimate of the distribution of objects, which we'll call $q(i|l)$, is

$$q(i|l) = \frac{q_i}{Q_k} P(k|l) \tag{2}$$

where $Q_k \equiv \sum_{i \in k} q_i$ and the notation $i \in k$ means sum over only those i such that i is in group k. The number of yes/no questions she will have to ask to guess the object is now

$$N_I(p,q) = -\sum_l P(l) \sum_i p(i|l) \log q(i|l) \tag{3}$$

where the subscript I means that I bits were sent (I will be computed shortly), $P(l)$ is the probability that person B received string l, and $p(i|l)$ is the true probability that object i was chosen given that string l was received. Analogous to Eq. (2), this last quantity is given by

$$p(i|l) = \frac{p_i}{P_k} P(k|l) \tag{4}$$

where $P_k \equiv \sum_{i \in k} p_i$.

Inserting Eqs. (2) and (4) into (3), rearranging terms slightly, and using $P_k = \sum_l P(k|l) P(l)$, we find that

$$N_I(p,q) = N(p,q) - \sum_l P(l) \sum_k P(k|l) \log \frac{P(k|l)}{P_k} - \sum_k P_k \log \frac{P_k}{Q_k} .$$

The second term is $I(k;l)$, the amount of information transmitted through the noisy channel (also the average string length), and the third term is $D(P||Q)$, the Kullback-Leibler distance between P and Q. We thus have

$$N_I(p,q) = N(p,q) - I(k;l) - D(P||Q) .$$

Since $D(P||Q)$ is non-negative, giving I bits to person B reduces the number of yes/no questions she has to ask by *at least* I. The reduction could be larger, though. To find out how much larger, we minimize $N_I(p, q)$ with respect to q, subject to the constraint that the q_i sum to 1. When we do this, we find that the minimum value of $N_I(p, q)$, which occurs when $q_i = p_i Q_k / P_k$, is $H(p) - I(k;l)$. The maximum *reduction* in yes/no questions, $N(p, q) - [H(p) - I(k;l)]$, is thus $I(k;l) + \Delta I$ (see Eq. (1)). Consequently, giving I bits to person B reduces the number of yes/no questions she has to ask by an amount somewhere between I and $I + \Delta I$, inclusive. Alternatively,

> Giving ΔI bits **or less** to person B allows her to guess the object in exactly the same number of yes/no questions as someone who knows the true distribution.

This is our main result, and it's why we interpret ΔI as an upper bound on information loss.

This looks sort of odd: we can, in principle, give person B an arbitrarily small amount of information and produce a potentially large reduction in the number of yes/no questions. Can this really happen? The answer is yes, as the following example shows.

Let $p_i = 1/M$, $i = 1, ..., M$, $q_1 = 1/M - \alpha/M$, and $q_{i>1} = 1/M + \alpha/M(M-1)$. For these distributions

$$\Delta I = -\frac{1}{M} \log(1 - \alpha) - \frac{M-1}{M} \log\left[1 + \frac{\alpha}{M-1}\right] .$$

For fixed M, we can make ΔI arbitrarily large by letting α approach 1.

Now let's inject a little information by telling person B whether or not element 1 was chosen. We'll use a lossless channel, so this results in a transmission of

$$I = -\frac{1}{M} \log \frac{1}{M} - \frac{M-1}{M} \log \frac{M-1}{M}$$

bits, which, for large M, approaches $\log M/M$. Thus, we can provide person B with an arbitrarily small amount of information (by letting M go to infinity), while reducing the number of yes/no questions she would have to ask by an arbitrarily large amount (by letting α go to 1).

Information differences and ΔI

Suppose you want to compare two neural codes – say spike timing and spike count. The natural thing to do is compute the information using one code, compute the information using the other, and then take the difference. What we show here is that when one of the neural codes is a sub-code of the other (as spike count is a sub-code of spike timing), then this difference is exactly equal to the ΔI, the cost function used to assess the role of correlations in Nirenberg et al. (2001) [*Nature* **411**:698-701].

As usual in information calculations in the brain, you show a set of stimuli and measure neuronal responses. The latter are denoted $\mathbf{r} \equiv (r_1, r_2, ...)$ where the r_i can be any aspect of the code – 1s and 0s in small bins to indicate the presence or absence of a spike, for example. A sub-code of $\mathbf{r}$ is any function of $\mathbf{r}$: if $\mathbf{z} = \mathbf{f}(\mathbf{r})$ then $\mathbf{z}$ is a sub-code of $\mathbf{r}$. If you observe just $\mathbf{z}$, you get no more information than if you observe $\mathbf{r}$, and you usually get less. The difference is

$$\Delta \hat{I} = I(\mathbf{r}; s) - I(\mathbf{z}; s) \,,$$

where s is the stimulus and

$$\begin{aligned} I(\mathbf{r}; s) &= -\sum_{\mathbf{r}} P(\mathbf{r}) \log P(\mathbf{r}) + \sum_{s} P(s) \sum_{\mathbf{r}} P(\mathbf{r}|s) \log P(\mathbf{r}|s) \\ I(\mathbf{z}; s) &= -\sum_{\mathbf{z}} P(\mathbf{z}) \log P(\mathbf{z}) + \sum_{s} P(s) \sum_{\mathbf{z}} P(\mathbf{z}|s) \log P(\mathbf{z}|s) \,. \end{aligned}$$

The probability distributions $P(\mathbf{z}|s)$ and $P(\mathbf{z})$ are given by the usual formulae

$$P(\mathbf{z}) = \sum_{\mathbf{r}} P(\mathbf{r}) \delta(\mathbf{z} - \mathbf{f}(\mathbf{r})) \tag{1a}$$

$$P(\mathbf{z}|s) = \sum_{\mathbf{r}} P(\mathbf{r}|s) \delta(\mathbf{z} - \mathbf{f}(\mathbf{r})) \,. \tag{1b}$$

Here δ is a Kronecker δ-like object: $\delta(\mathbf{z} - \mathbf{f}(\mathbf{r})) = 1$ if $\mathbf{z} = \mathbf{f}(\mathbf{r})$ and 0 otherwise. Had these been continuous distributions, we would have used Dirac δ-functions and the sums would have been integrals.

Let's compare $\Delta \hat{I}$ to ΔI, the latter being the number of extra yes-no questions it would take to guess the stimulus given that you observed only $\mathbf{z} = \mathbf{f}(\mathbf{r})$ rather than the full set of responses, $\mathbf{r}$. This quantity is given by [Nirenberg et al., *Nature* **411**:698-701 (2001)]

$$\Delta I = \sum_{\mathbf{r}} P(\mathbf{r}) \sum_s P(s|\mathbf{r}) \log \left[\frac{P(s|\mathbf{r})}{P(s|\mathbf{f}(\mathbf{r}))} \right] .$$

Using Bayes' theorem and rearranging terms slightly leads to

$$\Delta I = \sum_s P(s) \sum_{\mathbf{r}} P(\mathbf{r}|s) \log \left[\frac{P(\mathbf{r}|s)P(s)}{P(\mathbf{r})} \; \frac{P(\mathbf{f}(\mathbf{r}))}{P(\mathbf{f}(\mathbf{r})|s)P(s)} \right] .$$

Canceling the $P(s)$ that appears in the numerator and denominator inside the logs, and again rearranging terms, we find that

$$\Delta I = I(\mathbf{r}; s) - \left[-\sum_{\mathbf{r}} P(\mathbf{r}) \log P(\mathbf{f}(\mathbf{r})) + \sum_s P(s) \sum_{\mathbf{r}} P(\mathbf{r}|s) \log P(\mathbf{f}(\mathbf{r})|s) \right] . \tag{2}$$

We can rewrite the first term in brackets as

$$\sum_{\mathbf{r}} P(\mathbf{r}) \log P(\mathbf{f}(\mathbf{r})) = \sum_{\mathbf{r}} P(\mathbf{r}) \sum_{\mathbf{z}} \delta(\mathbf{z} - \mathbf{f}(\mathbf{r})) \log P(\mathbf{z}) .$$

Rearranging terms one last time, we have

$$\sum_{\mathbf{r}} P(\mathbf{r}) \log P(\mathbf{f}(\mathbf{r})) = \sum_{\mathbf{z}} \log P(\mathbf{z}) \sum_{\mathbf{r}} P(\mathbf{r}) \delta(\mathbf{z} - \mathbf{f}(\mathbf{r})) = \sum_{\mathbf{z}} P(\mathbf{z}) \log P(\mathbf{z}) \tag{3}$$

where the last equality follows from Eq. (1a). Using identical logic,

$$\sum_{\mathbf{r}} P(\mathbf{r}|s) \log P(\mathbf{f}(\mathbf{r})|s) = \sum_{\mathbf{z}} P(\mathbf{z}|s) \log P(\mathbf{z}|s) . \tag{4}$$

Finally, Inserting Eqs. (3) and (4) into Eq. (2), we find that

$$\Delta I = I(\mathbf{r}; s) - \left[-\sum_{\mathbf{z}} P(\mathbf{z}) \log P(\mathbf{z}) + \sum_{\mathbf{z}} P(\mathbf{z}|s) \log P(\mathbf{z}|s) \right] = I(\mathbf{r}; s) - I(\mathbf{z}; s) = \Delta \hat{I} .$$

Thus, $\Delta \hat{I}$ and ΔI are one and the same.