Optimal cost functions and $\Delta I$

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, $r$, to an estimate of the stimulus, $\hat{s}(r)$, such that the difference between the true stimulus, $s$, and the estimated stimulus, $\hat{s}(r)$, is as small as possible. “As small as possible”, of course, means with respect to some cost function, $C(\hat{s}(r), s)$. The total cost is some functional of $C(\hat{s}(r), s)$; here we’ll use the average, denoted $\langle C(\hat{s}) \rangle_p$,

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

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

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

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

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

The difference between the two costs, denoted $\Delta C$, is given by

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

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

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

$$\int ds \, q(s|r) \nabla C(\hat{s}_q(r), s) = 0 \quad (3)$$
where the gradient is with respect to \( \hat{s} \): \( \nabla C(\hat{s}, s) \equiv \partial C(\hat{s}, s)/\partial \hat{s} \). Expanding \( \hat{s}_q \) around \( \hat{s}_p \) and \( q \) around \( p \), and working to lowest order in \( (p - q) \), Eq. (3) becomes

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

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

\[
\hat{s}_q - \hat{s}_p = \langle \nabla \nabla C(\hat{s}_p, s) \rangle_p^{-1} \cdot \langle \nabla C(\hat{s}_p, s)|q(s|r) - p(s|r)|/p(s|r)\rangle_p . \] (5)

The notation \( \langle ... \rangle_p \) means average over \( s \) with respect to the distribution \( p(s|r) \).

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

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

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

\[
\Delta C = \langle (\hat{s}_q - \hat{s}_p) \cdot \nabla \nabla C(\hat{s}_p) \cdot (\hat{s}_q - \hat{s}_p) \rangle_p . \] (7)

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

\[
\Delta C = \int dr \ p(r) \langle (\delta p/p)\nabla C(\hat{s}_p, s) \rangle_{p(s|r)} \cdot \langle \nabla \nabla C(\hat{s}_p, s) \rangle_p^{-1} \cdot \langle \nabla C(\hat{s}_p, s)(\delta p/p) \rangle_{p(s|r)} \) \] (8)

where \( \delta p/p \) is shorthand for \( \frac{p(s|r) - q(s|r)}{p(s|r)} \).

What we want to do now is compare this expression for \( \Delta C \) to the one for \( \Delta I \). The latter is defined to be

\[
\Delta I = \langle \log \frac{p(s|r)}{q(s|r)} \rangle_p . \] (9)
Expanding this to lowest order in \((p - q)\) and using \(\langle (p - q)/p \rangle_p = 0\), \(\Delta I\) becomes, to lowest nonvanishing order in \((p - q)\),

\[
\Delta I = \langle (\delta p/p)^2 \rangle_p .
\]  

To compare \(\Delta I\) to \(\Delta C\), we need the following inequality. If \(A\) is symmetric and positive semi-definite, then, for any functions \(f\) and \(g\),

\[
\langle fg \rangle \cdot A \cdot \langle gf \rangle = \langle fg \rangle \cdot \left( \sum_k \lambda_k v_k v_k \right) \cdot \langle gf \rangle = \sum_k \lambda_k \langle fg \cdot v_k \rangle^2 \\
\leq \sum_k \lambda_k \langle f^2 \rangle \langle g \cdot v_k^2 \rangle \\
= \langle f^2 \rangle \left( g \cdot \left( \sum_k \lambda_k \cdot v_k v_k \right) \cdot g \right) = \langle f^2 \rangle \langle g \cdot A \cdot g \rangle .
\]  

where the lone inequality in the above list of expressions follows from the Schwarz inequality, and \(\lambda_k\) and \(v_k\) are the eigenvalues and eigenvectors of \(A\).

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

\[
\Delta C \leq \int d\mathbf{r} p(\mathbf{r}) \langle (\delta p/p)^2 \rangle_p \left[ \langle \nabla C(\hat{s}_p, s) \cdot \langle \nabla \nabla C(\hat{s}_p, s) \rangle_p^{-1} \cdot \nabla C(\hat{s}_p, s) \rangle_p \right] .
\]  

Comparing Eqs. (10) and (12), we see that, so long as \(\langle \nabla \nabla C(\hat{s}_p, s) \rangle_p\) is invertible and \(\Delta I\) is sufficiently small,

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

where

\[
\tilde{p}(\mathbf{r}) \equiv \frac{p(\mathbf{r}) \langle (\delta p/p)^2 \rangle_p}{\int d\mathbf{r} p(\mathbf{r}) \langle (\delta p/p)^2 \rangle_p} .
\]