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

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

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

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

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 $\mathbf{g}$,

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

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 $\langle \nabla\nabla C(\hat{\mathbf{s}}_p, \mathbf{s})\rangle^{-1}_{p(\mathbf{s}|\mathbf{r})}$ 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

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

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,

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

where

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