Derivations NIP course

Mark van Rossum

Version 1. 8th February 2017

Some notation:

- δ_{ij}: Kronecker delta. $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. $i, j \in \mathbb{N}$.
- $\delta(x-y)$: Dirac delta. $\int dx\, f(x)\delta(x-y) = f(y)$. Can be seen as infinitely narrow (and high) Gaussian distribution, or a as continuous version of the Kronecker delta.
- $\partial_x = \frac{\partial}{\partial x}$: partial derivative w.r.t. x.
- $\langle x \rangle$: expectation (see lectures for interpretation).
- $\mathcal{N}(\mu, \sigma^2)$: Gaussian distribution with mean μ and variance σ^2.

1 Encoding lectures

1.1 Higher moments of Gaussian

Q: Consider an uncorrelated (white) Gaussian signal $s(t)$. Calculate its 3rd and 4th order moments, $m_3 = \langle s(t_1)s(t_2)s(t_3)\rangle$ and $m_4 = \langle s(t_1)s(t_2)s(t_3)s(t_4)\rangle$. Confirm with simulation.

The third moment $m_3 = \langle s(t_1)s(t_2)s(t_3)\rangle$ is zero. Namely, if $t_1 \neq t_2 \neq t_3$ then $m_3 = \langle s\rangle^3 = 0^3 = 0$. If $t_1 = t_2 = t_3$, $m_3 = \langle s^3\rangle = \int \mathcal{N}(0, \sigma^2)x^3 dx = 0$, as this is the integral over an odd function. For $t_1 = t_2 \neq t_3$, $m_3 = \langle s\rangle \langle s^2\rangle = 0.\sigma^2 = 0$.

Consider $m_4 = \langle s(t_1)s(t_2)s(t_3)s(t_4)\rangle$. Here we have contributions when $t_1 = t_2$ and $t_3 = t_4$ and permutation

$$\begin{aligned} m_4 &= \delta(t_1-t_2)\delta(t_3-t_4)\left\langle s^2\right\rangle^2 + \delta(t_1-t_3)\delta(t_2-t_4)\left\langle s^2\right\rangle^2 + \delta(t_1-t_4)\delta(t_2-t_3)\left\langle s^2\right\rangle^2 \\ &= [\delta(t_1-t_2)\delta(t_3-t_4) + \delta(t_1-t_3)\delta(t_2-t_4) + \delta(t_1-t_4)\delta(t_2-t_3)]\sigma^4 \end{aligned}$$

The fact that there is no additional contribution when $t_1 = t_2 = t_3 = t_4$, is related to the fourth cumulant of the Gaussian being zero.

In Matlab you can test this by trying things like `mean(randn(1,n).*randn(1,n).*randn(1,n).*randn(1,n))` and `mean(randn(1,n).^2*randn(1,n).*randn(1,n))`.

1.2 Wiener Kernel

Q: Suppose that the response of a system is given by a 2nd order Wiener approximation:

$$r(t) = g_0 + \int dt_1 g_1(t_1)s(t-t_1) + \int dt_1 dt_2 g_2(t_1, t_2)s(t-t_1)s(t-t_2) - \sigma^2 \int dt_1 g_2(t_1, t_1)$$

Show that when the stimulus is Gaussian with variance σ^2, one indeed can extract the kernels from the correlates of stimulus with the response without contamination of higher orders, i.e. show that $\langle r\rangle = g_0$ and $\langle r(t)s(t-\tau)\rangle = \sigma^2 g_1(\tau)$.

Zero order term:

$$
\begin{aligned}
\langle r\rangle &= g_0 + \left\langle \int dt_1 g_1(t_1)s(t-t_1)\right\rangle + \left\langle \int dt_1 dt_2 g_2(t_1,t_2)s(t-t_1)s(t-t_2) - \sigma^2 \int dt_1 g_2(t_1,t_1)\right\rangle \\
&= g_0 + \int dt_1 \langle s\rangle\, g_1(t_1) + \int dt_1 dt_2 g_2(t_1,t_2)\, \langle s(t-t_1)s(t-t_2)\rangle - \sigma^2 \int dt_1 g_2(t_1,t_1) \\
&= g_0 + 0 + \int dt_1 dt_2 g_2(t_1,t_2)\delta(t_1,t_2)\sigma^2 - \sigma^2 \int dt_1 g_2(t_1,t_1) \\
&= g_0
\end{aligned}
$$

First order. Similar calculation gives

$$
\begin{aligned}
\langle r(t)s(t-\tau)\rangle &= 0 + \left\langle \int dt_1 g_1(t_1)s(t-t_1)s(t-\tau)\right\rangle + 0) \\
&= \int dt_1 g_1(t_1)\langle s(t-t_1)s(t-\tau)\rangle = \sigma^2 \int dt_1 g_1(t_1)\delta(\tau-t_1) = \sigma^2 g_1(\tau)
\end{aligned}
$$

1.3 Discrete time kernel

Q: Assume a linear system, discrete in time, described by 0th and 1st order kernels, so that we can write $\hat{\mathbf{r}} = S\mathbf{g}$ *(see lecture slides for the construction of* S *and definition of* $\mathbf{g}$*). Derive the kernels* g_0 *and* g_1^i *that minimizes mean square error* $E = (\boldsymbol{r} - S\boldsymbol{g})^T(\boldsymbol{r} - S\boldsymbol{g})$*.*

Minimize $E = (\boldsymbol{r} - S\boldsymbol{g})^T(\boldsymbol{r} - S\boldsymbol{g})$ w.r.t. to $\boldsymbol{g}$, where $\boldsymbol{g} = (g_0, g_1^1, \ldots, g_1^N)$. For definition of S see lecture notes.

$$
\begin{aligned}
\frac{dE}{dg_j} &= \frac{d}{dg_j}[(\boldsymbol{r} - S\boldsymbol{g})^T(\boldsymbol{r} - S\boldsymbol{g})] \\
&= -2(\boldsymbol{r} - S\boldsymbol{g})\frac{dS\boldsymbol{g}}{dg_j} \\
&= -2\sum_i (\boldsymbol{r} - S\boldsymbol{g})_i \frac{d}{dg_j}\sum_k S_{ik}g_k \\
&= -2\sum_i (\boldsymbol{r} - S\boldsymbol{g})_i \sum_k S_{ik}\delta_{j,k} \\
&= -2\sum_i (\boldsymbol{r} - S\boldsymbol{g})_i S_{ij} \\
&= -2(S^T\boldsymbol{r})_j + 2(S^T S\boldsymbol{g})_j
\end{aligned}
$$

As this needs to be zero for all j, $S^T\boldsymbol{r} = S^T S\boldsymbol{g}$, or $\boldsymbol{g} = (S^T S)^{-1}S^T\boldsymbol{r}$. This is an equation you often see in linear regression. Note that at this point we have not used anything about the stimulus.

$S^T S = \sum_j S_{ij}S_{kj}$, i.e. it equals the dot product of the columns of S. If S is a design matrix for a Gaussian stimulus then the expected value is $\left\langle S^T S\right\rangle = \mathrm{diag}(\mathrm{n}, \mathrm{n}\sigma^2, \mathrm{n}\sigma^2, \mathrm{n}\sigma^2, \ldots)$, or, $\langle S^T S\rangle_{ij} = n\delta_{i,k}(\sigma^2 + (1-\sigma^2)\delta_{i,1})$. So

$$
\begin{aligned}
g_0 =& \frac{1}{n}\sum_i S_{0i}r_i = \langle r\rangle \\
g_1^i =& \frac{1}{n\sigma^2}\sum S_{ij}r_j = \frac{1}{n\sigma^2}\sum S_{i-j}r_j
\end{aligned}
$$

This is a discrete formulation of the first two Wiener kernels.

1.4 Optimality of STA as stimulus

Q: Assume a linear system described by 0th and 1st order kernels, $r(t) = r_0 + \int dt_1 g_1(t_1)s(t-t_1)$*. Derive the stimulus that maximizes the response* $r(t)$*. Constrain the 'energy' of the stimulus to be* $\int s(t)^2 dt = 1$*.*

The response of linear system is $r(t) = r_0 + \int dt_1 g_1(t_1) s(t - t_1)$, with Lagrange multiplier to constrain $|s|^2$, we maximize $f(t) = r(t) - \lambda \int dt_1 s^2(t_1)$ wrt $s(t)$. Note, this is a functional derivative; if unfamiliar, you can discretize in time (sums become integrals).

$$
\begin{aligned}
\frac{\delta f(t)}{\delta s(t_0)} &= \frac{\delta r(t)}{\delta s(t_0)} - \lambda \frac{\delta \int dt_1 s^2(t_1)}{\delta s(t_0)} \\
&= 0 + \int dt_1 g_1(t_1) \frac{\delta s(t - t_1)}{\delta s(t_0)} - \lambda \int dt_1 \delta s^2(t_1)/\delta s(t_0) \delta(t_0 - t_1) \\
&= \int dt_1 g_1(t_1) \delta(t_0 - t + t_1) - 2\lambda \int dt_1 s(t_1) \delta(t_0 - t_1) \\
&= g_1(t - t_0) - 2\lambda s(t_o)
\end{aligned}
$$

Set to zero for all t_0, hence $s(t_0) \propto g_1(t - t_0)$.

Extra: what happens for different constraints such as $\int s(t)^k dt = 1$?

2 Decoding lecture

2.1 Fisher Information Gaussian noise

Q: Assume N neurons with Gaussian independent noise and tuning curves $f_i(\theta)$ $(i = 1 \dots N)$. Show that the Fisher info equals $I_f(\theta) = \frac{1}{\sigma^2}\sum_{i=1}^N [f_i'(\theta)^2]$.

The Fisher information is defined as $I_f \equiv -\int P(\mathbf{r}|\theta)\frac{\partial^2 \log P(\mathbf{r}|\theta)}{\partial\theta^2}d\mathbf{r}$. For independent. Gaussian noise $P(\mathbf{r}|\theta) = \prod_{i=1}^N P(r_i|\theta) = \prod_{i=1}^N \frac{1}{Z}\exp(-[r_i - f_i(\theta)]^2/2\sigma^2)$.

So that

$$\begin{aligned}\frac{\partial^2 \log P(\mathbf{r}|\theta)}{\partial\theta^2} &= \left(-N\log Z - \sum_i [r_i - f_i(\theta)]^2/2\sigma^2\right)'' \\ &= \frac{1}{\sigma^2}\left(\sum_i [r_i - f_i(\theta)]f_i'(\theta)\right)' \\ &= \frac{1}{\sigma^2}\sum_i \left[(r_i - f_i(\theta))f_i''(\theta) - (f_i'(\theta))^2\right]\end{aligned}$$

Now use that $\int P(\mathbf{r}|\theta)d\mathbf{r} = 1$ and $\int P(\mathbf{r}|\theta)r_i d\mathbf{r} = f_i$. Also note that for this factorizing probability and a general function $g(r_i)$ then $\int d\mathbf{r}P(\mathbf{r}|\theta)\sum_i g(r_i) = \sum_i \int dr_i P(r_i|\theta)g(r_i)$. So that

$$\begin{aligned}I_f(\theta) &= -\int P(\mathbf{r}|\theta)\frac{1}{\sigma^2}\sum_i [(r_i - f_i(\theta))f_i''(\theta) - (f_i'(\theta))^2]d\mathbf{r} \\ &= -\frac{1}{\sigma^2}\sum_i \left\{f_i''(\theta)\int P(r_i|\theta)[r_i - f_i(\theta)]dr_i - (f_i'(\theta))^2\int P(r_i|\theta)dr_i\right\} \\ &= \frac{1}{\sigma^2}\sum_{i=1}^N [f_i'(\theta)^2].\end{aligned}$$

2.2 Fisher Information Poisson noise

Q: Assume N neurons with independent Poisson noise and homogeneous tuning curves $f_i(\theta)$ $(i = 1 \dots N)$. Assume that the coding is dense (i.e. each stimulus leads to the response of many neurons).

First show that the only stimulus dependent term of $\sum_i \log P(n_i)$ equals $\sum_i n_i \log f_i(\theta)$. Next, show that the Fisher info equals $I_f = T\sum_i \frac{f_i'^2}{f_i} - f_i''$

Finally, show that the Fisher info equals $I_f(\theta) = \frac{1}{\sigma^2}\sum_{i=1}^N \frac{[f_i''(\theta)]^2}{f_i(\theta)}$.

The number of spikes that a Poisson neuron at rate f_i fires in a time T is, $P = [f_i(\theta)T]^{n_i}\exp(-f_i T)/n_i!$.

To calculate FisherInfo:

$\sum_i \log P(n_i|\theta) = \sum_i [n_i \log f_i T - \log n_i! - f_i T]$. Now $\sum_i f_i \approx const$ for dense coding, so the only θ-dependent term is $\sum_i n_i \log f_i(\theta)$. Therefore

$\frac{\partial^2 \log P}{\partial\theta^2} = \sum_i n_i(\frac{f_i'}{f_i})' = \sum_i n_i[(\frac{f_i'}{f_i})^2 - \frac{f_i''}{f_i}]$ and $I_f = \sum_{n^{(1)},n^{(2)},\dots,n^{(N)}}^{\infty} p(n_i)n_i[(\frac{f_i'}{f_i})^2 - \frac{f_i''}{f_i}]$.

Note that $\sum_n p(n)n = Tf$ (the average of the firing rate is the tuning curve). So that

$$I_f(\theta) = T\sum_i \frac{f_i'(\theta)^2}{f_i(\theta)} - f_i''(\theta)$$

For dense tuning tuning curves we can replace the sum over units by an integral over centres of the tuning curves. This cancels out the second term. You can see this by for instance substitution of Gaussian shape tuning curves $f_i(\theta) = A\exp(-(\Delta i - s)^2/2\sigma^2)$.

2.3 Example Fisher information correlated neurons

Q: Consider two neurons, with response r_1 and r_2 with correlated Gaussian noise $r_i = f_i(\theta) + \eta_i$. Assume the correlation matrix of noise η equals $Q_{ij} = \sigma^2(\delta_{ij} + c(1-\delta_{ij}))$. The neurons have tuning curves $f_1(\theta)$ and $f_2(\theta) = \alpha f_1(\theta)$. Calculate Fisher info $I_F(\theta)$. You can use that for stimulus-independent noise $I_F = \boldsymbol{f}'(\theta)Q^{-1}\boldsymbol{f}'(\theta)$.

For correlated Gaussian noise (e.g. Abbott and Dayan, Neural comp 1999),

$$I = \boldsymbol{f}'(\theta)Q^{-1}(\theta)\boldsymbol{f}'(\theta) + \frac{1}{2}\mathrm{Tr}(Q'(\theta)Q^{-1}(\theta)Q'(\theta)Q^{-1}(\theta))$$

which in our case simplifies to $I = \boldsymbol{f}'(\theta)Q^{-1}\boldsymbol{f}'(\theta)$.

As $Q^{-1} = \frac{1}{\sigma^2(1-c^2)}\begin{pmatrix} 1 & -c \\ -c & 1 \end{pmatrix}$, so $I = \frac{1}{\sigma^2(1-c^2)}[|f_1'|^2 + |f_2'|^2 - 2cf_1'.f_2']$. Now assume that $f_2(\theta) = \alpha f_1(\theta)$ (so the neurons have identical tuning apart from a scale factor, that could be negative).

$$I = \frac{1}{\sigma^2(1-c^2)}|f_1'|^2[1 - 2c\alpha + \alpha^2]$$

If $\alpha = 1$, then $I = \frac{2}{1+c}\frac{|f_1'|^2}{\sigma^2}$. Note that if $\alpha = -1$, $I = \frac{2}{1-c}\frac{|f_1'|^2}{\sigma^2}$, so now correlation helps the information.

Note that if $\alpha = 0$ (one neuron does not code for the stimulus), the information diverges for $c \to \pm 1$. How would you build a perfect decoder in that limit?

2.4 Cramer-Rao bound with bias

Q: Given data $\mathbf{x}$, define estimator $T(\mathbf{x})$ of a quantity θ. If unbiased $\langle T \rangle = \theta$, but if the estimator is biased $\langle T \rangle = \theta + b(\theta)$, where $b(\theta)$ denotes the bias. Derive the Cramer-Rao bound for a biased estimator.

Define 'score' $V \equiv \partial_\theta \log P(x|\theta) = \frac{\partial_\theta P(x|\theta)}{P(x|\theta)}$. Note that for a 1D Gaussian $V \propto (x-\theta)/\sigma^2$. The Fisher info is defined as $I_f = \langle V^2 \rangle$.

According to Cauchy-Schwartz, $(\boldsymbol{x}.\boldsymbol{y})^2 \leq |x|^2|y|^2$. This is usually done for vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. Here we replace the sum over the dimensions in the inner-products with the sum over trials: $\langle (V - \langle V \rangle)(T - \langle T \rangle) \rangle^2 \leq \langle (V - \langle V \rangle)^2 \rangle \langle (T - \langle T \rangle)^2 \rangle$.

Note that $\langle V \rangle = \int dx p(x) \frac{\partial_\theta P(x|\theta)}{P(x|\theta)} = \partial_\theta \int dx P(x|\theta) = \partial_\theta 1 = 0$. So

$$\langle VT - V\langle T \rangle \rangle^2 \leq \langle V^2 \rangle \langle (T - \langle T \rangle)^2 \rangle$$
$$\langle VT \rangle^2 \leq \langle V^2 \rangle var(T)$$

Now $\langle VT \rangle = \int dx p(x) \frac{\partial_\theta P(x|\theta)}{P(x|\theta)} T(x) = \partial_\theta \int dx P(x|\theta)T(x) = \partial_\theta \langle T \rangle = 1 + b'(\theta)$. So we have $\langle (T-\theta)^2 \rangle \geq \frac{[1+b'(\theta)]^2}{I_f} - b^2(\theta)$ or

$$var(T) \geq \frac{[1+b'(\theta)]^2}{I_f}.$$

2.5 Convolution in Fourier domain

Q: Show that the Fourier transform of a convolution of two function equals the product of the Fourier transforms. The Fourier transform of a function is defined as $\widetilde{f(\omega)} = \int dt \exp(i\omega t) f(t)$.

Define Fourier Transform $\widetilde{f(\omega)} = \int dt \exp(i\omega t) f(t)$ (there are are various normalization conventions for the Fourier transform. In the end this should not matter). $(f \star g)(\tau) \equiv \int dt f(t) g(\tau - t)$. So $\widetilde{f \star g}(\omega) = \int d\tau \int dt f(t) g(\tau - t) \exp(i\omega\tau)$.

On the other hand

$$
\begin{aligned}
\tilde{f}(\omega).\tilde{g}(\omega) &= \int dt_1 dt_2 \exp(i\omega t_1) f(t_1) \exp(i\omega t_2) g(t_2) \\
&= \int dt \int d\tau \exp(i\omega t) \exp(-i\omega t + i\omega\tau) f(t) g(\tau - t) \\
&= \int dt \int d\tau \exp(i\omega\tau) f(t) g(\tau - t)
\end{aligned}
$$

(using $t_1 = t$,$t_2 = \tau - t$). So that indeed $\widetilde{f \star g}(\omega) = \tilde{f}(\omega).\tilde{g}(\omega)$

3 Information Theory

3.1 Derivation of mutual information

Q: Derive the expression for the mutual information $I_m = H(R) - H(R|S)$ in terms of the distribution $p(r, s)$.

Use that $p(r|s)p(s) = p(r, s)$, $p(r) = \sum_s p(r, s)$.

$$
\begin{aligned}
I_m &= H(R) - H(R|S) \\
&= -\sum_r p(r) \log p(r) + \sum_{r,s} p(r|s)p(s) \log p(r|s) \\
&= -\sum_{r,s} p(r, s) \log p(r) + \sum_{r,s} p(r, s)p(s)/p(s) \log[p(r, s)/p(s)] \\
&= \sum_{r,s} p(r, s) \log \frac{p(r, s)}{p(r)p(s)}
\end{aligned}
$$

3.2 Mutual information 2 correlated Gaussians

Q: Assume two correlated Gaussian variables y_1 and y_2. Show that mutual information $I(Y_1, Y_2) = -\frac{1}{2}\log(1 - \rho^2)$ with Pearson correlation coefficient $\rho = \frac{\sigma_{12}^2}{\sigma_1\sigma_2}$.

Introduce $\mathbf{y}$ =(y_1, y_2) and, if needed, translate the integrals so that they are centred around zero-mean.

$$
\begin{aligned}
I(Y_1, Y_2) &= \int P(y_1, y_2) \log \frac{P(y_1, y_2)}{P(y_1)P(y_2)} dy_1 dy_2 \\
&= \int N_{12} \exp(-\mathbf{y}^T C^{-1} \mathbf{y}/2) \log \left[\frac{N_{12}}{N_1 N_2} \exp(-\mathbf{y}^T C^{-1} \mathbf{y}/2 + y_1^2/2\sigma_1^2 + y_2^2/2\sigma_2^2) \right] \\
&= \int N_{12} \exp(-\mathbf{y}^T C^{-1} \mathbf{y}/2) \left[-\mathbf{y}^T C^{-1} \mathbf{y}/2 + y_1^2/\sigma_1^2 + y_2^2/\sigma_2^2 + \log \frac{N_{12}}{N_1 N_2} \right] \\
&= -1 + 1/2 + 1/2 + \log \frac{N_{12}}{N_1 N_2}
\end{aligned}
$$

where the normalisation factors are $1/N_{12} = 2\pi\sqrt{\det C}$, $1/N_1 = 2\pi\sigma_1$.

$\left(\frac{N_{12}}{N_1 N_2}\right)^2 = \frac{\sigma_1^2\sigma_2^2}{\det C} = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2\sigma_2^2 - \sigma_{12}^4} = [1 - \rho^2]^{-1}$, with Pearson correlation coefficient $\rho = \frac{\sigma_{12}^2}{\sigma_1\sigma_2}$. So that $I(Y_1, Y_2) = -\frac{1}{2}\log(1 - \rho^2)$. Note, if $\rho \to \pm 1$ then $I \to \infty$, and if $\rho \to 0$ then $I \to 0$.

3.3 Example of synergistic code

Q: Consider the following table of stimulus s and the response of two units r_1 and r_2 and their probability P.

s	r_1	r_2	$P(s, r_1, r_2)$
0	0	0	0
1	0	0	1/4
0	0	1	1/4
1	0	1	0
0	1	0	1/4
1	1	0	0
0	1	1	0
1	1	1	1/4

Show that this a synergistic code.

Note that $P(\boldsymbol{r}) = \prod_{i=1,2} P(r_i)$. $I_m(r_1, s) = I_m(r_2, s) = 0$, but $I_m((r_1, r_2), s) > 0$. In other words, observation of just r_1 or r_2 provides no information about the stimulus. However, observing both does.

3.4 Maximal entropy distributions

(Discrete distributions) Q: Which distribution $p(r_i)$ maximized the entropy $H(R) = -\sum_i p(r_i) \log_2 p(r_i)$? What if the mean is constrained? Assume that the bins are linearly spaced and start at $r = 0$. Thus $r_i = i\Delta r$. What if the variance is constrained?

We write $p_i = p(r_i)$.

- Only constrained is that p is a distribution, so that $\sum p_i = 1$. Maximize $E = H - \lambda(\sum p_i - 1)$, where λ is a Lagrange multiplier. $\frac{dE}{dp_j} = 0 \forall_j$ (using different indices i and j to prevent errors) so $\frac{dE}{dp_j} = 0 = \sum_i \frac{d}{dp_j}(-p_i \log p_i - \lambda p_i) = \sum_i \delta_{ij} - \log p_i - \frac{p_i}{p_i} - \lambda = 0$. So $\log p_i = -1 - \lambda \forall_i$ so $p_i = constant$ i.e. a uniform distribution
- Constrain mean to be $\bar{r}$. Now $E = H - \lambda(\sum p_i - 1) - \lambda_2(\sum p_i r_i - \bar{r})$. So $\log p_i = -1 - \lambda - \lambda_2 r_i$, so $p_i \propto \exp(-\lambda_2 r_i)$. Which after normalization and setting the mean gives $p_i = 1/\bar{r} \exp(-r_i/\bar{r})$.
- Constrain variance, same as above but now add term $E = H - \lambda(\sum p_i - 1) - \lambda_2(\sum p_i r_i^2 - \sigma_r^2)$ (assuming zero mean). Now $p_i \propto \exp(-c r_i^2)$.

3.5 Gaussian variable with noise

Q: Consider a noise response $r = s + \eta$, where both signal s and noise η are Gaussian distributed. Calculate the mutual information between r and s.

What is the MAP estimate of s given r?

$H_{noise} = \frac{1}{2}\log_2(2\pi e \sigma_\eta^2)$, $H_r = \frac{1}{2}\log_2(2\pi e(\sigma_{eta}^2 + \sigma_s^2))$. Hence $I = H - H_{noise} = \frac{1}{2}\log_2(\frac{\sigma_\eta^2 + \sigma_s^2}{\sigma_\eta^2}) = \log_2 \sqrt{1 + \sigma_s^2/\sigma_\eta^2} = \log_2 \sqrt{1 + SNR}$, with $SNR \equiv \sigma_s^2/\sigma_\eta^2$.

Note if $r = r(f)$, $s = s(f)$ $I = \int df \log_2(1 + SNR(f))$.

MAP estimate of s

$$
\begin{aligned}
P(s|r) &= P(s)P(r|s)/P(r) \\
&= \frac{1}{P(r)} N \exp[-s^2/2\sigma_s^2] \exp[-(r-s)^2/2\sigma_\eta^2] \\
&= \frac{1}{P(r)} N \exp[-\frac{1}{2}s^2(\sigma_s^{-2} + \sigma_\eta^{-2}) + sr\sigma_\eta^{-2}]
\end{aligned}
$$

Maximal $d/ds = 0$, if $s = r\frac{\sigma_s^2}{\sigma_s^2 + \sigma_\eta^2}$. Interpretation: because r can be large due to the noise, r "conservatively" estimates s.