Information Theory

Maneesh Sahani

Gatsby Computational Neuroscience Unit
University College London

February 2019

Quantifying a Code

- How much information does a neural response carry about a stimulus?
- How efficient is a hypothetical code, given the statistical behaviour of the components?
- How much better could another code do, given the same components?
- Is the information carried by different neurons complementary, synergistic (whole is greater than sum of parts), or redundant?
- Can further processing extract more information about a stimulus?

Information theory is the mathematical framework within which questions such as these can be framed and answered.

Quantifying a Code

- How much information does a neural response carry about a stimulus?
- How efficient is a hypothetical code, given the statistical behaviour of the components?
- How much better could another code do, given the same components?
- Is the information carried by different neurons complementary, synergistic (whole is greater than sum of parts), or redundant?
- Can further processing extract more information about a stimulus?

Information theory is the mathematical framework within which questions such as these can be framed and answered.

Information theory does not directly address:

- estimation (but there are some relevant bounds)
- computation (but “information bottleneck” might provide a motivating framework)
- representation (but redundancy reduction has obvious information theoretic connections)

Uncertainty and Information

Information is related to the removal of uncertainty.

Uncertainty and Information

Information is related to the removal of uncertainty.

$$S \rightarrow R \rightarrow P(S|R)$$

How informative is R about S?

Uncertainty and Information

Information is related to the removal of uncertainty.

$$S \rightarrow R \rightarrow P(S|R)$$

How informative is R about S?

$$P(S|R) = \left[0, 0, 1, 0, \ldots, 0\right] \qquad \Rightarrow \text{high information?}$$

$$P(S|R) = \left[\frac{1}{M}, \frac{1}{M}, \ldots, \frac{1}{M}\right] \qquad \Rightarrow \text{low information?}$$

Uncertainty and Information

Information is related to the removal of uncertainty.

$$S \to R \to P(S|R)$$

How informative is R about S?

$$P(S|R) = \left[0, 0, 1, 0, \ldots, 0\right] \qquad \Rightarrow \text{high information?}$$

$$P(S|R) = \left[\frac{1}{M}, \frac{1}{M}, \ldots, \frac{1}{M}\right] \qquad \Rightarrow \text{low information?}$$

But also depends on $P(S)$.

We need to start by considering the uncertainty in a probability distribution $\to$ called the **entropy**

Uncertainty and Information

Information is related to the removal of uncertainty.

$$S \rightarrow R \rightarrow P(S|R)$$

How informative is R about S?

$$P(S|R) = \left[0, 0, 1, 0, \ldots, 0\right] \qquad \Rightarrow \text{high information?}$$

$$P(S|R) = \left[\frac{1}{M}, \frac{1}{M}, \ldots, \frac{1}{M}\right] \qquad \Rightarrow \text{low information?}$$

But also depends on $P(S)$.

We need to start by considering the uncertainty in a probability distribution $\rightarrow$ called the **entropy**

Let $S \sim P(S)$. The entropy is the minimum number of bits needed, on average, to specify the value S takes, assuming $P(S)$ is known.

Equivalently, the minimum average number of yes/no questions needed to guess S.

Entropy

Entropy

- Suppose there are M equiprobable stimuli: $P(s_m) = 1/M$.

  To specify which stimulus appears on a given trial, we would need assign each a (binary) number. This would take,

$$
\begin{aligned}
B_s \leq & \quad \log_2 M + 1 \qquad [2^B \geq M] \\
= & -\log_2 \frac{1}{M} + 1 \text{ bits}
\end{aligned}
$$

Entropy

- Suppose there are M equiprobable stimuli: $P(s_m) = 1/M$.

  To specify which stimulus appears on a given trial, we would need assign each a (binary) number. This would take,

$$\begin{aligned} B_s \leq & \quad \log_2 M + 1 \qquad [2^B \geq M] \\ & = -\log_2 \frac{1}{M} + 1 \text{ bits} \end{aligned}$$

- Now suppose we code N such stimuli, drawn iid, at once.

$$\begin{aligned} B_N \leq & \quad \log_2 M^N + 1 \\ & \to -N \log_2 \frac{1}{M} \qquad \text{as } N \to \infty \\ \Rightarrow B_s & \to -\log_2 p \text{ bits} \end{aligned}$$

  This is called block coding. It is useful for extracting theoretical limits. The nervous system is unlikely to use block codes in time, but may in space.

Entropy

- Now suppose stimuli are not equiprobable. Write $P(s_m) = p_m$. Then

$$P(S_1, S_2, \ldots, S_N) = \prod_m p_m^{n_m} \qquad \text{[where } n_m = (\text{\# of } S_i = s_m)\text{].}$$

Entropy

- Now suppose stimuli are not equiprobable. Write $P(s_m) = p_m$. Then

$$P(S_1, S_2, \ldots, S_N) = \prod_m p_m^{n_m} \qquad \text{[where } n_m = (\text{\# of } S_i = s_m)\text{]}.$$

  As $N \to \infty$ only “typical” sequences, with $n_m = p_m N$, have non-zero probability of occuring; and they are all equally likely. This is called the Asymptotic Equipartition Property (or AEP).

Entropy

- Now suppose stimuli are not equiprobable. Write $P(s_m) = p_m$. Then

$$P(S_1, S_2, \ldots, S_N) = \prod_m p_m^{n_m} \qquad\qquad [\text{where } n_m = (\#\text{ of } S_i = s_m)].$$

As $N \to \infty$ only "typical" sequences, with $n_m = p_m N$, have non-zero probability of occuring; and they are all equally likely. This is called the Asymptotic Equipartition Property (or AEP).

Thus,

$$\begin{aligned} B_N &\to -\log_2 \textstyle\prod_m p_m^{n_m} &&= -\textstyle\sum_m n_m \log_2 p_m \\ &= -\textstyle\sum_m p_m N \log_2 p_m &&= -N \underbrace{\sum_m p_m \log_2 p_m}_{-\mathbf{H}[s]} \end{aligned}$$

Entropy

- Now suppose stimuli are not equiprobable. Write $P(s_m) = p_m$. Then

$$P(S_1, S_2, \ldots, S_N) = \prod_m p_m^{n_m} \qquad \text{[where } n_m = (\#\text{ of } S_i = s_m)\text{]}.$$

As $N \to \infty$ only "typical" sequences, with $n_m = p_m N$, have non-zero probability of occuring; and they are all equally likely. This is called the Asymptotic Equipartition Property (or AEP).

Thus,

$$\begin{aligned} B_N &\to -\log_2 \textstyle\prod_m p_m^{n_m} &&= -\textstyle\sum_m n_m \log_2 p_m \\ &= -\textstyle\sum_m p_m N \log_2 p_m &&= -N \underbrace{\sum_m p_m \log_2 p_m}_{-\mathbf{H}[s]} \end{aligned}$$

$\mathbf{H}[S] = \mathbb{E}\left[-\log_2 P(S)\right]$, also written $\mathbf{H}[P(S)]$, is the **entropy** of the stimulus distribution.

Entropy

- Now suppose stimuli are not equiprobable. Write $P(s_m) = p_m$. Then

$$P(S_1, S_2, \ldots, S_N) = \prod_m p_m^{n_m} \qquad [\text{where } n_m = (\# \text{ of } S_i = s_m)].$$

As $N \to \infty$ only "typical" sequences, with $n_m = p_m N$, have non-zero probability of occuring; and they are all equally likely. This is called the Asymptotic Equipartition Property (or AEP).

Thus,

$$\begin{aligned} B_N &\to -\log_2 \textstyle\prod_m p_m^{n_m} &&= -\textstyle\sum_m n_m \log_2 p_m \\ &= -\textstyle\sum_m p_m N \log_2 p_m &&= -N \underbrace{\sum_m p_m \log_2 p_m}_{-\mathbf{H}[s]} \end{aligned}$$

$\mathbf{H}[S] = \mathbb{E}\left[-\log_2 P(S)\right]$, also written $\mathbf{H}[P(S)]$, is the **entropy** of the stimulus distribution.

Rather than appealing to typicality, we could instead have used the law of large numbers directly:

$$\frac{1}{N} \log_2 P(S_1, S_2, \ldots S_N) = \frac{1}{N} \log_2 \prod_i P(S_i) = \frac{1}{N} \sum_i \log_2 P(S_i) \overset{N \to \infty}{\to} \mathrm{E}[\log_2 P(S_i)]$$

Conditional Entropy

Entropy is a measure of “available information” in the stimulus ensemble.

Conditional Entropy

Entropy is a measure of “available information” in the stimulus ensemble. Now suppose we measure a particular response r which depends on the stimulus according to $P(R|S)$.

How uncertain is the stimulus once we know r?

Conditional Entropy

Entropy is a measure of “available information” in the stimulus ensemble. Now suppose we measure a particular response r which depends on the stimulus according to $P(R|S)$.

How uncertain is the stimulus once we know r? Bayes rule gives us

$$P(S|r) = \frac{P(r|S)P(S)}{\sum_s P(r|s)P(s)}$$

so we can write

$$\mathbf{H}[S|r] = -\sum_s P(s|r) \log_2 P(s|r)$$

Conditional Entropy

Entropy is a measure of “available information” in the stimulus ensemble. Now suppose we measure a particular response r which depends on the stimulus according to $P(R|S)$.

How uncertain is the stimulus once we know r? Bayes rule gives us

$$P(S|r) = \frac{P(r|S)P(S)}{\sum_s P(r|s)P(s)}$$

so we can write

$$\mathbf{H}[S|r] = -\sum_s P(s|r)\log_2 P(s|r)$$

The *average* uncertainty in S for $r \sim P(R) = \sum_s P(R|s)p(s)$ is then

$$\mathbf{H}[S|R] = \sum_r P(r)\left[-\sum_s P(s|r)\log_2 P(s|r)\right] = -\sum_{s,r} P(s,r)\log_2 P(s|r)$$

Conditional Entropy

Entropy is a measure of "available information" in the stimulus ensemble. Now suppose we measure a particular response r which depends on the stimulus according to $P(R|S)$.

How uncertain is the stimulus once we know r? Bayes rule gives us

$$P(S|r) = \frac{P(r|S)P(S)}{\sum_s P(r|s)P(s)}$$

so we can write

$$\mathbf{H}[S|r] = -\sum_s P(s|r) \log_2 P(s|r)$$

The *average* uncertainty in S for $r \sim P(R) = \sum_s P(R|s)p(s)$ is then

$$\mathbf{H}[S|R] = \sum_r P(r) \left[-\sum_s P(s|r) \log_2 P(s|r) \right] = -\sum_{s,r} P(s,r) \log_2 P(s|r)$$

It is easy to show that:

1. $\mathbf{H}[S|R] \leq \mathbf{H}[S]$
2. $\mathbf{H}[S|R] = \mathbf{H}[S,R] - \mathbf{H}[R]$
3. $\mathbf{H}[S|R] = \mathbf{H}[S]$ iff $S \perp\!\!\!\perp R$

Average Mutual Information

A natural definition of the average information gained about S from R is

$$\mathbf{I}[S;R] = \mathbf{H}[S] - \mathbf{H}[S|R]$$

Measures *reduction in uncertainty* due to R.

Average Mutual Information

A natural definition of the average information gained about S from R is

$$\mathbf{I}[S;R] = \mathbf{H}[S] - \mathbf{H}[S|R]$$

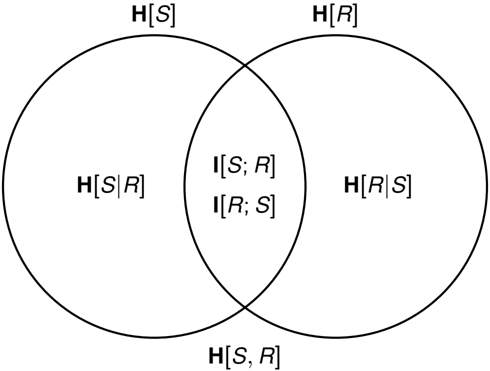
Measures *reduction in uncertainty* due to R.

It follows from the definition that

$$\begin{aligned}
\mathbf{I}[S;R] &= \sum_{s} P(s)\log\frac{1}{P(s)} - \sum_{s,r} P(s,r)\log\frac{1}{P(s|r)} \\
&= \sum_{s,r} P(s,r)\log\frac{1}{P(s)} + \sum_{s,r} P(s,r)\log P(s|r) \\
&= \sum_{s,r} P(s,r)\log\frac{P(s|r)}{P(s)} \\
&= \sum_{s,r} P(s,r)\log\frac{P(s,r)}{P(s)P(r)} \\
&= \mathbf{I}[R;S]
\end{aligned}$$

Average Mutual Information

The symmetry suggests a Venn-like diagram.



All of the additive and equality relationships implied by this picture hold for two variables. Unfortunately, we will see that this does not generalise to any more than two.

Kullback-Leibler Divergence

Another useful information theoretic quantity measures the difference between two distributions.

$$
\begin{aligned}
\mathbf{KL}[P(S) \| Q(S)] &= \sum_s P(s) \log \frac{P(s)}{Q(s)} \\
&= \underbrace{\sum_s P(s) \log \frac{1}{Q(s)}}_{\text{cross entropy}} - \mathbf{H}[P]
\end{aligned}
$$

Excess cost in bits paid by encoding according to Q instead of P.

Kullback-Leibler Divergence

Another useful information theoretic quantity measures the difference between two distributions.

$$\begin{aligned}\mathbf{KL}[P(S)\|Q(S)] &= \sum_s P(s)\log\frac{P(s)}{Q(s)}\\ &= \underbrace{\sum_s P(s)\log\frac{1}{Q(s)}}_{\text{cross entropy}} - \mathbf{H}[P]\end{aligned}$$

Excess cost in bits paid by encoding according to Q instead of P.

$$\begin{aligned}-\mathbf{KL}[P\|Q] &= \sum_s P(s)\log\frac{Q(s)}{P(s)}\\ &\leq \log\sum_s P(s)\frac{Q(s)}{P(s)} \qquad \text{by Jensen}\\ &= \log\sum_s Q(s) = \log 1 = 0\end{aligned}$$

So $\mathbf{KL}[P\|Q] \geq 0$. Equality iff $P = Q$

Mutual Information and KL

$$\mathbf{I}[S;R] = \sum_{s,r} P(s,r) \log \frac{P(s,r)}{P(s)P(r)} = \mathbf{KL}[P(S,R) \| P(S)P(R)]$$

Mutual Information and KL

$$\mathbf{I}[S;R] = \sum_{s,r} P(s,r) \log \frac{P(s,r)}{P(s)P(r)} = \mathbf{KL}[P(S,R) \| P(S)P(R)]$$

Thus:

1. Mutual information is always non-negative

$$\mathbf{I}[S;R] \geq 0$$

Mutual Information and KL

$$\mathbf{I}[S;R] = \sum_{s,r} P(s,r) \log \frac{P(s,r)}{P(s)P(r)} = \mathbf{KL}[P(S,R) \| P(S)P(R)]$$

Thus:

1. Mutual information is always non-negative

$$\mathbf{I}[S;R] \geq 0$$

2. Conditioning never increases entropy

$$\mathbf{H}[S|R] \leq \mathbf{H}[S]$$

Multiple Responses

Two responses to the same stimulus, R_1 and R_2, may provide either more or less information jointly than independently.

Multiple Responses

Two responses to the same stimulus, R_1 and R_2, may provide either more or less information jointly than independently.

$$I_{12} = \mathbf{I}[S; R_1, R_2] = \mathbf{H}[R_1, R_2] - \mathbf{H}[R_1, R_2|S]$$

$$R_1 \perp\!\!\!\perp R_2 \Rightarrow \mathbf{H}[R_1, R_2] = \mathbf{H}[R_1] + \mathbf{H}[R_2]$$
$$R_1 \perp\!\!\!\perp R_2|S \Rightarrow \mathbf{H}[R_1, R_2|S] = \mathbf{H}[R_1|S] + \mathbf{H}[R_2|S]$$

Multiple Responses

Two responses to the same stimulus, R_1 and R_2, may provide either more or less information jointly than independently.

$$I_{12} = \mathbf{I}[S; R_1, R_2] = \mathbf{H}[R_1, R_2] - \mathbf{H}[R_1, R_2|S]$$

$$R_1 \perp\!\!\!\perp R_2 \Rightarrow \mathbf{H}[R_1, R_2] = \mathbf{H}[R_1] + \mathbf{H}[R_2]$$
$$R_1 \perp\!\!\!\perp R_2|S \Rightarrow \mathbf{H}[R_1, R_2|S] = \mathbf{H}[R_1|S] + \mathbf{H}[R_2|S]$$

$R_1 \perp\!\!\!\perp R_2$	$R_1 \perp\!\!\!\perp R_2 \vert S$		
no	yes	$I_{12} < I_1 + I_2$	redundant

Multiple Responses

Two responses to the same stimulus, R_1 and R_2, may provide either more or less information jointly than independently.

$$I_{12} = \mathbf{I}[S; R_1, R_2] = \mathbf{H}[R_1, R_2] - \mathbf{H}[R_1, R_2|S]$$

$$R_1 \perp\!\!\!\perp R_2 \Rightarrow \mathbf{H}[R_1, R_2] = \mathbf{H}[R_1] + \mathbf{H}[R_2]$$
$$R_1 \perp\!\!\!\perp R_2|S \Rightarrow \mathbf{H}[R_1, R_2|S] = \mathbf{H}[R_1|S] + \mathbf{H}[R_2|S]$$

$R_1 \perp\!\!\!\perp R_2$	$R_1 \perp\!\!\!\perp R_2 \mid S$		
no	yes	$I_{12} < I_1 + I_2$	redundant
yes	yes	$I_{12} = I_1 + I_2$	independent

Multiple Responses

Two responses to the same stimulus, R_1 and R_2, may provide either more or less information jointly than independently.

$$I_{12} = \mathbf{I}[S; R_1, R_2] = \mathbf{H}[R_1, R_2] - \mathbf{H}[R_1, R_2|S]$$

$$R_1 \perp\!\!\!\perp R_2 \Rightarrow \mathbf{H}[R_1, R_2] = \mathbf{H}[R_1] + \mathbf{H}[R_2]$$
$$R_1 \perp\!\!\!\perp R_2|S \Rightarrow \mathbf{H}[R_1, R_2|S] = \mathbf{H}[R_1|S] + \mathbf{H}[R_2|S]$$

$R_1 \perp\!\!\!\perp R_2$	$R_1 \perp\!\!\!\perp R_2\|S$		
no	yes	$I_{12} < I_1 + I_2$	redundant
yes	yes	$I_{12} = I_1 + I_2$	independent
yes	no	$I_{12} > I_1 + I_2$	synergistic

Multiple Responses

Two responses to the same stimulus, R_1 and R_2, may provide either more or less information jointly than independently.

$$I_{12} = \mathbf{I}[S; R_1, R_2] = \mathbf{H}[R_1, R_2] - \mathbf{H}[R_1, R_2|S]$$

$$R_1 \perp\!\!\!\perp R_2 \Rightarrow \mathbf{H}[R_1, R_2] = \mathbf{H}[R_1] + \mathbf{H}[R_2]$$
$$R_1 \perp\!\!\!\perp R_2|S \Rightarrow \mathbf{H}[R_1, R_2|S] = \mathbf{H}[R_1|S] + \mathbf{H}[R_2|S]$$

$R_1 \perp\!\!\!\perp R_2$	$R_1 \perp\!\!\!\perp R_2 \vert S$		
no	yes	$I_{12} < I_1 + I_2$	redundant
yes	yes	$I_{12} = I_1 + I_2$	independent
yes	no	$I_{12} > I_1 + I_2$	synergistic
no	no	?	any of the above

Multiple Responses

Two responses to the same stimulus, R_1 and R_2, may provide either more or less information jointly than independently.

$$I_{12} = \mathbf{I}[S; R_1, R_2] = \mathbf{H}[R_1, R_2] - \mathbf{H}[R_1, R_2|S]$$

$$R_1 \perp\!\!\!\perp R_2 \Rightarrow \mathbf{H}[R_1, R_2] = \mathbf{H}[R_1] + \mathbf{H}[R_2]$$
$$R_1 \perp\!\!\!\perp R_2|S \Rightarrow \mathbf{H}[R_1, R_2|S] = \mathbf{H}[R_1|S] + \mathbf{H}[R_2|S]$$

$R_1 \perp\!\!\!\perp R_2$	$R_1 \perp\!\!\!\perp R_2 \vert S$		
no	yes	$I_{12} < I_1 + I_2$	redundant
yes	yes	$I_{12} = I_1 + I_2$	independent
yes	no	$I_{12} > I_1 + I_2$	synergistic
no	no	?	any of the above

$I_{12} > \max(I_1, I_2)$: the second response cannot destroy information.

Multiple Responses

Two responses to the same stimulus, R_1 and R_2, may provide either more or less information jointly than independently.

$$I_{12} = \mathbf{I}[S; R_1, R_2] = \mathbf{H}[R_1, R_2] - \mathbf{H}[R_1, R_2|S]$$

$$R_1 \perp\!\!\!\perp R_2 \Rightarrow \mathbf{H}[R_1, R_2] = \mathbf{H}[R_1] + \mathbf{H}[R_2]$$
$$R_1 \perp\!\!\!\perp R_2|S \Rightarrow \mathbf{H}[R_1, R_2|S] = \mathbf{H}[R_1|S] + \mathbf{H}[R_2|S]$$

$R_1 \perp\!\!\!\perp R_2$	$R_1 \perp\!\!\!\perp R_2 \| S$		
no	yes	$I_{12} < I_1 + I_2$	redundant
yes	yes	$I_{12} = I_1 + I_2$	independent
yes	no	$I_{12} > I_1 + I_2$	synergistic
no	no	?	any of the above

$I_{12} > \max(I_1, I_2)$: the second response cannot destroy information.

Thus, the Venn-like diagram with three variables is misleading.

Data Processing Inequality

Data Processing Inequality

Suppose $S \to R_1 \to R_2$ form a Markov chain; that is, $R_2 \perp\!\!\!\perp S \mid R_1$.

Then,

$$\begin{aligned} &P(R_2, S|R_1) = P(R_2|R_1)P(S|R_1) \\ \Rightarrow\ &P(S|R_1, R_2) = P(S|R_1) \end{aligned}$$

Data Processing Inequality

Suppose $S \rightarrow R_1 \rightarrow R_2$ form a Markov chain; that is, $R_2 \perp\!\!\!\perp S \mid R_1$.

Then,

$$\begin{aligned} &P(R_2, S|R_1) = P(R_2|R_1)P(S|R_1) \\ \Rightarrow\ &P(S|R_1, R_2) = P(S|R_1) \end{aligned}$$

Thus,

$$\mathbf{H}[S|R_2] \geq \mathbf{H}[S|R_1, R_2] = \mathbf{H}[S|R_1]$$

$$\Rightarrow \mathbf{I}[S; R_2] \leq \mathbf{I}[S; R_1]$$

So any computation based on R_1 that does not have separate access to S cannot add information (in the Shannon sense) about the world.

Data Processing Inequality

Suppose $S \to R_1 \to R_2$ form a Markov chain; that is, $R_2 \perp\!\!\!\perp S \mid R_1$.

Then,

$$\begin{aligned} & P(R_2, S|R_1) = P(R_2|R_1)P(S|R_1) \\ \Rightarrow\ & P(S|R_1, R_2) = P(S|R_1) \end{aligned}$$

Thus,

$$\mathbf{H}[S|R_2] \geq \mathbf{H}[S|R_1, R_2] = \mathbf{H}[S|R_1]$$

$$\Rightarrow \mathbf{I}[S; R_2] \leq \mathbf{I}[S; R_1]$$

So any computation based on R_1 that does not have separate access to S cannot add information (in the Shannon sense) about the world.

Equality holds iff $S \to R_2 \to R_1$ as well. In this case R_2 is called a **sufficient statistic** for S.

Entropy Rate

So far we have discussed S and R as single (or iid) random variables. But real stimuli and responses form a time series.

Entropy Rate

So far we have discussed S and R as single (or iid) random variables. But real stimuli and responses form a time series.

Let $\mathcal{S} = \{S_1, S_2, S_3 \ldots\}$ form a stochastic process.

$$
\begin{aligned}
\mathbf{H}[S_1, S_2, \ldots, S_n] &= \mathbf{H}[S_n | S_1, S_2, \ldots, S_{n-1}] + \mathbf{H}[S_1, S_2, \ldots, S_{n-1}] \\
&= \mathbf{H}[S_n | S_1, S_2, \ldots, S_{n-1}] + \mathbf{H}[S_{n-1} | S_1, S_2, \ldots, S_{n-2}] + \ldots + \mathbf{H}[S_1]
\end{aligned}
$$

Entropy Rate

So far we have discussed S and R as single (or iid) random variables. But real stimuli and responses form a time series.

Let $\mathcal{S} = \{S_1, S_2, S_3 \ldots\}$ form a stochastic process.

$$\begin{aligned}\mathbf{H}[S_1, S_2, \ldots, S_n] &= \mathbf{H}[S_n | S_1, S_2, \ldots, S_{n-1}] + \mathbf{H}[S_1, S_2, \ldots, S_{n-1}] \\ &= \mathbf{H}[S_n | S_1, S_2, \ldots, S_{n-1}] + \mathbf{H}[S_{n-1} | S_1, S_2, \ldots, S_{n-2}] + \ldots + \mathbf{H}[S_1]\end{aligned}$$

The **entropy rate** of $\mathcal{S}$ is defined as

$$\mathbf{H}[\mathcal{S}] = \lim_{n \to \infty} \frac{\mathbf{H}[S_1, S_2, \ldots, S_n]}{N}$$

or alternatively as

$$\mathbf{H}[\mathcal{S}] = \lim_{n \to \infty} \mathbf{H}[S_n | S_1, S_2, \ldots, S_{n-1}]$$

Entropy Rate

So far we have discussed S and R as single (or iid) random variables. But real stimuli and responses form a time series.

Let $\mathcal{S} = \{S_1, S_2, S_3 \ldots\}$ form a stochastic process.

$$\begin{aligned}\mathbf{H}[S_1, S_2, \ldots, S_n] &= \mathbf{H}[S_n|S_1, S_2, \ldots, S_{n-1}] + \mathbf{H}[S_1, S_2, \ldots, S_{n-1}] \\ &= \mathbf{H}[S_n|S_1, S_2, \ldots, S_{n-1}] + \mathbf{H}[S_{n-1}|S_1, S_2, \ldots, S_{n-2}] + \ldots + \mathbf{H}[S_1]\end{aligned}$$

The **entropy rate** of $\mathcal{S}$ is defined as

$$\mathbf{H}[\mathcal{S}] = \lim_{n\to\infty} \frac{\mathbf{H}[S_1, S_2, \ldots, S_n]}{N}$$

or alternatively as

$$\mathbf{H}[\mathcal{S}] = \lim_{n\to\infty} \mathbf{H}[S_n|S_1, S_2, \ldots, S_{n-1}]$$

If $S_i \overset{\text{iid}}{\sim} P(S)$ then $\mathbf{H}[\mathcal{S}] = \mathbf{H}[S]$.
If $\mathcal{S}$ is Markov (and stationary) then $\mathbf{H}[\mathcal{S}] = \mathbf{H}[S_n|S_{n-1}]$.

Continuous Random Variables

The discussion so far has involved discrete S and R. Now, let $S \in \mathbb{R}$ with density p(s). What is its entropy?

Continuous Random Variables

The discussion so far has involved discrete S and R. Now, let $S \in \mathbb{R}$ with density p(s). What is its entropy?

Suppose we discretise with length Δs:

$$\begin{aligned} \mathrm{H}_{\Delta}[S] &= -\sum_i p(s_i)\Delta s \log p(s_i)\Delta s \\ &= -\sum_i p(s_i)\Delta s(\log p(s_i) + \log \Delta s) \end{aligned}$$

Continuous Random Variables

The discussion so far has involved discrete S and R. Now, let $S \in \mathbb{R}$ with density p(s). What is its entropy?

Suppose we discretise with length Δs:

$$\begin{aligned}
\mathrm{H}_\Delta[S] &= -\sum_i p(s_i)\Delta s \log p(s_i)\Delta s \\
&= -\sum_i p(s_i)\Delta s(\log p(s_i) + \log \Delta s) \\
&= -\sum_i p(s_i)\Delta s \log p(s_i) - \log \Delta s \sum_i p(s_i)\Delta s
\end{aligned}$$

Continuous Random Variables

The discussion so far has involved discrete S and R. Now, let $S \in \mathbb{R}$ with density p(s). What is its entropy?

Suppose we discretise with length Δs:

$$\begin{aligned}
\mathrm{H}_\Delta[S] &= -\sum_i p(s_i)\Delta s \log p(s_i)\Delta s \\
&= -\sum_i p(s_i)\Delta s(\log p(s_i) + \log \Delta s) \\
&= -\sum_i p(s_i)\Delta s \log p(s_i) - \log \Delta s \sum_i p(s_i)\Delta s \\
&= -\sum_i \Delta s\, p(s_i) \log p(s_i) - \log \Delta s
\end{aligned}$$

Continuous Random Variables

The discussion so far has involved discrete S and R. Now, let $S \in \mathbb{R}$ with density p(s). What is its entropy?

Suppose we discretise with length Δs:

$$
\begin{aligned}
\mathrm{H}_\Delta[S] &= -\sum_i p(s_i)\Delta s \log p(s_i)\Delta s \\
&= -\sum_i p(s_i)\Delta s(\log p(s_i) + \log \Delta s) \\
&= -\sum_i p(s_i)\Delta s \log p(s_i) - \log \Delta s \sum_i p(s_i)\Delta s \\
&= -\sum_i \Delta s\, p(s_i) \log p(s_i) - \log \Delta s \\
&\to -\int ds\, p(s) \log p(s) + \infty
\end{aligned}
$$

Continuous Random Variables

The discussion so far has involved discrete S and R. Now, let $S \in \mathbb{R}$ with density p(s). What is its entropy?

Suppose we discretise with length Δs:

$$
\begin{aligned}
\mathrm{H}_\Delta[S] &= -\sum_i p(s_i)\Delta s \log p(s_i)\Delta s \\
&= -\sum_i p(s_i)\Delta s(\log p(s_i) + \log \Delta s) \\
&= -\sum_i p(s_i)\Delta s \log p(s_i) - \log \Delta s \sum_i p(s_i)\Delta s \\
&= -\sum_i \Delta s\, p(s_i) \log p(s_i) - \log \Delta s \\
&\to -\int ds\, p(s) \log p(s) + \infty
\end{aligned}
$$

We define the **differential entropy**:

$$h(S) = -\int ds\, p(s) \log p(s).$$

Note that $h(S)$ can be < 0, and can be $\pm\infty$.

Continuous Random Variables

We can define other information theoretic quantities similarly.

Continuous Random Variables

We can define other information theoretic quantities similarly.

The conditional differential entropy is

$$h(S|R) = -\int ds\, dr\; p(s,r) \log p(s|r)$$

and, like the differential entropy itself, may be poorly behaved.

Continuous Random Variables

We can define other information theoretic quantities similarly.

The conditional differential entropy is

$$h(S|R) = -\int ds\, dr\ p(s,r) \log p(s|r)$$

and, like the differential entropy itself, may be poorly behaved.

The mutual information, however, is well-defined

$$\begin{aligned}
\mathrm{I}_\Delta[S;R] &= \mathrm{H}_\Delta[S] - \mathrm{H}_\Delta[S|R] \\
&= -\sum_i \Delta s\ p(s_i) \log p(s_i) - \log \Delta s \\
&\qquad - \int dr\ p(r) \left(-\sum_i \Delta s\ p(s_i|r) \log p(s_i|r) - \log \Delta s \right) \\
&\to h(S) - h(S|R)
\end{aligned}$$

as are other KL divergences.

Maximum Entropy Distributions

1. $\mathbf{H}[R_1, R_2] = \mathbf{H}[R_1] + \mathbf{H}[R_2]$ with equality iff $R_1 \perp\!\!\!\perp R_2$.

Maximum Entropy Distributions

1. $\mathbf{H}[R_1, R_2] = \mathbf{H}[R_1] + \mathbf{H}[R_2]$ with equality iff $R_1 \perp\!\!\!\perp R_2$.
2. Let $\int ds\, p(s) f(s) = a$ for some function f. What distribution has maximum entropy?

Maximum Entropy Distributions

1. $\mathbf{H}[R_1, R_2] = \mathbf{H}[R_1] + \mathbf{H}[R_2]$ with equality iff $R_1 \perp\!\!\!\perp R_2$.
2. Let $\int ds\, p(s)f(s) = a$ for some function f. What distribution has maximum entropy? Use Lagrange multipliers:

$$\mathcal{L} = \int ds\, p(s) \log p(s) - \lambda_0 \left[\int ds\, p(s) - 1\right] - \lambda_1 \left[\int ds\, p(s)f(s) - a\right]$$

$$\frac{\delta \mathcal{L}}{\delta p(s)} = 1 + \log p(s) - \lambda_0 - \lambda_1 f(s) = 0$$

$$\Rightarrow \log p(s) = \lambda_0 + \lambda_1 f(s) - 1$$

$$\Rightarrow p(s) = \frac{1}{Z} e^{\lambda_1 f(s)}$$

The constants λ_0 and λ_1 can be found by solving the constraint equations.

Maximum Entropy Distributions

1. $\mathbf{H}[R_1, R_2] = \mathbf{H}[R_1] + \mathbf{H}[R_2]$ with equality iff $R_1 \perp\!\!\!\perp R_2$.
2. Let $\int ds\; p(s) f(s) = a$ for some function f. What distribution has maximum entropy?
   Use Lagrange multipliers:

$$\mathcal{L} = \int ds\; p(s) \log p(s) - \lambda_0 \left[\int ds\; p(s) - 1 \right] - \lambda_1 \left[\int ds\; p(s) f(s) - a \right]$$

$$\frac{\delta \mathcal{L}}{\delta p(s)} = 1 + \log p(s) - \lambda_0 - \lambda_1 f(s) = 0$$

$$\Rightarrow \log p(s) = \lambda_0 + \lambda_1 f(s) - 1$$

$$\Rightarrow p(s) = \frac{1}{Z} e^{\lambda_1 f(s)}$$

   The constants λ_0 and λ_1 can be found by solving the constraint equations.
   Thus,

$$f(s) = s \quad \Rightarrow \quad p(s) = \tfrac{1}{Z} e^{\lambda_1 s}. \quad \text{Exponential (need } p(s) = 0 \text{ for } s < T\text{).}$$

$$f(s) = s^2 \quad \Rightarrow \quad p(s) = \tfrac{1}{Z} e^{\lambda_1 s^2}. \quad \text{Gaussian.}$$

Maximum Entropy Distributions

1. $\mathbf{H}[R_1, R_2] = \mathbf{H}[R_1] + \mathbf{H}[R_2]$ with equality iff $R_1 \perp\!\!\!\perp R_2$.
2. Let $\int ds\, p(s) f(s) = a$ for some function f. What distribution has maximum entropy? Use Lagrange multipliers:

$$\mathcal{L} = \int ds\, p(s) \log p(s) - \lambda_0 \left[\int ds\, p(s) - 1\right] - \lambda_1 \left[\int ds\, p(s) f(s) - a\right]$$

$$\frac{\delta \mathcal{L}}{\delta p(s)} = 1 + \log p(s) - \lambda_0 - \lambda_1 f(s) = 0$$

$$\Rightarrow \log p(s) = \lambda_0 + \lambda_1 f(s) - 1$$

$$\Rightarrow p(s) = \frac{1}{Z} e^{\lambda_1 f(s)}$$

The constants λ_0 and λ_1 can be found by solving the constraint equations. Thus,

$$\begin{array}{lcll} f(s) = s & \Rightarrow & p(s) = \frac{1}{Z} e^{\lambda_1 s}. & \text{Exponential (need } p(s) = 0 \text{ for } s < T). \\ f(s) = s^2 & \Rightarrow & p(s) = \frac{1}{Z} e^{\lambda_1 s^2}. & \text{Gaussian.} \end{array}$$

Both results together $\Rightarrow$ maximum entropy point process (for fixed mean arrival rate) is homogeneous Poisson – independent, exponentially distributed ISIs.

Channels

We now direct our focus to the conditional $P(R|S)$ which defines the **channel** linking S to R.

$$S \xrightarrow{P(R|S)} R$$

Channels

We now direct our focus to the conditional $P(R|S)$ which defines the **channel** linking S to R.

$$S \xrightarrow{P(R|S)} R$$

The mutual information

$$\mathbf{I}[S;R] = \sum_{s,r} P(s,r) \log \frac{P(s,r)}{P(s)P(r)} = \sum_{s,r} P(s) P(r|s) \log \frac{P(r|s)}{P(r)}$$

depends on marginals $P(s)$ and $P(r) = \sum_s P(r|s)P(s)$ as well and thus is unsuitable to characterise the conditional alone.

Channels

We now direct our focus to the conditional $P(R|S)$ which defines the **channel** linking S to R.

$$S \xrightarrow{P(R|S)} R$$

The mutual information

$$\mathbf{I}[S;R] = \sum_{s,r} P(s,r) \log \frac{P(s,r)}{P(s)P(r)} = \sum_{s,r} P(s)P(r|s) \log \frac{P(r|s)}{P(r)}$$

depends on marginals $P(s)$ and $P(r) = \sum_s P(r|s)P(s)$ as well and thus is unsuitable to characterise the conditional alone.

Instead, we characterise the channel by its **capacity**

$$\mathbf{C}_{R|S} = \sup_{P(s)} \mathbf{I}[S;R]$$

Thus the capacity gives the theoretical limit on the amount of information that can be transmitted over a channel. Clearly, this is limited by the properties of the noise.

Joint source-channel coding theorem

The remarkable central result of information theory.

$$S \xrightarrow{\textit{encoder}} \widetilde{S} \xrightarrow[\mathbf{C}_{R|\widetilde{S}}]{\textit{channel}} R \xrightarrow{\textit{decoder}} \widehat{T}$$

Any source ensemble S with entropy $\mathbf{H}[S] < \mathbf{C}_{R|\widetilde{S}}$ can be transmitted (in sufficiently long blocks) with $P_{error} \to 0$.

The proof is beyond our scope.

Some of the key ideas that appear in the proof are:

- block coding
- error correction
- joint typicality
- random codes

The channel coding problem

$$S \xrightarrow{\textit{encoder}} \widetilde{S} \xrightarrow[\mathbf{C}_{R|\widetilde{S}}]{\textit{channel}} R \xrightarrow{\textit{decoder}} \widehat{T}$$

Given channel $P(R|\widetilde{S})$ and source $P(S)$, find **encoding** $P(\widetilde{S}|S)$ (may be deterministic) to maximise $\mathbf{I}[S; R]$.

By data processing inequality, and defn of capacity:

$$\mathbf{I}[S; R] \leq \mathbf{I}[\widetilde{S}; R] \leq \mathbf{C}_{R|\widetilde{S}}$$

By JSCT, equality can be achieved (in the limit of increasing block size).

Thus $\mathbf{I}[\widetilde{S}; R]$ should saturate $\mathbf{C}_{R|\widetilde{S}}$.

See homework for an algorithm (Blahut-Arimoto) to find $P(\widetilde{S})$ that saturates $\mathbf{C}_{R|\widetilde{S}}$ for a general discrete channel.

Entropy maximisation

$$\mathbf{I}[\widetilde{S}; R] = \underbrace{\mathbf{H}[R]}_{\text{marginal entropy}} - \underbrace{\mathbf{H}\left[R|\widetilde{S}\right]}_{\text{noise entropy}}$$

Entropy maximisation

$$\mathbf{I}[\widetilde{S}; R] = \underbrace{\mathbf{H}[R]}_{\text{marginal entropy}} - \underbrace{\mathbf{H}\left[R|\widetilde{S}\right]}_{\text{noise entropy}}$$

If noise is small and "constant" $\Rightarrow$ maximise marginal entropy $\Rightarrow$ maximise $\mathbf{H}\left[\widetilde{S}\right]$

Entropy maximisation

$$\mathbf{I}[\widetilde{S}; R] = \underbrace{\mathbf{H}[R]}_{\text{marginal entropy}} - \underbrace{\mathbf{H}\left[R|\widetilde{S}\right]}_{\text{noise entropy}}$$

If noise is small and "constant" $\Rightarrow$ maximise marginal entropy $\Rightarrow$ maximise $\mathbf{H}\left[\widetilde{S}\right]$

Consider a (rate coding) neuron with $r \in [0, r_{\max}]$.

$$h(r) = -\int_0^{r_{\max}} dr\, p(r) \log p(r)$$

Entropy maximisation

$$\mathbf{I}[\widetilde{S}; R] = \underbrace{\mathbf{H}[R]}_{\text{marginal entropy}} - \underbrace{\mathbf{H}\left[R|\widetilde{S}\right]}_{\text{noise entropy}}$$

If noise is small and "constant" $\Rightarrow$ maximise marginal entropy $\Rightarrow$ maximise $\mathbf{H}\left[\widetilde{S}\right]$

Consider a (rate coding) neuron with $r \in [0, r_{\text{max}}]$.

$$h(r) = -\int_0^{r_{\text{max}}} dr\, p(r) \log p(r)$$

To maximise the marginal entropy, we add a Lagrange multiplier (μ) to enforce normalisation and then differentiate

$$\frac{\delta}{\delta p(r)}\left[h(r) - \mu \int_0^{r_{\text{max}}} p(r)\right] = \begin{cases} -\log p(r) - 1 - \mu & r \in [0, r_{\text{max}}] \\ 0 & \text{otherwise} \end{cases}$$

Entropy maximisation

$$\mathbf{I}[\widetilde{S}; R] = \underbrace{\mathbf{H}[R]}_{\text{marginal entropy}} - \underbrace{\mathbf{H}\left[R|\widetilde{S}\right]}_{\text{noise entropy}}$$

If noise is small and "constant" $\Rightarrow$ maximise marginal entropy $\Rightarrow$ maximise $\mathbf{H}\left[\widetilde{S}\right]$

Consider a (rate coding) neuron with $r \in [0, r_{\text{max}}]$.

$$h(r) = -\int_0^{r_{\text{max}}} dr\, p(r) \log p(r)$$

To maximise the marginal entropy, we add a Lagrange multiplier (μ) to enforce normalisation and then differentiate

$$\frac{\delta}{\delta p(r)} \left[h(r) - \mu \int_0^{r_{\text{max}}} p(r) \right] = \begin{cases} -\log p(r) - 1 - \mu & r \in [0, r_{\text{max}}] \\ 0 & \text{otherwise} \end{cases}$$

$\Rightarrow p(r) = \mathit{const}$ for $r \in [0, r_{\text{max}}]$

Entropy maximisation

$$\mathbf{I}[\widetilde{S}; R] = \underbrace{\mathbf{H}[R]}_{\text{marginal entropy}} - \underbrace{\mathbf{H}\left[R|\widetilde{S}\right]}_{\text{noise entropy}}$$

If noise is small and "constant" $\Rightarrow$ maximise marginal entropy $\Rightarrow$ maximise $\mathbf{H}\left[\widetilde{S}\right]$

Consider a (rate coding) neuron with $r \in [0, r_{\max}]$.

$$h(r) = -\int_0^{r_{\max}} dr\, p(r) \log p(r)$$

To maximise the marginal entropy, we add a Lagrange multiplier (μ) to enforce normalisation and then differentiate

$$\frac{\delta}{\delta p(r)} \left[h(r) - \mu \int_0^{r_{\max}} p(r) \right] = \begin{cases} -\log p(r) - 1 - \mu & r \in [0, r_{\max}] \\ 0 & \text{otherwise} \end{cases}$$

$\Rightarrow p(r) = const$ for $r \in [0, r_{\max}]$
i.e.

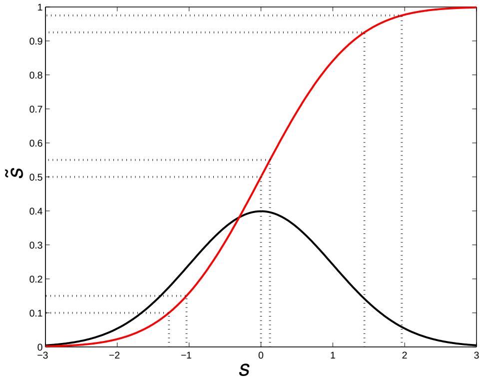
$$p(r) = \begin{cases} \frac{1}{r_{\max}} & r \in [0, r_{\max}] \\ 0 & \text{otherwise} \end{cases}$$

Histogram Equalisation

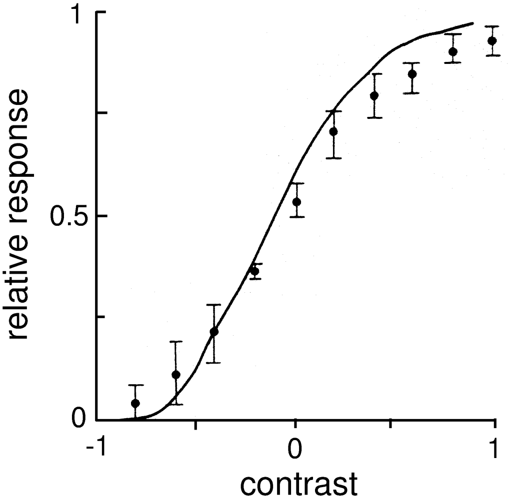
Suppose $r = \tilde{s} + \eta$ where η represents a (relatively small) source of noise. Consider deterministic encoding $\tilde{s} = f(s)$. How do we ensure that $p(r) = 1/r_{\max}$?

$$\frac{1}{r_{\max}} = p(r) \approx p(\tilde{s}) = \frac{p(s)}{f'(s)} \qquad \Rightarrow f'(s) = r_{\max}\ p(s)$$

$$\Rightarrow f(s) = r_{\max} \int_{-\infty}^{s} ds'\ p(s')$$



Histogram Equalisation



Laughlin (1981)

Gaussian channel

A similar idea of output-entropy maximisation appears in the theory of Gaussian channel coding, where it is called the **water filling** algorithm.

Gaussian channel

A similar idea of output-entropy maximisation appears in the theory of Gaussian channel coding, where it is called the **water filling** algorithm.

We will need the differential entropy of a (multivariate) Gaussian distribution:

Gaussian channel

A similar idea of output-entropy maximisation appears in the theory of Gaussian channel coding, where it is called the **water filling** algorithm.

We will need the differential entropy of a (multivariate) Gaussian distribution:

Let

$$p(\mathbf{Z}) = |2\pi\Sigma|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{Z}-\boldsymbol{\mu})^{\mathsf{T}}\Sigma^{-1}(\mathbf{Z}-\boldsymbol{\mu})\right],$$

then,

$$h(\mathbf{Z}) = -\int d\mathbf{Z}\; p(\mathbf{Z}) \left[-\frac{1}{2}\log|2\pi\Sigma| - \frac{1}{2}(\mathbf{Z}-\boldsymbol{\mu})^{\mathsf{T}}\Sigma^{-1}(\mathbf{Z}-\boldsymbol{\mu})\right]$$

Gaussian channel

A similar idea of output-entropy maximisation appears in the theory of Gaussian channel coding, where it is called the **water filling** algorithm.

We will need the differential entropy of a (multivariate) Gaussian distribution:

Let

$$p(\mathbf{Z}) = |2\pi\Sigma|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{Z}-\boldsymbol{\mu})^{\mathsf{T}}\Sigma^{-1}(\mathbf{Z}-\boldsymbol{\mu})\right],$$

then,

$$\begin{aligned} h(\mathbf{Z}) &= -\int d\mathbf{Z}\, p(\mathbf{Z}) \left[-\frac{1}{2}\log|2\pi\Sigma| - \frac{1}{2}(\mathbf{Z}-\boldsymbol{\mu})^{\mathsf{T}}\Sigma^{-1}(\mathbf{Z}-\boldsymbol{\mu})\right] \\ &= \frac{1}{2}\log|2\pi\Sigma| + \frac{1}{2}\int d\mathbf{Z}\, p(\mathbf{Z}) \mathrm{Tr}\left[\Sigma^{-1}(\mathbf{Z}-\boldsymbol{\mu})(\mathbf{Z}-\boldsymbol{\mu})^{\mathsf{T}}\right] \end{aligned}$$

Gaussian channel

A similar idea of output-entropy maximisation appears in the theory of Gaussian channel coding, where it is called the **water filling** algorithm.

We will need the differential entropy of a (multivariate) Gaussian distribution:

Let

$$p(\mathbf{Z}) = |2\pi\Sigma|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{Z}-\boldsymbol{\mu})^{\mathsf{T}}\Sigma^{-1}(\mathbf{Z}-\boldsymbol{\mu})\right],$$

then,

$$\begin{aligned} h(\mathbf{Z}) &= -\int d\mathbf{Z}\, p(\mathbf{Z}) \left[-\frac{1}{2}\log|2\pi\Sigma| - \frac{1}{2}(\mathbf{Z}-\boldsymbol{\mu})^{\mathsf{T}}\Sigma^{-1}(\mathbf{Z}-\boldsymbol{\mu})\right] \\ &= \frac{1}{2}\log|2\pi\Sigma| + \frac{1}{2}\int d\mathbf{Z}\, p(\mathbf{Z}) \mathrm{Tr}\left[\Sigma^{-1}(\mathbf{Z}-\boldsymbol{\mu})(\mathbf{Z}-\boldsymbol{\mu})^{\mathsf{T}}\right] \\ &= \frac{1}{2}\log|2\pi\Sigma| + \frac{1}{2}\mathrm{Tr}\left[\Sigma^{-1}\Sigma\right] \end{aligned}$$

Gaussian channel

A similar idea of output-entropy maximisation appears in the theory of Gaussian channel coding, where it is called the **water filling** algorithm.

We will need the differential entropy of a (multivariate) Gaussian distribution:

Let

$$p(\mathbf{Z}) = |2\pi\Sigma|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{Z}-\boldsymbol{\mu})^{\mathsf{T}}\Sigma^{-1}(\mathbf{Z}-\boldsymbol{\mu})\right],$$

then,

$$\begin{aligned}
h(\mathbf{Z}) &= -\int d\mathbf{Z}\; p(\mathbf{Z}) \left[-\frac{1}{2}\log|2\pi\Sigma| - \frac{1}{2}(\mathbf{Z}-\boldsymbol{\mu})^{\mathsf{T}}\Sigma^{-1}(\mathbf{Z}-\boldsymbol{\mu})\right] \\
&= \frac{1}{2}\log|2\pi\Sigma| + \frac{1}{2}\int d\mathbf{Z}\; p(\mathbf{Z}) \mathrm{Tr}\left[\Sigma^{-1}(\mathbf{Z}-\boldsymbol{\mu})(\mathbf{Z}-\boldsymbol{\mu})^{\mathsf{T}}\right] \\
&= \frac{1}{2}\log|2\pi\Sigma| + \frac{1}{2}\mathrm{Tr}\left[\Sigma^{-1}\Sigma\right] \\
&= \frac{1}{2}\log|2\pi\Sigma| + \frac{1}{2}d \quad (\log e)
\end{aligned}$$

Gaussian channel

A similar idea of output-entropy maximisation appears in the theory of Gaussian channel coding, where it is called the **water filling** algorithm.

We will need the differential entropy of a (multivariate) Gaussian distribution:

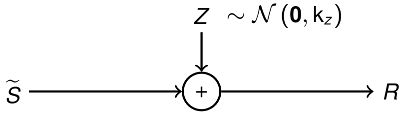
Let

$$p(\mathbf{Z}) = |2\pi\Sigma|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{Z}-\boldsymbol{\mu})^{\mathsf{T}}\Sigma^{-1}(\mathbf{Z}-\boldsymbol{\mu})\right],$$

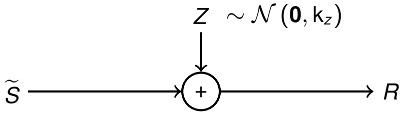
then,

$$\begin{aligned}
h(\mathbf{Z}) &= -\int d\mathbf{Z}\, p(\mathbf{Z}) \left[-\frac{1}{2}\log|2\pi\Sigma| - \frac{1}{2}(\mathbf{Z}-\boldsymbol{\mu})^{\mathsf{T}}\Sigma^{-1}(\mathbf{Z}-\boldsymbol{\mu})\right] \\
&= \frac{1}{2}\log|2\pi\Sigma| + \frac{1}{2}\int d\mathbf{Z}\, p(\mathbf{Z})\mathrm{Tr}\left[\Sigma^{-1}(\mathbf{Z}-\boldsymbol{\mu})(\mathbf{Z}-\boldsymbol{\mu})^{\mathsf{T}}\right] \\
&= \frac{1}{2}\log|2\pi\Sigma| + \frac{1}{2}\mathrm{Tr}\left[\Sigma^{-1}\Sigma\right] \\
&= \frac{1}{2}\log|2\pi\Sigma| + \frac{1}{2}d \quad (\log e) \\
&= \frac{1}{2}\log|2\pi e\Sigma|
\end{aligned}$$

Gaussian channel – white noise

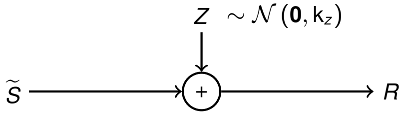


Gaussian channel – white noise



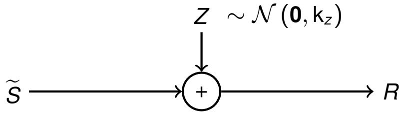
$$\mathbf{I}[\widetilde{S}; R] = h(R) - h(R|\widetilde{S})$$

Gaussian channel – white noise



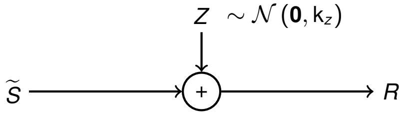
$$
\begin{aligned}
\mathbf{I}[\widetilde{S}; R] &= h(R) - h(R|\widetilde{S}) \\
&= h(R) - h(\widetilde{S} + Z|\widetilde{S})
\end{aligned}
$$

Gaussian channel – white noise



$$\begin{aligned}\mathbf{I}[\widetilde{S}; R] &= h(R) - h(R|\widetilde{S}) \\ &= h(R) - h(\widetilde{S} + Z|\widetilde{S}) \\ &= h(R) - h(Z)\end{aligned}$$

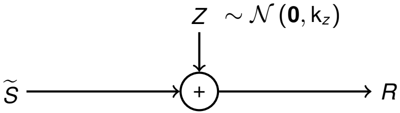
Gaussian channel – white noise



$$
\begin{aligned}
\mathbf{I}[\widetilde{S}; R] &= h(R) - h(R|\widetilde{S}) \\
&= h(R) - h(\widetilde{S} + Z|\widetilde{S}) \\
&= h(R) - h(Z)
\end{aligned}
$$

$$\Rightarrow \mathbf{I}[\widetilde{S}; R] = h(R) - \frac{1}{2}\log 2\pi e \mathrm{k}_z.$$

Gaussian channel – white noise

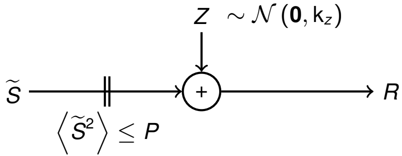


$$
\begin{aligned}
\mathbf{I}[\widetilde{S}; R] &= h(R) - h(R|\widetilde{S}) \\
&= h(R) - h(\widetilde{S} + Z|\widetilde{S}) \\
&= h(R) - h(Z)
\end{aligned}
$$

$$
\Rightarrow \mathbf{I}[\widetilde{S}; R] = h(R) - \frac{1}{2}\log 2\pi e \mathrm{k}_z.
$$

Without constraint, $h(R) \to \infty$ and $\mathbf{C}_{R|\widetilde{S}} = \infty$.

Gaussian channel – white noise



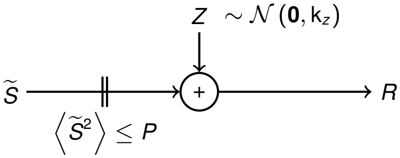
$$\begin{aligned}
\mathbf{I}[\widetilde{S}; R] &= h(R) - h(R|\widetilde{S}) \\
&= h(R) - h(\widetilde{S} + Z|\widetilde{S}) \\
&= h(R) - h(Z)
\end{aligned}$$

$$\Rightarrow \mathbf{I}[\widetilde{S}; R] = h(R) - \frac{1}{2} \log 2\pi e \mathrm{k}_z.$$

Without constraint, $h(R) \to \infty$ and $\mathbf{C}_{R|\widetilde{S}} = \infty$.

Therefore, constrain $\frac{1}{n} \sum_{i=1}^{n} \tilde{s}_i^2 \leq \mathrm{P}$.

Gaussian channel – white noise



$$\begin{aligned}\mathbf{I}[\widetilde{S}; R] &= h(R) - h(R|\widetilde{S})\\ &= h(R) - h(\widetilde{S} + Z|\widetilde{S})\\ &= h(R) - h(Z)\end{aligned}$$

$$\Rightarrow \mathbf{I}[\widetilde{S}; R] = h(R) - \frac{1}{2}\log 2\pi e\mathrm{k}_z.$$

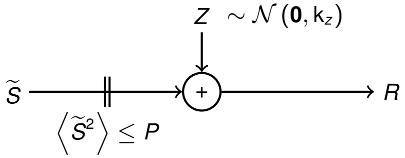
Without constraint, $h(R) \to \infty$ and $\mathbf{C}_{R|\widetilde{S}} = \infty$.

Therefore, constrain $\frac{1}{n}\sum_{i=1}^{n} \tilde{s}_i^2 \leq \mathrm{P}$.

Then,

$$\left\langle R^2 \right\rangle = \left\langle (\widetilde{S} + Z)^2 \right\rangle = \left\langle \widetilde{S}^2 + Z^2 + 2\widetilde{S}Z \right\rangle$$

Gaussian channel – white noise



$$\begin{aligned}
\mathbf{I}[\widetilde{S}; R] &= h(R) - h(R|\widetilde{S}) \\
&= h(R) - h(\widetilde{S} + Z|\widetilde{S}) \\
&= h(R) - h(Z)
\end{aligned}$$

$$\Rightarrow \mathbf{I}[\widetilde{S}; R] = h(R) - \frac{1}{2}\log 2\pi e \mathrm{k}_z.$$

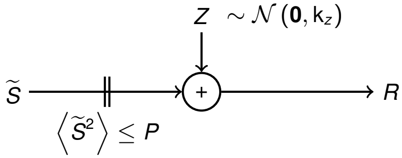
Without constraint, $h(R) \to \infty$ and $\mathbf{C}_{R|\widetilde{S}} = \infty$.

Therefore, constrain $\frac{1}{n}\sum_{i=1}^{n} \tilde{s}_i^2 \leq \mathrm{P}$.

Then,

$$\left\langle R^2 \right\rangle = \left\langle (\widetilde{S} + Z)^2 \right\rangle = \left\langle \widetilde{S}^2 + Z^2 + 2\widetilde{S}Z \right\rangle \leq \mathrm{P} + \mathrm{k}_z + 0$$

Gaussian channel – white noise



$$\begin{aligned} \mathbf{I}[\widetilde{S}; R] &= h(R) - h(R|\widetilde{S}) \\ &= h(R) - h(\widetilde{S} + Z|\widetilde{S}) \\ &= h(R) - h(Z) \end{aligned}$$

$$\Rightarrow \mathbf{I}[\widetilde{S}; R] = h(R) - \frac{1}{2}\log 2\pi e \mathrm{k}_z.$$

Without constraint, $h(R) \to \infty$ and $\mathbf{C}_{R|\widetilde{S}} = \infty$.

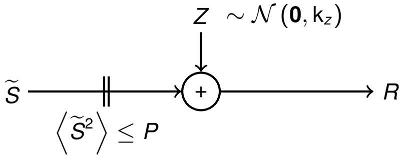
Therefore, constrain $\frac{1}{n}\sum_{i=1}^{n} \tilde{s}_i^2 \leq \mathrm{P}$.

Then,

$$\left\langle R^2 \right\rangle = \left\langle (\widetilde{S} + Z)^2 \right\rangle = \left\langle \widetilde{S}^2 + Z^2 + 2\widetilde{S}Z \right\rangle \leq \mathrm{P} + \mathrm{k}_z + 0$$

$$\Rightarrow h(R) \leq h(\mathcal{N}(0, \mathrm{P} + \mathrm{k}_z))$$

Gaussian channel – white noise



$$
\begin{aligned}
\mathbf{I}[\widetilde{S}; R] &= h(R) - h(R|\widetilde{S}) \\
&= h(R) - h(\widetilde{S} + Z|\widetilde{S}) \\
&= h(R) - h(Z)
\end{aligned}
$$

$$\Rightarrow \mathbf{I}[\widetilde{S}; R] = h(R) - \frac{1}{2} \log 2\pi e \mathrm{k}_z .$$

Without constraint, $h(R) \to \infty$ and $\mathbf{C}_{R|\widetilde{S}} = \infty$.

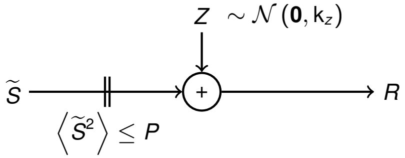
Therefore, constrain $\frac{1}{n} \sum_{i=1}^{n} \tilde{s}_i^2 \leq \mathrm{P}$.

Then,

$$\left\langle R^2 \right\rangle = \left\langle (\widetilde{S} + Z)^2 \right\rangle = \left\langle \widetilde{S}^2 + Z^2 + 2\widetilde{S}Z \right\rangle \leq \mathrm{P} + \mathrm{k}_z + 0$$

$$\Rightarrow h(R) \leq h(\mathcal{N}(0, \mathrm{P} + \mathrm{k}_z)) = \frac{1}{2} \log 2\pi e(\mathrm{P} + \mathrm{k}_z)$$

Gaussian channel – white noise



$$\begin{aligned}\mathbf{I}[\widetilde{S}; R] &= h(R) - h(R|\widetilde{S}) \\ &= h(R) - h(\widetilde{S} + Z|\widetilde{S}) \\ &= h(R) - h(Z)\end{aligned}$$

$$\Rightarrow \mathbf{I}[\widetilde{S}; R] = h(R) - \frac{1}{2}\log 2\pi e \mathrm{k}_z.$$

Without constraint, $h(R) \to \infty$ and $\mathbf{C}_{R|\widetilde{S}} = \infty$.

Therefore, constrain $\frac{1}{n}\sum_{i=1}^{n} \tilde{s}_i^2 \le \mathrm{P}$.

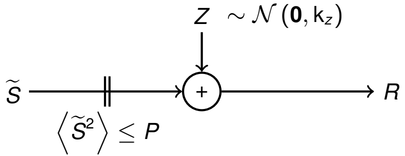
Then,

$$\left\langle R^2 \right\rangle = \left\langle (\widetilde{S} + Z)^2 \right\rangle = \left\langle \widetilde{S}^2 + Z^2 + 2\widetilde{S}Z \right\rangle \le \mathrm{P} + \mathrm{k}_z + 0$$

$$\Rightarrow h(R) \le h(\mathcal{N}(0, \mathrm{P} + \mathrm{k}_z)) = \frac{1}{2}\log 2\pi e(\mathrm{P} + \mathrm{k}_z)$$

$$\Rightarrow \mathbf{I}[\widetilde{S}; R] \le \frac{1}{2}\log 2\pi e(\mathrm{P} + \mathrm{k}_z) - \frac{1}{2}\log 2\pi e \mathrm{k}_z$$

Gaussian channel – white noise



$$\begin{aligned}
\mathbf{I}[\widetilde{S}; R] &= h(R) - h(R|\widetilde{S}) \\
&= h(R) - h(\widetilde{S} + Z|\widetilde{S}) \\
&= h(R) - h(Z)
\end{aligned}$$

$$\Rightarrow \mathbf{I}[\widetilde{S}; R] = h(R) - \frac{1}{2} \log 2\pi e \mathrm{k}_z.$$

Without constraint, $h(R) \to \infty$ and $\mathbf{C}_{R|\widetilde{S}} = \infty$.

Therefore, constrain $\displaystyle \frac{1}{n} \sum_{i=1}^{n} \tilde{s}_i^2 \le \mathrm{P}$.

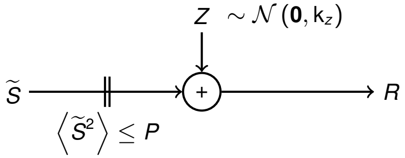
Then,

$$\left\langle R^2 \right\rangle = \left\langle (\widetilde{S} + Z)^2 \right\rangle = \left\langle \widetilde{S}^2 + Z^2 + 2\widetilde{S}Z \right\rangle \le \mathrm{P} + \mathrm{k}_z + 0$$

$$\Rightarrow h(R) \le h(\mathcal{N}(0, \mathrm{P} + \mathrm{k}_z)) = \frac{1}{2} \log 2\pi e(\mathrm{P} + \mathrm{k}_z)$$

$$\Rightarrow \mathbf{I}[\widetilde{S}; R] \le \frac{1}{2} \log 2\pi e(\mathrm{P} + \mathrm{k}_z) - \frac{1}{2} \log 2\pi e \mathrm{k}_z = \frac{1}{2} \log 2\pi e \left(1 + \frac{\mathrm{P}}{\mathrm{k}_z}\right)$$

Gaussian channel – white noise



$$\begin{aligned}
\mathbf{I}[\widetilde{S}; R] &= h(R) - h(R|\widetilde{S}) \\
&= h(R) - h(\widetilde{S} + Z|\widetilde{S}) \\
&= h(R) - h(Z)
\end{aligned}$$

$$\Rightarrow \mathbf{I}[\widetilde{S}; R] = h(R) - \frac{1}{2} \log 2\pi e \mathrm{k}_z.$$

Without constraint, $h(R) \to \infty$ and $\mathbf{C}_{R|\widetilde{S}} = \infty$.

Therefore, constrain $\frac{1}{n} \sum_{i=1}^{n} \tilde{s}_i^2 \leq \mathrm{P}$.

Then,

$$\left\langle R^2 \right\rangle = \left\langle (\widetilde{S} + Z)^2 \right\rangle = \left\langle \widetilde{S}^2 + Z^2 + 2\widetilde{S}Z \right\rangle \leq \mathrm{P} + \mathrm{k}_z + 0$$

$$\Rightarrow h(R) \leq h(\mathcal{N}(0, \mathrm{P} + \mathrm{k}_z)) = \frac{1}{2} \log 2\pi e(\mathrm{P} + \mathrm{k}_z)$$

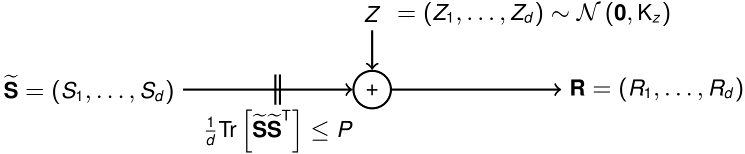
$$\Rightarrow \mathbf{I}[\widetilde{S}; R] \leq \frac{1}{2} \log 2\pi e(\mathrm{P} + \mathrm{k}_z) - \frac{1}{2} \log 2\pi e \mathrm{k}_z = \frac{1}{2} \log 2\pi e \left(1 + \frac{\mathrm{P}}{\mathrm{k}_z}\right)$$

$$\mathbf{C}_{R|\widetilde{S}} = \frac{1}{2} \log 2\pi e \left(1 + \frac{\mathrm{P}}{\mathrm{k}_z}\right)$$

The capacity is achieved iff $R \sim \mathcal{N}(0, \mathrm{P} + \mathrm{k}_z) \quad \Rightarrow \widetilde{S} \sim \mathcal{N}(0, \mathrm{P})$.

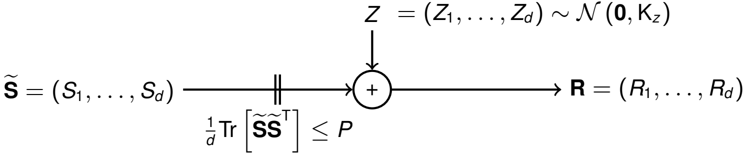
Gaussian channel – correlated noise

Now consider a vector Gaussian channel:



Gaussian channel – correlated noise

Now consider a vector Gaussian channel:

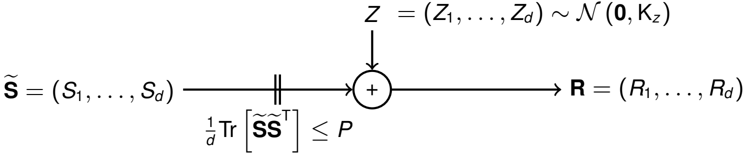


Following the same approach as before:

$$\mathbf{I}[\widetilde{\mathbf{S}}; \mathbf{R}] = h(\mathbf{R}) - h(\mathbf{Z}) \leq \frac{1}{2}\log\left[(2\pi e)^d \; |\mathrm{K}_{\tilde{s}} + \mathrm{K}_z|\right] - \frac{1}{2}\log\left[(2\pi e)^d \; |\mathrm{K}_z|\right],$$

Gaussian channel – correlated noise

Now consider a vector Gaussian channel:



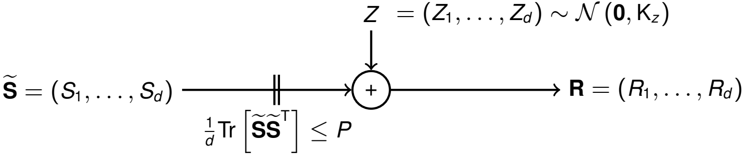
Following the same approach as before:

$$\mathbf{I}[\widetilde{\mathbf{S}}; \mathbf{R}] = h(\mathbf{R}) - h(\mathbf{Z}) \leq \frac{1}{2}\log\left[(2\pi e)^d \; |\mathrm{K}_{\tilde{s}} + \mathrm{K}_z|\right] - \frac{1}{2}\log\left[(2\pi e)^d \; |\mathrm{K}_z|\right],$$

$\Rightarrow \mathbf{C}_{R|S}$ achieved when $\widetilde{\mathbf{S}}$ (and thus $\mathbf{R}$) $\sim \mathcal{N}$, with $|\mathrm{K}_{\tilde{s}} + \mathrm{K}_z|$ max given $\frac{1}{d}\mathrm{Tr}\left[\mathrm{K}_{\tilde{s}}\right] \leq P$.

Gaussian channel – correlated noise

Now consider a vector Gaussian channel:



Following the same approach as before:

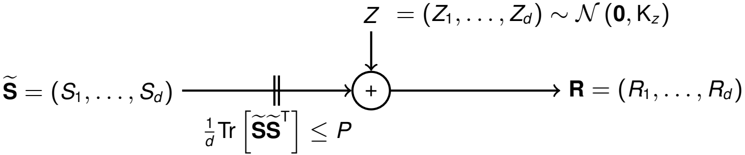
$$\mathbf{I}[\widetilde{\mathbf{S}}; \mathbf{R}] = h(\mathbf{R}) - h(\mathbf{Z}) \leq \frac{1}{2}\log\left[(2\pi e)^d \; |\mathrm{K}_{\tilde{s}} + \mathrm{K}_z|\right] - \frac{1}{2}\log\left[(2\pi e)^d \; |\mathrm{K}_z|\right],$$

$\Rightarrow$ $\mathbf{C}_{R|S}$ achieved when $\widetilde{\mathbf{S}}$ (and thus $\mathbf{R}$) $\sim \mathcal{N}$, with $|\mathrm{K}_{\tilde{s}} + \mathrm{K}_z|$ max given $\frac{1}{d}\mathrm{Tr}\left[\mathrm{K}_{\tilde{s}}\right] \leq P$.

Diagonalise K_z $\Rightarrow$$\mathrm{K}_{\tilde{s}}$ is diagonal in same basis.

Gaussian channel – correlated noise

Now consider a vector Gaussian channel:



Following the same approach as before:

$$\mathbf{I}[\widetilde{\mathbf{S}}; \mathbf{R}] = h(\mathbf{R}) - h(\mathbf{Z}) \le \frac{1}{2}\log\left[(2\pi e)^d \; |\mathsf{K}_{\tilde{s}} + \mathsf{K}_z|\right] - \frac{1}{2}\log\left[(2\pi e)^d \; |\mathsf{K}_z|\right],$$

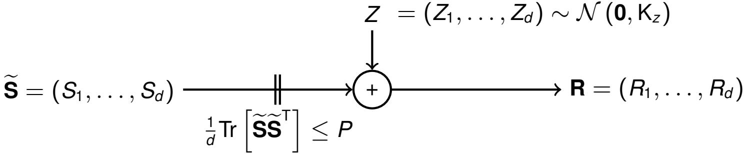
$\Rightarrow$ $\mathbf{C}_{R|S}$ achieved when $\widetilde{\mathbf{S}}$ (and thus $\mathbf{R}$) $\sim \mathcal{N}$, with $|\mathsf{K}_{\tilde{s}} + \mathsf{K}_z|$ max given $\frac{1}{d}\mathrm{Tr}\left[\mathsf{K}_{\tilde{s}}\right] \le P$.

Diagonalise K_z $\Rightarrow$$\mathsf{K}_{\tilde{s}}$ is diagonal in same basis.

For **stationary** noise (wrt dimension indexed by d) this can be achieved by a Fourier transform $\Rightarrow$ index diagonal elements by ω.

Gaussian channel – correlated noise

Now consider a vector Gaussian channel:



Following the same approach as before:

$$\mathbf{I}[\widetilde{\mathbf{S}}; \mathbf{R}] = h(\mathbf{R}) - h(\mathbf{Z}) \leq \frac{1}{2}\log\left[(2\pi e)^d \; |\mathrm{K}_{\tilde{s}} + \mathrm{K}_z|\right] - \frac{1}{2}\log\left[(2\pi e)^d \; |\mathrm{K}_z|\right],$$

$\Rightarrow$ $\mathbf{C}_{R|S}$ achieved when $\widetilde{\mathbf{S}}$ (and thus $\mathbf{R}$) $\sim \mathcal{N}$, with $|\mathrm{K}_{\tilde{s}} + \mathrm{K}_z|$ max given $\frac{1}{d}\mathrm{Tr}\left[\mathrm{K}_{\tilde{s}}\right] \leq P$.

Diagonalise K_z $\Rightarrow$$\mathrm{K}_{\tilde{s}}$ is diagonal in same basis.

For **stationary** noise (wrt dimension indexed by d) this can be achieved by a Fourier transform $\Rightarrow$ index diagonal elements by ω.

$$\mathrm{k}^*_{\tilde{s}}(\omega) = \operatorname{argmax} \prod_{\omega} \left(\mathrm{k}_{\tilde{s}}(\omega) + \mathrm{k}_z(\omega)\right) \qquad \text{such that } \frac{1}{d}\sum \mathrm{k}_{\tilde{s}}(\omega) \leq P$$

Water filling

Assume that optimum is achieved for max. input power.

$$\mathrm{k}_{\tilde{s}}^{*}(\omega) = \operatorname{argmax} \left[\sum_{\omega} \log\left(\mathrm{k}_{\tilde{s}}(\omega) + \mathrm{k}_{z}(\omega)\right) - \lambda \left(\frac{1}{d} \sum_{\omega} \mathrm{k}_{\tilde{s}}(\omega) - P \right) \right]$$

Water filling

Assume that optimum is achieved for max. input power.

$$\mathrm{k}_{\tilde{s}}^{*}(\omega) = \operatorname{argmax}\left[\sum_{\omega} \log\left(\mathrm{k}_{\tilde{s}}(\omega) + \mathrm{k}_{z}(\omega)\right) - \lambda\left(\frac{1}{d}\sum_{\omega} \mathrm{k}_{\tilde{s}}(\omega) - P\right)\right]$$

$$\Rightarrow \quad \frac{1}{\mathrm{k}_{\tilde{s}}^{*}(\omega) + \mathrm{k}_{z}(\omega)} - \frac{\lambda}{d} = 0$$

Water filling

Assume that optimum is achieved for max. input power.

$$\mathrm{k}^*_{\tilde{s}}(\omega) = \operatorname{argmax}\left[\sum_\omega \log\left(\mathrm{k}_{\tilde{s}}(\omega) + \mathrm{k}_z(\omega)\right) - \lambda\left(\frac{1}{d}\sum_\omega \mathrm{k}_{\tilde{s}}(\omega) - P\right)\right]$$

$$\Rightarrow \quad \frac{1}{\mathrm{k}^*_{\tilde{s}}(\omega) + \mathrm{k}_z(\omega)} - \frac{\lambda}{d} = 0$$

$$\Rightarrow \quad \mathrm{k}^*_{\tilde{s}}(\omega) + \mathrm{k}_z(\omega) = \nu \quad (const.)$$

Water filling

Assume that optimum is achieved for max. input power.

$$\mathrm{k}_{\tilde{s}}^*(\omega) = \mathrm{argmax}\left[\sum_\omega \log\left(\mathrm{k}_{\tilde{s}}(\omega) + \mathrm{k}_z(\omega)\right) - \lambda\left(\frac{1}{d}\sum_\omega \mathrm{k}_{\tilde{s}}(\omega) - P\right)\right]$$

$$\Rightarrow \quad \frac{1}{\mathrm{k}_{\tilde{s}}^*(\omega) + \mathrm{k}_z(\omega)} - \frac{\lambda}{d} = 0$$

$$\Rightarrow \quad \mathrm{k}_{\tilde{s}}^*(\omega) + \mathrm{k}_z(\omega) = \nu \quad (const.)$$

$$(\mathrm{k}_{\tilde{s}} \geq 0) \Rightarrow \quad \mathrm{k}_{\tilde{s}}^*(\omega) = [\nu - \mathrm{k}_z(\omega)]^+$$

Water filling

Assume that optimum is achieved for max. input power.

$$\mathrm{k}^*_{\tilde{s}}(\omega) = \operatorname{argmax}\left[\sum_{\omega} \log\left(\mathrm{k}_{\tilde{s}}(\omega) + \mathrm{k}_z(\omega)\right) - \lambda\left(\frac{1}{d}\sum_{\omega}\mathrm{k}_{\tilde{s}}(\omega) - P\right)\right]$$

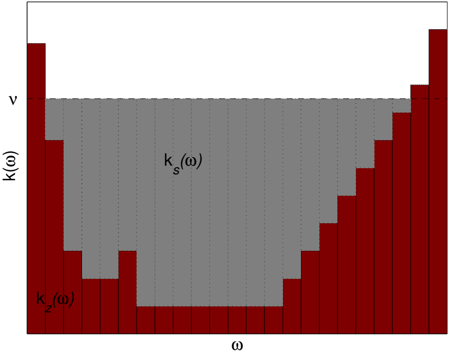
$$\Rightarrow \quad \frac{1}{\mathrm{k}^*_{\tilde{s}}(\omega) + \mathrm{k}_z(\omega)} - \frac{\lambda}{d} = 0$$

$$\Rightarrow \quad \mathrm{k}^*_{\tilde{s}}(\omega) + \mathrm{k}_z(\omega) = \nu \quad (const.)$$

$$(\mathrm{k}_{\tilde{s}} \geq 0) \Rightarrow \quad \mathrm{k}^*_{\tilde{s}}(\omega) = [\nu - \mathrm{k}_z(\omega)]^+$$

Waterfilling: choose ν so

$$\sum_{\omega} \mathrm{k}_{\tilde{s}}(\omega) = d \cdot P$$



R is white or decorrelated (within power budget) ⇒variance equalisation.

Decorrelation at the retina

Atick and Redlich (1992) argued that the retina decorrelates natural spatial statistics.

Decorrelation at the retina

Atick and Redlich (1992) argued that the retina decorrelates natural spatial statistics.

RGCs exhibit roughly linear (centre-surround) processing:

$$r_{\mathbf{a}} - \langle r_{\mathbf{a}} \rangle = \int d\mathbf{x} \; \underbrace{D_s(\mathbf{x} - \mathbf{a})}_{\text{filter}} \; \underbrace{s(\mathbf{x})}_{\text{stimulus}}$$

Decorrelation at the retina

Atick and Redlich (1992) argued that the retina decorrelates natural spatial statistics.

RGCs exhibit roughly linear (centre-surround) processing:

$$r_{\mathbf{a}} - \langle r_{\mathbf{a}} \rangle = \int d\mathbf{x} \; \underbrace{D_s(\mathbf{x} - \mathbf{a})}_{\text{filter}} \; \underbrace{s(\mathbf{x})}_{\text{stimulus}}$$

Therefore the correlation (covariance) between cells is

$$\begin{aligned} Q_r(\mathbf{a}, \mathbf{b}) &= \left\langle \int d\mathbf{x} \; d\mathbf{y} \; D_s(\mathbf{x} - \mathbf{a}) D_s(\mathbf{y} - \mathbf{b}) s(\mathbf{x}) s(\mathbf{y}) \right\rangle \\ &= \int d\mathbf{x} \; d\mathbf{y} \; D_s(\mathbf{x} - \mathbf{a}) D_s(\mathbf{y} - \mathbf{b}) \underbrace{\langle s(\mathbf{x}) s(\mathbf{y}) \rangle}_{Q_s(\mathbf{x}, \mathbf{y})} \end{aligned}$$

Decorrelation at the retina

Atick and Redlich (1992) argued that the retina decorrelates natural spatial statistics.

RGCs exhibit roughly linear (centre-surround) processing:

$$r_{\mathbf{a}} - \langle r_{\mathbf{a}} \rangle = \int d\mathbf{x} \; \underbrace{D_s(\mathbf{x} - \mathbf{a})}_{\text{filter}} \; \underbrace{s(\mathbf{x})}_{\text{stimulus}}$$

Therefore the correlation (covariance) between cells is

$$\begin{aligned} Q_r(\mathbf{a}, \mathbf{b}) &= \left\langle \int d\mathbf{x} \; d\mathbf{y} \; D_s(\mathbf{x} - \mathbf{a}) D_s(\mathbf{y} - \mathbf{b}) s(\mathbf{x}) s(\mathbf{y}) \right\rangle \\ &= \int d\mathbf{x} \; d\mathbf{y} \; D_s(\mathbf{x} - \mathbf{a}) D_s(\mathbf{y} - \mathbf{b}) \underbrace{\langle s(\mathbf{x}) s(\mathbf{y}) \rangle}_{Q_s(\mathbf{x},\mathbf{y})} \end{aligned}$$

Using (spatial) stationarity, we can transform to the Fourier domain:

$$\widetilde{Q}_r(\mathbf{k}) = |\widetilde{D}_s(\mathbf{k})|^2 \widetilde{Q}_s(\mathbf{k})$$

Decorrelation at the retina

Atick and Redlich (1992) argued that the retina decorrelates natural spatial statistics.

RGCs exhibit roughly linear (centre-surround) processing:

$$r_{\mathbf{a}} - \langle r_{\mathbf{a}} \rangle = \int d\mathbf{x} \; \underbrace{D_s(\mathbf{x} - \mathbf{a})}_{\text{filter}} \; \underbrace{s(\mathbf{x})}_{\text{stimulus}}$$

Therefore the correlation (covariance) between cells is

$$\begin{aligned} Q_r(\mathbf{a}, \mathbf{b}) &= \left\langle \int d\mathbf{x} \; d\mathbf{y} \; D_s(\mathbf{x} - \mathbf{a}) D_s(\mathbf{y} - \mathbf{b}) s(\mathbf{x}) s(\mathbf{y}) \right\rangle \\ &= \int d\mathbf{x} \; d\mathbf{y} \; D_s(\mathbf{x} - \mathbf{a}) D_s(\mathbf{y} - \mathbf{b}) \underbrace{\langle s(\mathbf{x}) s(\mathbf{y}) \rangle}_{Q_s(\mathbf{x}, \mathbf{y})} \end{aligned}$$

Using (spatial) stationarity, we can transform to the Fourier domain:

$$\widetilde{Q}_r(\mathbf{k}) = |\widetilde{D}_s(\mathbf{k})|^2 \widetilde{Q}_s(\mathbf{k})$$

and thus output decorrelation requires

$$|\widetilde{D}_s(\mathbf{k})|^2 \propto \frac{1}{\widetilde{Q}_s(\mathbf{k})}$$

Decorrelation at the retina

Spatial correlations of natural images fall off with f^{-2}:

$$\widetilde{Q}_s(\mathbf{k}) \propto \frac{1}{|\mathbf{k}|^2 + k_0^2}$$

and the optical filter of the eye introduces (crudely) a low-pass term $\propto e^{-\alpha|\mathbf{k}|}$.
So decorrelation requires

$$|\widetilde{D}_s(\mathbf{k})|^2 \propto \frac{|\mathbf{k}|^2 + k_0^2}{e^{-\alpha|\mathbf{k}|}}$$

Decorrelation at the retina

Spatial correlations of natural images fall off with f^{-2}:

$$\widetilde{Q}_s(\mathbf{k}) \propto \frac{1}{|\mathbf{k}|^2 + k_0^2}$$

and the optical filter of the eye introduces (crudely) a low-pass term $\propto e^{-\alpha|\mathbf{k}|}$.
So decorrelation requires

$$|\widetilde{D}_s(\mathbf{k})|^2 \propto \frac{|\mathbf{k}|^2 + k_0^2}{e^{-\alpha|\mathbf{k}|}}$$

But: not all input is signal.

Decorrelation at the retina

Spatial correlations of natural images fall off with f^{-2}:

$$\widetilde{Q}_s(\mathbf{k}) \propto \frac{1}{|\mathbf{k}|^2 + k_0^2}$$

and the optical filter of the eye introduces (crudely) a low-pass term $\propto e^{-\alpha|\mathbf{k}|}$.
So decorrelation requires

$$|\widetilde{D}_s(\mathbf{k})|^2 \propto \frac{|\mathbf{k}|^2 + k_0^2}{e^{-\alpha|\mathbf{k}|}}$$

But: not all input is signal.
Photodetection introduces noise. Therefore, cascade linear filters:

$$\mathbf{s} + \boldsymbol{\eta} \xrightarrow[D_\eta]{} \hat{\mathbf{s}} \xrightarrow[D_s]{} \mathbf{r}$$

with

$$\widetilde{D}_\eta(\mathbf{k}) = \frac{\widetilde{Q}_s(\mathbf{k})}{\widetilde{Q}_s(\mathbf{k}) + \widetilde{Q}_\eta(\mathbf{k})} \qquad \text{(Wiener filter)}$$

Decorrelation at the retina

Spatial correlations of natural images fall off with f^{-2}:

$$\widetilde{Q}_s(\mathbf{k}) \propto \frac{1}{|\mathbf{k}|^2 + k_0^2}$$

and the optical filter of the eye introduces (crudely) a low-pass term $\propto e^{-\alpha|\mathbf{k}|}$.
So decorrelation requires

$$|\widetilde{D}_s(\mathbf{k})|^2 \propto \frac{|\mathbf{k}|^2 + k_0^2}{e^{-\alpha|\mathbf{k}|}}$$

But: not all input is signal.
Photodetection introduces noise. Therefore, cascade linear filters:

$$\mathbf{s} + \boldsymbol{\eta} \xrightarrow[D_\eta]{} \hat{\mathbf{s}} \xrightarrow[D_s]{} \mathbf{r}$$

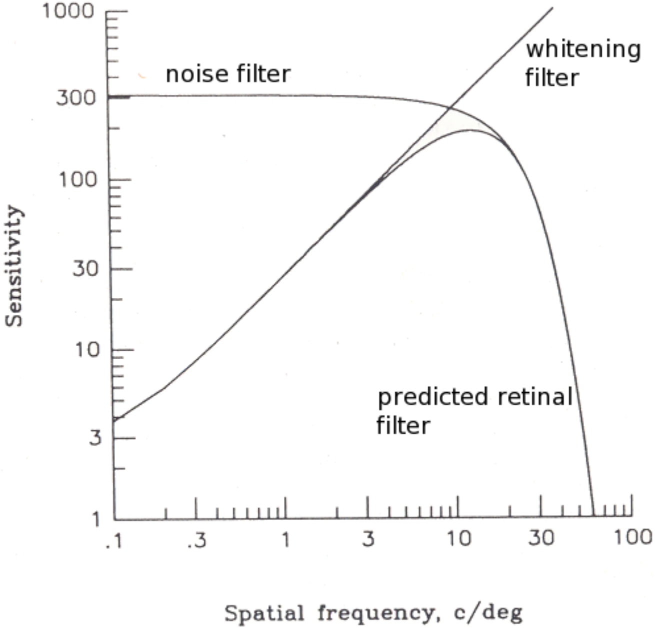
with

$$\widetilde{D}_\eta(\mathbf{k}) = \frac{\widetilde{Q}_s(\mathbf{k})}{\widetilde{Q}_s(\mathbf{k}) + \widetilde{Q}_\eta(\mathbf{k})} \qquad \text{(Wiener filter)}$$

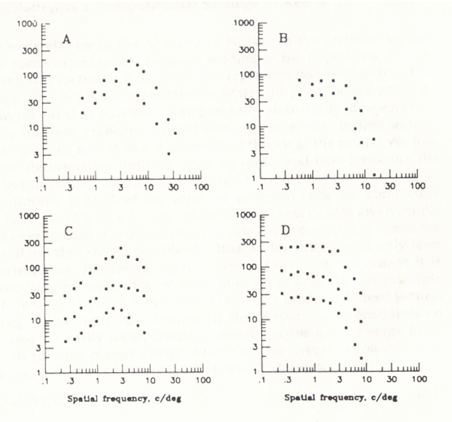
Thus the combined RGC filter is predicted to be:

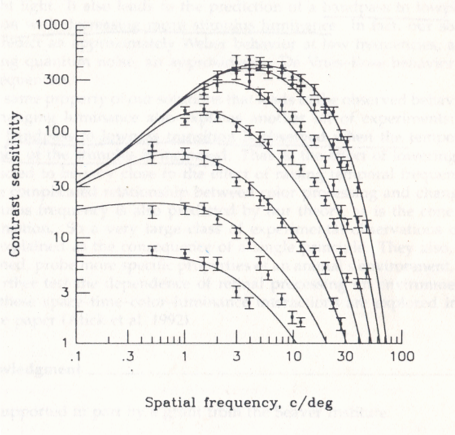
$$|\widetilde{D}_s(\mathbf{k})|\widetilde{D}_\eta(\mathbf{k}) \propto \frac{\sqrt{\widetilde{Q}_s(\mathbf{k})}}{\widetilde{Q}_s(\mathbf{k}) + \widetilde{Q}_\eta(\mathbf{k})}$$

Decorrelation at the retina



Decorrelation at the retina





Related ideas

- efficient channel utilisation
- output entropy maximisation
- variance equalisation
- redundancy reduction
- decorrelation
- discovery of independent projections or components