Statistical Physics of the Mutual Information

Neri Merhav

EE Department

Partly joint work with D. Guo (Northwestern U.) and S. Shamai.

Physics Colloquium, February 19, 2009

General Background

Relations between information theory and statistical physics:

- The maximum entropy principle: Jaynes, Shore & Johnson, Burg, ...
- Physics of information: Landauer, Bennet, Maroney, Plenio & Vitelli, ...
- Large deviations theory: Ellis, Oono, McAllester, ...
- Random matrix theory: Wigner, Balian, Foschini, Telatar, Tse, Hanly, Shamai, Verdú, Tulino, ...
- Coding and spin glasses: Sourlas, Kabashima, Saad, Kanter, Mézard, Montanari, Nishimori, Tanaka, ...

Physical insights and analysis tools are ‘imported’ to IT (and vice versa).

In This Talk We:

- Briefly review basic background in Information Theory.
- Explore relations between information measures and free energy.
- Present mutual information calculation as equilibrium between systems.
- Provide some background in estimation theory.
- Relate mutual information to estimation error from a physics viewpoint.
- Show examples where this error is analyzable via statistical physics.

Background in Information Theory

Source Coding – Data Compression

An information source generates random bits $S_1, S_2, \ldots, S_N$ with

Pr$\{S_i = 1\} = q$.

Q: How much can we compress and still reconstruct perfectly?

A: Shannon's Lossless Source Coding Theorem: For large N, $(S_1, S_2, \ldots, S_N)$ can best be compressed to $\sim N h_2(q)$ bits, where:

$$h_2(q) = -q \log_2 q - (1-q) \log_2 (1-q).$$

Many practical algorithms asymptotically achieve $h_2(q)$.

Q: Can we compress further if we allow a bit error rate D?

A: Yes, we can reduce from $h_2(q)$ to the rate–distortion function:

$$R(D) = h_2(q) - h_2(D).$$

Backgd in Info Theory (Cont'd) – Channel Coding

Suppose we have to transmit a message m of k bits over a noisy channel, which flips each transmitted bit with probability p.

Reliable transmission – only if m is encoded, i.e., mapped (sophisticatedly) into a codeword $\boldsymbol{x}(m)$ of $n > k$ bits before transmission.

$$R = \frac{k}{n} = \text{coding rate}.$$

Let $\boldsymbol{y} = (y_1, \ldots, y_n)$ be the received binary channel output sequence. Optimum decoder for minimum decoding error probability = Maximum Aposteriori Probability (MAP):

$$\hat{m} = \text{arg max}_m P(m|\boldsymbol{y}) = \text{arg max}_m [P(m)P(\boldsymbol{y}|\boldsymbol{x}(m))].$$

Backgd in Info Theory – Channel Coding (Cont'd)

If $P(m) = 2^{-k}$ for all m, MAP decoding = Maximum Likelihood (ML) decoding:

$$\hat{m} = \text{arg max}_m P(\boldsymbol{y}|\boldsymbol{x}(m)).$$

Channel capacity, $C \triangleq$ max R s.t. $\exists$ encoder & decoder with $\lim_{n\to\infty} P_e = 0$.

Q: What is C for this bit–flipping channel?

A: Shannon's Channel Coding Theorem:

$$C = 1 - h_2(p) = 1 + p\log p + (1-p)\log(1-p).$$

Backgd in Info Theory – Channel Coding (Cont’d)

- $\exists$ good codes with $R \approx C$: normally proved by random coding:
  $\boldsymbol{x}(1), \boldsymbol{x}(2), \ldots, \boldsymbol{x}(2^k)$ are selected independently at random. ‘Most’ codes are good – except those that we can think of...
- Ensemble of codes – ensembles that govern large systems – natural relation to statistical mechanics: the code randomness is quenched.
- Mainstream efforts in IT research: seeking good codes with $R \approx C$ & low complexity:
  Low complexity $\Longleftrightarrow$ structure $\Longleftrightarrow$ Low randomness $\Longleftrightarrow$ bad performance.
- Turbo/LDPC codes – good compromise.

Bckgd in IT (Cont'd) – Joint Source–Channel Coding

Consider a Ber(q) source and a bit–flipping channel with parameter p.

A joint source–channel code maps $\boldsymbol{s} = (s_1, \ldots, s_N)$ to a channel input $\boldsymbol{x}(\boldsymbol{s})$ of length $n = \lambda N$. Reliable communication $\iff$ $h_2(q) < \lambda C$.

The decoder estimates $\boldsymbol{u}$ from $\boldsymbol{y} = (y_1, \ldots, y_n)$:

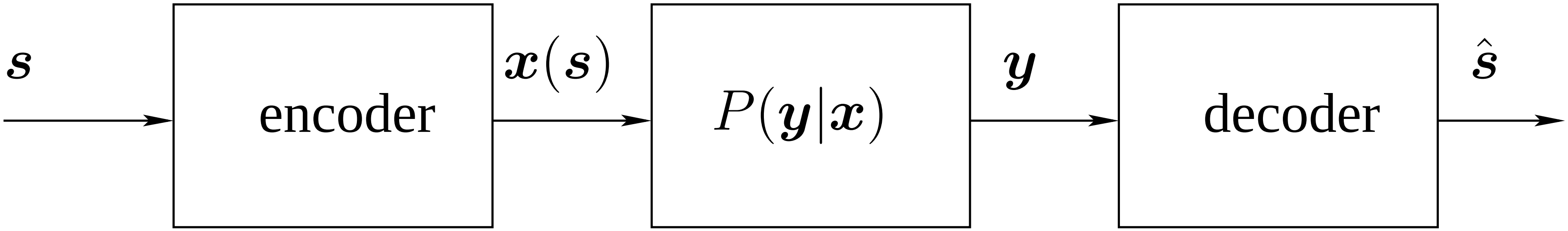
Word MAP decoder $\iff$ min. word error probability

$$\hat{s} = \text{arg max}_{\boldsymbol{s}} P(\boldsymbol{s}|\boldsymbol{y}) = \text{arg max}_{\boldsymbol{s}} [P(\boldsymbol{s}) P(\boldsymbol{y}|\boldsymbol{x}(\boldsymbol{s}))].$$

Bit MAP decoder $\iff$ min. bit error probability:

$$\hat{s}_i = \text{arg max}_{s} P(s_i = s|\boldsymbol{y}) = \text{arg max}_{\boldsymbol{s}} \sum_{\boldsymbol{s}:\ s_i = s} P(\boldsymbol{s}) P(\boldsymbol{y}|\boldsymbol{x}(\boldsymbol{s})).$$

The posterior $P(\boldsymbol{s}|\boldsymbol{y})$ plays a key role.


$\boldsymbol{s}$
encoder
$\boldsymbol{x}(\boldsymbol{s})$
$P(\boldsymbol{y}|\boldsymbol{x})$
$\boldsymbol{y}$
decoder
$\hat{\boldsymbol{s}}$

Background in Information Theory (Cont'd)

A key notion in IT is the mutual information:

Let $(U, V) \sim P(u, v)$:

$$I(U;V) \equiv \left\langle \log \frac{P(U,V)}{P(U)P(V)} \right\rangle = H(U) + H(V) - H(U,V)$$

where

$$H(U) = -\langle \log P(U) \rangle, \quad H(V) = -\langle \log P(V) \rangle, \quad H(U,V) = -\langle \log P(U,V) \rangle.$$

$I(U, V)$ – statistical dependence between U and V.

Other forms:

$$I(U;V) = H(U) - H(U|V) = H(V) - H(V|U)$$

where $H(U|V) = -\langle \log P(U|V) \rangle$.

The Second Law and the Data Processing Thm

A very fundamental inequality in IT: data processing theorem (DPT):

$$A \to B \to C \text{ Markov chain } \implies I(A;B) \geq I(A;C).$$

Virtually, in the proof of every negative result (converse theorem) in IT, the DPT is used. Equivalent to Gibbs' inequality, which can be represented as:

$$\text{avg work of 'abrupt' force} \implies \langle W \rangle \geq \Delta F \impliedby \text{free energy increase}$$

relating coded comm. systems with thermodynamical processes:

- Suboptimum commun. system $\iff$ irreversible process.
- Info rate loss $\iff$ dissipated work $\to$ entropy $\uparrow$
- Fundamental limits of IT $\iff$ second law.

Mutual Information

We will be interested in

$$I(\boldsymbol{X};\boldsymbol{Y}) \text{ – pure channel coding}$$

or

$$I(\boldsymbol{S};\boldsymbol{Y}) \text{ – joint source–channel coding.}$$

Measures how much can one learn from $\boldsymbol{Y}$ about $\boldsymbol{X}$ or $\boldsymbol{S}$, resp.
Suppose

$$\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{Z}$$

where $\boldsymbol{Z} =$ noise: independent of $\boldsymbol{X}$ and $\boldsymbol{Z} \sim \prod_i P(z_i)$.

$$I(\boldsymbol{X};\boldsymbol{Y}) = H(\boldsymbol{Y}) - H(\boldsymbol{Y}|\boldsymbol{X}) \qquad \Longleftarrow \text{ second term is easy:}$$

$$H(\boldsymbol{Y}|\boldsymbol{X}) = H(\boldsymbol{Z}) = nH(Z).$$

$H(\boldsymbol{Y})$ is more difficult to handle..

Channel Output Entropy and Free Energy

Suppose $Z_i \sim \mathcal{N}(0, 1/\beta)$. Then, in pure channel coding:

$$\begin{aligned} H(\boldsymbol{Y}) &= -\langle \log P(\boldsymbol{Y}) \rangle \\ &= -\left\langle \log \left[\sum_m P(m) P(\boldsymbol{Y} | \boldsymbol{x}(m)) \right] \right\rangle \\ &= \text{const.} - \left\langle \log \left[\sum_m e^{-\beta \|\boldsymbol{Y} - \boldsymbol{x}(m)\|^2/2} \right] \right\rangle \\ &\equiv \text{const.} - \langle \log Z(\beta | \boldsymbol{Y}) \rangle \end{aligned}$$

Calculation of $\langle \log Z(\beta | \boldsymbol{Y}) \rangle$ – using statistical mechanical methods.

$\boldsymbol{Y}$ & code are quenched.

A Slightly Different Look

Introduce

$$P(\boldsymbol{s}) \propto \exp\{-\beta \mathcal{E}_S(\boldsymbol{s})\}; \quad P(\boldsymbol{y}|\boldsymbol{x}) \propto \exp\{-\beta \mathcal{E}_C(\boldsymbol{x}, \boldsymbol{y})\}.$$

Thus,

$$P(\boldsymbol{s}|\boldsymbol{y}) = \frac{P(\boldsymbol{s})P(\boldsymbol{y}|\boldsymbol{x}(\boldsymbol{s}))}{\sum_{\boldsymbol{s}'} P(\boldsymbol{s}')P(\boldsymbol{y}|\boldsymbol{x}(\boldsymbol{s}'))} = \frac{\exp\{-\beta[\mathcal{E}_S(\boldsymbol{s}) + \mathcal{E}_C(\boldsymbol{x}(\boldsymbol{s}), \boldsymbol{y})]\}}{\sum_{\boldsymbol{s}'} \exp\{-\beta[\mathcal{E}_S(\boldsymbol{s}') + \mathcal{E}_C(\boldsymbol{x}(\boldsymbol{s}'), \boldsymbol{y})]\}}$$

where

$$Z(\beta|\boldsymbol{y}) \equiv \text{denominator}$$

$\Longrightarrow$ partition function of a system in equilibrium between source and channel at "temperature" $T = 1/\beta$.

$$I(\boldsymbol{S};\boldsymbol{Y}) = H(\boldsymbol{S}) - H(\boldsymbol{S}|\boldsymbol{Y})$$

where $H(\boldsymbol{S})$ = entropy of $Z_S(\beta) = \sum_{\boldsymbol{s}} e^{-\beta \mathcal{E}_S(\boldsymbol{s})}$ and $H(\boldsymbol{S}|\boldsymbol{Y})$ = entropy of $Z(\beta|\boldsymbol{Y})$.

Source–Channel Equilibrium

Let

$$\Sigma_S(\epsilon_1) \equiv \frac{1}{N} \log \left[\#\text{ of } \{\boldsymbol{s}\}\text{: } \mathcal{E}_S(\boldsymbol{s}) \approx N\epsilon_1\right].$$

For a randomly selected code, let

$$\text{Pr}\{\mathcal{E}_C(\boldsymbol{X}, \boldsymbol{y}) \approx n\epsilon_2\} = e^{n\phi_n(\epsilon_2|\boldsymbol{y})}.$$

In many cases, ϕ_n converges and it is self-averaging:

$$\phi_n(\epsilon_2|\boldsymbol{Y}) \rightarrow \phi(\epsilon_2).$$

Finally, let

$$\Sigma(\epsilon|\boldsymbol{y}) \equiv \frac{1}{N+n} \log \left[\#\text{ of } \{\boldsymbol{s}\}\text{: } \mathcal{E}_S(\boldsymbol{s}) + \mathcal{E}_C(\boldsymbol{x}(\boldsymbol{s}), \boldsymbol{y}) \approx (N+n)\epsilon\right].$$

Source–Channel Equilibrium (Cont'd)

For the typical code and for

$$(N+n)\epsilon = N\epsilon_1 + n\epsilon_2 \implies (1+\lambda)\epsilon = \epsilon_1 + \lambda\epsilon_2,$$

$$\begin{aligned}
e^{(N+n)\Sigma(\epsilon|\boldsymbol{Y})} &= \sum_{\epsilon_1} e^{N\Sigma_S(\epsilon_1)} \cdot \mathsf{Pr}\{\mathcal{E}_C(\boldsymbol{X},\boldsymbol{y}) \approx n\epsilon_2\} \\
&\approx \sum_{\epsilon_1} e^{N\Sigma_S(\epsilon_1)} \cdot \exp\left\{n\phi\left(\frac{(1+\lambda)\epsilon-\epsilon_1}{1+\lambda}\right)\right\} \\
&\approx \exp\left\{(N+n)\max_{\epsilon_1}\left[\frac{\Sigma_S(\epsilon_1)}{1+\lambda} + \frac{\lambda}{1+\lambda}\cdot\phi\left(\frac{(1+\lambda)\epsilon-\epsilon_1}{1+\lambda}\right)\right]\right\}
\end{aligned}$$

sum & max over ϵ_1: in the range where $[\cdots] > 0$.

Mutual Info via Source–Channel Equilibrium

$$\Sigma(\epsilon|\boldsymbol{Y}) \quad = \quad \frac{\Sigma_S(\epsilon_1^*)}{1+\lambda} + \frac{\lambda}{1+\lambda} \cdot \phi\left(\frac{(1+\lambda)\epsilon - \epsilon_1^*}{1+\lambda}\right)$$

Let $\epsilon = \epsilon^*$ maximize $\Sigma(\epsilon|\boldsymbol{Y}) - \beta\epsilon$:

- Large β: $\Sigma(\epsilon^*|\boldsymbol{Y}) = \bar{H}(\boldsymbol{S}|\boldsymbol{Y}) = 0 \ \rightarrow$ glassy/ferro ϕ; unreliable comm.
- Small β: $\Sigma(\epsilon^*|\boldsymbol{Y}) = \bar{H}(\boldsymbol{S}|\boldsymbol{Y}) > 0 \ \rightarrow$ disordered ϕ; unreliable comm.

$$\begin{aligned} H(\boldsymbol{S}|\boldsymbol{Y}) \quad &\approx \quad (N+n)\Sigma(\epsilon|\boldsymbol{Y}) = N\Sigma_S(\epsilon_1^*) + n\phi\left(\frac{(1+\lambda)\epsilon - \epsilon_1^*}{1+\lambda}\right) \\ &\approx \quad H(\boldsymbol{S}) + n\phi(\epsilon_2^*) \end{aligned}$$

and so,

$$\lim_{n\rightarrow\infty} \frac{I(\boldsymbol{S};\boldsymbol{Y})}{n} = -\phi(\epsilon_2^*)$$

$I(\boldsymbol{S};\boldsymbol{Y})$ via Source–Channel Equilibrium (Cont'd)

What is ϵ_2^*? Share of $\mathcal{E}_C$ per–particle at inv. temp. β. $\Longrightarrow$ solves the eqn: $\beta = \phi'(\epsilon_2)$. If the codevectors $\sim \mu(\boldsymbol{x})$:

$$\epsilon_2^* = \lim_{n\to\infty} \frac{1}{n} \left\langle \mathcal{E}_C(\boldsymbol{X},\boldsymbol{Y}) \right\rangle_{\mu \times P_{X\to Y}} .$$

Normalized mutual info $=$ exponential rate of the prob. that $\boldsymbol{X}' \perp \boldsymbol{Y}$ yields $\mathcal{E}_C(\boldsymbol{X}',\boldsymbol{Y}) \approx \langle \mathcal{E}_C(\boldsymbol{X},\boldsymbol{Y}) \rangle$, where $(\boldsymbol{X},\boldsymbol{Y})$ are related via the channel.

Gaussian Example: $\mathcal{E}_C(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2$. $\mathcal{E}_C(\boldsymbol{X},\boldsymbol{Y})$ is typically $n/(2\beta)$. If $\boldsymbol{X} \sim$ Surf$(\sqrt{n\sigma^2})$,

$$\text{Pr}\left\{\mathcal{E}_C(\boldsymbol{X}',\boldsymbol{Y}) \approx \frac{n}{2\beta}\right\} \sim e^{-nC}$$

where

$$C = \frac{1}{2}\log(1+\beta\sigma^2),$$

the capacity of the Gaussian channel with input power σ^2.

Signal Estimation – Background

We said that $I(\boldsymbol{X};\boldsymbol{Y})$ tells how much can we learn from $\boldsymbol{Y}$ about $\boldsymbol{X}$, e.g., $I(\boldsymbol{X};\boldsymbol{Y}) = H(\boldsymbol{X}) - H(\boldsymbol{X}|\boldsymbol{Y})$ = reduction in uncertainty of $\boldsymbol{X}$ as $\boldsymbol{Y}$ becomes available.

Can we estimate $\boldsymbol{X}$ better for I large?

First, a word of background in estimation theory:

An estimator is any $\hat{\boldsymbol{X}} = f(\boldsymbol{Y})$. We want $\hat{\boldsymbol{X}}$ as 'close' as possible to $\boldsymbol{X}$.

$$\text{mean square error} = \left\langle \|\boldsymbol{X} - \hat{\boldsymbol{X}}\|^2 \right\rangle = \left\langle \|\boldsymbol{X} - f(\boldsymbol{Y})\|^2 \right\rangle .$$

A fundamental result: minimum mean square error (MMSE) $=$ conditional mean:

$$\boldsymbol{X}^* = f^*(\boldsymbol{y}) = \langle \boldsymbol{X} \rangle_{\boldsymbol{Y}=\boldsymbol{y}} \equiv \int \mathrm{d}\boldsymbol{x} \cdot \boldsymbol{x} P(\boldsymbol{x}|\boldsymbol{y}).$$

Normally, difficult both to apply $\boldsymbol{X}^*$ and to assess performance.

The I–MMSE Relation

[Guo–Shamai–Verdú 2005]: for $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{Z}$, $\boldsymbol{Z} \sim \mathcal{N}(0, I \cdot 1/\beta)$, regardless of $P(\boldsymbol{x})$:

$$\text{mmse}(\boldsymbol{X}|\boldsymbol{Y}) = 2 \cdot \frac{\text{d}}{\text{d}\beta} I(\boldsymbol{X};\boldsymbol{Y}),$$

where $\text{mmse}(\boldsymbol{X}|\boldsymbol{Y}) \equiv \langle \|\boldsymbol{X} - f^*(\boldsymbol{Y})\|^2 \rangle$.

Example: If $\boldsymbol{X} \sim \mathcal{N}(0, \sigma^2 I)$,

$$\frac{I(\boldsymbol{X};\boldsymbol{Y})}{n} = \frac{1}{2} \log(1 + \beta\sigma^2)$$

$$\Longrightarrow \quad \frac{\text{mmse}(\boldsymbol{X}|\boldsymbol{Y})}{n} = \frac{\sigma^2}{1 + \beta\sigma^2}.$$

MMSE – now calculated using stat–mech via the mutual info and I–MMSE relation.

Analogue stat–mech system exhibits ϕ transitions $\longrightarrow$ irregularities in MMSE.

Statistical Physics of the MMSE

$$
\begin{aligned}
I(\boldsymbol{X};\boldsymbol{Y}) &= \left\langle \log \frac{P(\boldsymbol{X}|\boldsymbol{Y})}{P(\boldsymbol{X})} \right\rangle_\beta \\
&= \left\langle \log \frac{\exp\{-\beta\|\boldsymbol{Y}-\boldsymbol{X}\|^2/2\}}{\sum_{\boldsymbol{x}} P(\boldsymbol{x})\exp\{-\beta\|\boldsymbol{Y}-\boldsymbol{x}\|^2/2\}} \right\rangle_\beta \\
&= -\frac{n}{2} - \langle \log Z(\beta|\boldsymbol{Y})\rangle_\beta
\end{aligned}
$$

and so,

$$
\text{mmse}(\boldsymbol{X}|\boldsymbol{Y}) = 2\cdot\frac{\mathrm{d}I(\boldsymbol{X};\boldsymbol{Y})}{\mathrm{d}\beta} = -2\frac{\partial}{\partial\beta}\langle \log Z(\beta|\boldsymbol{Y})\rangle_\beta.
$$

Similar to internal energy, but here also $\langle\cdot\rangle_\beta$ depends on β.

Statistical Physics of the MMSE (Cont'd)

A more detailed derivation yields:

$$\text{mmse}(\boldsymbol{X}|\boldsymbol{Y}) = \frac{n}{\beta} + \text{Cov}\{\|\boldsymbol{Y}-\boldsymbol{X}\|^2, \log Z(\beta|\boldsymbol{Y})\}$$

- The term $n/\beta \sim$ energy equipartition theorem.
- Covariance term – dependence of $\langle\cdot\rangle_\beta$ on β.

Statistical Physics of the MMSE (Cont'd)

In stat. mech:
$$\Sigma(\beta) = \log Z(\beta) + \beta\langle\mathcal{E}(X)\rangle$$
$$= \log Z(\beta) - \beta\frac{\mathrm{d}\log Z(\beta)}{\mathrm{d}\beta} \quad \Longleftarrow \text{ diff. eq.}$$

$$\log Z(\beta) = -\beta E_0 + \beta\cdot\int_{\beta}^{\infty}\frac{\mathrm{d}\hat{\beta}\cdot\Sigma(\hat{\beta})}{\hat{\beta}^2}; \quad E_0 = \text{ground–state energy}$$

$$\Longrightarrow\ E = -\frac{\mathrm{d}\log Z(\beta)}{\mathrm{d}\beta} = \left[E_0 - \int_{\beta}^{\infty}\frac{\mathrm{d}\hat{\beta}\cdot\Sigma(\hat{\beta})}{\hat{\beta}^2}\right] + \frac{\Sigma(\beta)}{\beta}$$

Similarly for $\langle\log Z(\beta|\boldsymbol{Y})\rangle_{\beta}$ except that

$$\Sigma(\beta) \Longleftarrow \frac{\beta}{2}\mathrm{Cov}\{\|\boldsymbol{Y}-\boldsymbol{X}\|^2, \log Z(\beta|\boldsymbol{Y})\} - I(\boldsymbol{X};\boldsymbol{Y})$$

$$E_0 \Longleftarrow \frac{1}{2}\left\langle\min_{\boldsymbol{x}}\|\boldsymbol{Y}-\boldsymbol{x}\|^2\right\rangle_{\beta}.$$

Examples

Example 1 – Random Codebook on a Sphere Surface

$$\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{Z}; \qquad \boldsymbol{X} \sim \text{Unif}\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}, \;\; M = e^{nR}$$

Codewords: randomly drawn independently uniformly on Surf$(\sqrt{n\sigma^2})$.

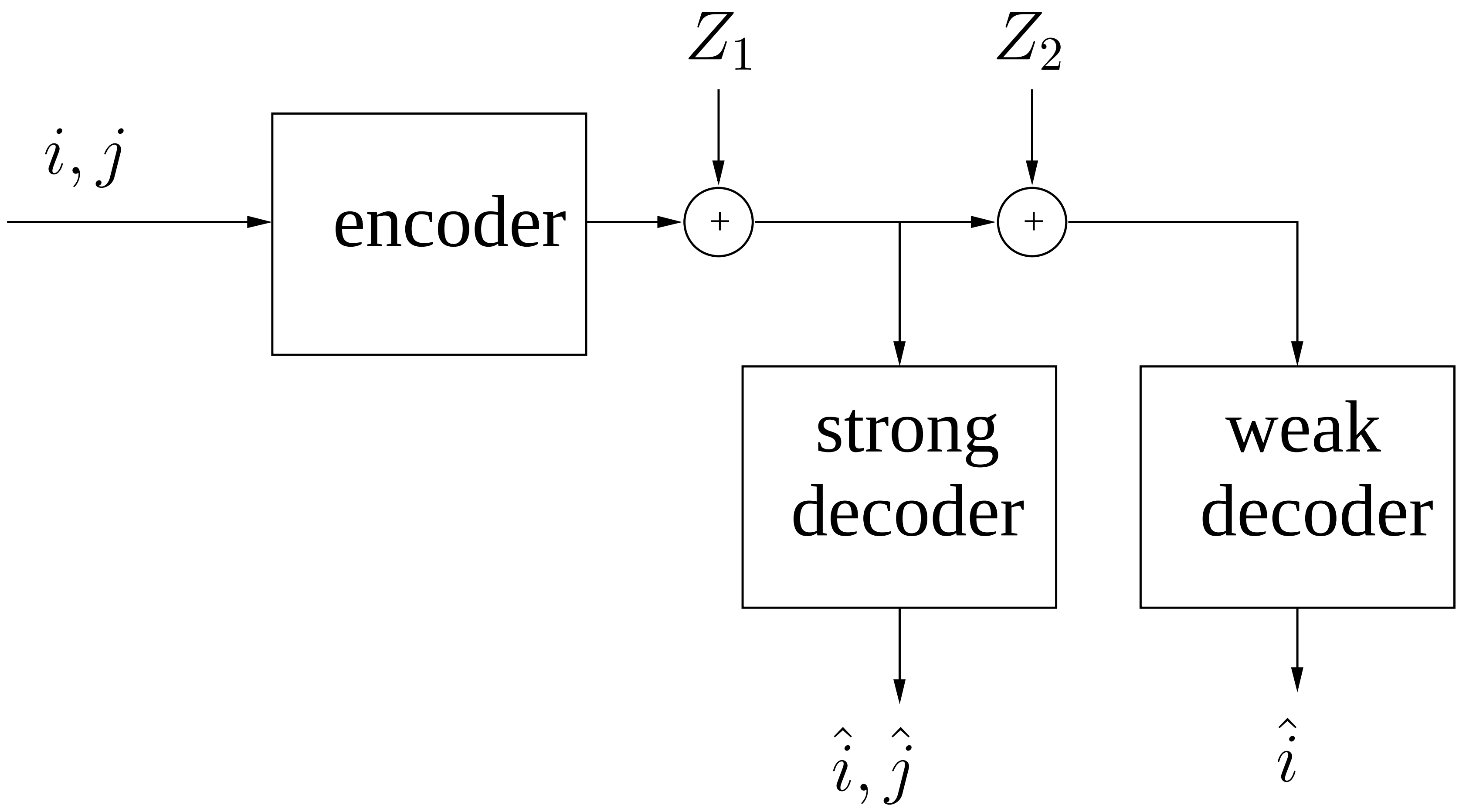
$$\lim_{n\to\infty} \frac{\langle I(\boldsymbol{X};\boldsymbol{Y})\rangle}{n} = \begin{cases} \frac{1}{2}\log(1+\beta\sigma^2) & \beta < \beta_R \\ R & \beta \geq \beta_R \end{cases}$$

where β_R is the solution to the eqn $R = \frac{1}{2}\log(1+\beta\sigma^2)$. Thus,

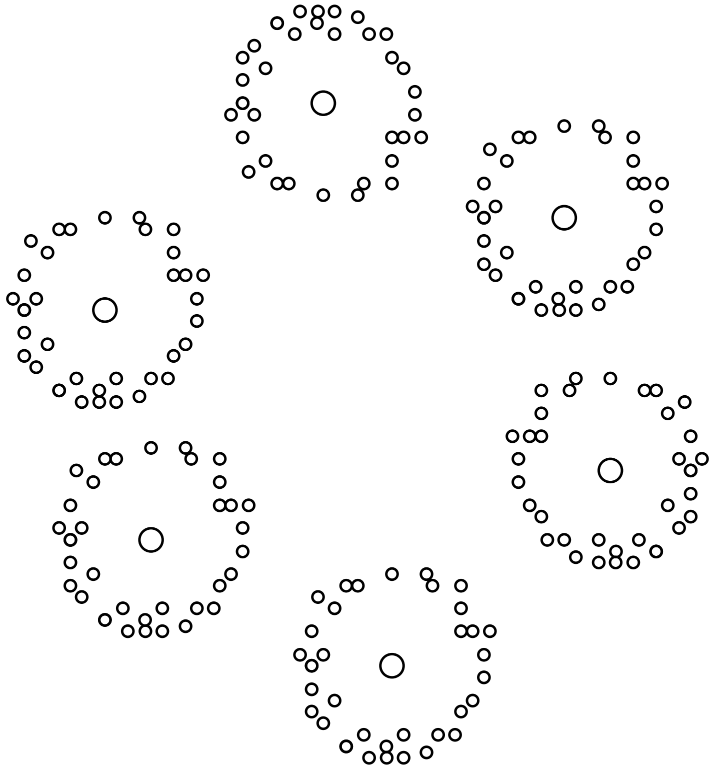
$$\lim_{n\to\infty} \frac{\text{mmse}(\boldsymbol{X}|\boldsymbol{Y})}{n} = \begin{cases} \frac{\sigma^2}{1+\beta\sigma^2} & \beta < \beta_R \\ 0 & \beta \geq \beta_R \end{cases}$$

A 1st–order ϕ transition in MMSE: At high temp. behaves as if $\boldsymbol{X}$ was Gaussian and at $\beta = \beta_R$ jumps to zero!

Example 2 − broadcast channel



Example 2 − broadcast channel (cont'd)



$i =$ index of 'cloud' center

$j =$ index of codeword within cloud

Example 2 (Cont'd)

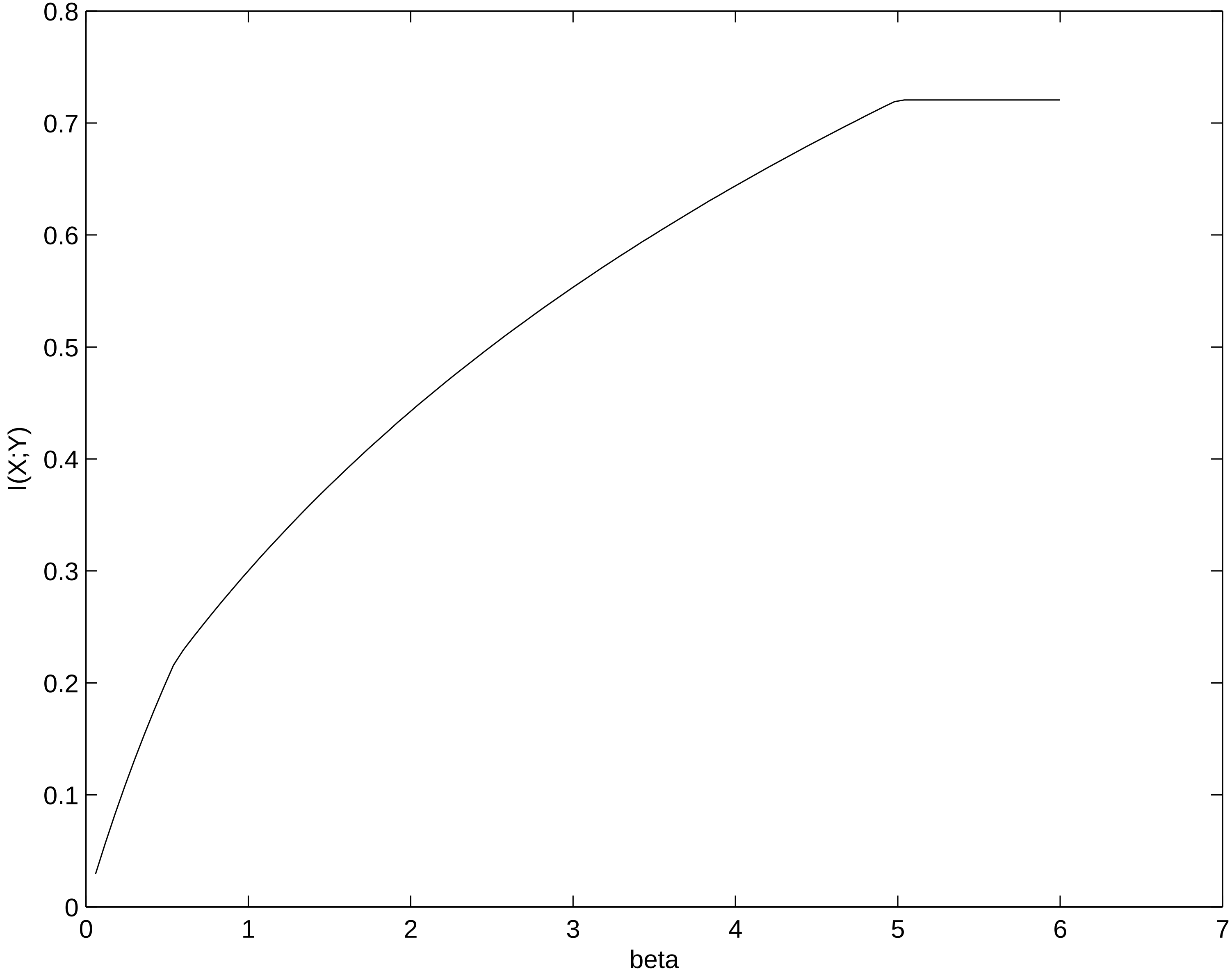
$$\boldsymbol{X} \sim \text{Unif}\{\boldsymbol{x}_{i,j}\}, \;\; 1 \leq i \leq e^{nR_1}; \;\; 1 \leq j \leq e^{nR_2}$$

$$\boldsymbol{x}_{i,j} = \alpha \boldsymbol{u}_i + \sqrt{1-\alpha^2}\boldsymbol{v}_{i,j}$$

where $\{\boldsymbol{u}_i\}$ and $\{\boldsymbol{v}_{i,j}\}$ are drawn independently on Surf$(\sqrt{n})$.
Two phase transitions:

$$\lim_{n\to\infty} \frac{\langle I(\boldsymbol{X};\boldsymbol{Y})\rangle}{n} = \begin{cases} \frac{1}{2}\log(1+\beta) & \beta < \beta_1 \\ R_1 + \frac{1}{2}\log[(1+\beta(1-\alpha^2)] & \beta_1 \leq \beta < \beta_2 \\ R = R_1 + R_2 & \beta \geq \beta_2 \end{cases}$$

$$\lim_{n\to\infty} \frac{\text{mmse}(\boldsymbol{X}|\boldsymbol{Y})}{n} = \begin{cases} \frac{1}{1+\beta} & \beta < \beta_1 \\ \frac{1-\alpha^2}{1+\beta(1-\alpha^2)} & \beta_1 \leq \beta < \beta_2 \\ 0 & \beta \geq \beta_2 \end{cases}$$



0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
I(X;Y)
0
1
2
3
4
5
6
7
beta

Examples (Cont'd)

Example 3 – Sparse Signals

$$X_i = S_i U_i, \quad i = 1, \ldots, n$$

where $\boldsymbol{S} = (S_1, \ldots, S_n) \sim P(\boldsymbol{s})$ is binary 0–1; $U_i \sim \mathcal{N}(0, \sigma^2)$ – i.i.d. $\perp$ $\boldsymbol{S}$.

$$\begin{aligned} Z(\beta|\boldsymbol{y}) &= \int_{\mathbb{R}^n} \mathrm{d}\boldsymbol{x} P(\boldsymbol{x}) \exp\{-\beta\|\boldsymbol{y} - \boldsymbol{x}\|^2/2\} \quad \Longleftarrow \quad P(\boldsymbol{x}) = \sum_{\boldsymbol{s}} P(\boldsymbol{s}) P(\boldsymbol{x}|\boldsymbol{s}) \\ &= \sum_{\boldsymbol{s}} P(\boldsymbol{s}) \exp\left\{-\frac{1}{2}\sum_{i=1}^n \text{func}(y_i, s_i, q)\right\} \quad \Longleftarrow \quad q \equiv \beta\sigma^2 \\ &= \text{const.} \times \sum_{\boldsymbol{\mu}} P(\boldsymbol{\mu}) \cdot \exp\left\{\sum_{i=1}^n \mu_i h_i\right\} \end{aligned}$$

$$\mu_i = 1 - 2s_i; \quad h_i = \text{func}(y_i).$$

Sum over $\{\boldsymbol{\mu}\} \equiv \hat{Z}(\beta|\boldsymbol{y})$: "partition function" of spins in a random field $\{h_i\}$.

Example 3 (Cont'd)

Let $P(\boldsymbol{\mu}) \propto \exp\{nf[m(\boldsymbol{\mu})]\}$ where $m(\boldsymbol{\mu}) \equiv \frac{1}{n}\sum_i \mu_i$ and $f[m]$ is 'nice'.

$$\hat{Z}(\beta|\boldsymbol{y}) \propto \sum_{\boldsymbol{\mu}} \exp\left\{n\left[f[m(\boldsymbol{\mu})] + \frac{1}{n}\sum_i \mu_i h_i\right]\right\}$$

$\hat{Z}$ is dominated by configurations with magnetization m^*, solving the zero–derivative equation

$$m = \langle \tanh(f'[m] + H)\rangle$$

where H is a RV pertaining to h_i. $m^* =$ local maximum if:

$$\left\langle \tanh^2(f'[m^*] + H)\right\rangle > 1 - \frac{1}{f''[m^*]}.$$

When this becomes equality (and then reversed), m^* ceases to dominate $\hat{Z}$ (critical point) $\Longrightarrow$ dominant magentization jumps elsewhere.

Example 3 (Cont'd)

Consider the case

$$f[m] = am + \frac{bm^2}{2}$$

$\hat{Z}$ – similar to the random–field Curie–Weiss (RFCW) model.

We analyze the mutual info using stat–mech methods, and then derive the MMSE using the I–MMSE relation:

Defining m_a to be the maximizer of

$$h_2\left(\frac{1+m}{2}\right) + am + \frac{bm^2}{2},$$

$$\text{mmse}(\boldsymbol{X}|\boldsymbol{Y}) = \text{closed-form-expression}(a, b, m_a, m^*, \sigma^2, \beta).$$

Example 3: Discussion

- MMSE depends on m^*: jumps of m^* yield discontinuities in MMSE.
- As m^* jumps, the response of $\boldsymbol{X}^*(\boldsymbol{Y})$ jumps as well.
- In the C–W model: 1st order transition w.r.t. mag. field and 2nd order transition w.r.t. β. Here – a 1st order transition w.r.t. β because dependence on β is via the “magnetic fields” $\{h_i\}$..
- $b = 0$: i.i.d. spins $\Longrightarrow$ no ϕ transitions $\Longrightarrow$ sparsity alone does not cause ϕ transitions.

Conclusion

- The mutual info, which is a fundamental quantity in IT, is a measure of the relevant info given in one RV on the other.
- We demonstrated that mutual info can be assessed from a stat–mech perspective: one approach is via source–channel thermal equilibrium.
- The mutual info is related to the MMSE.
- $\Longrightarrow$ the MMSE is calculated using stat–mech tools.
- Statistical–mech techniques can be used to inspect inherent irregularities in the estimation error, via phase transitions.

MMSE for Example 3

$$\begin{aligned}\overline{\text{mmse}} \quad = \quad & \frac{\sigma^2 q}{2(1+q)^2} + \frac{(1-m_a)\sigma^2}{2}\left[1 - \frac{q(1+q/2)}{(1+q)^2}\right] + \\ & \frac{1+m_a}{2}\Big[\text{Cov}_0\{Y^2, \log[2\cosh(bm^* + a + H)]\} + \\ & \left\langle H' \tanh(bm^* + a + H)\right\rangle_0\Big] + \\ & + \frac{1-m_a}{2}\left[\frac{1}{(1+q)^2}\cdot\text{Cov}_1\{Y^2, \log[2\cosh(bm^* + a + H)]\} + \right. \\ & \left. \left\langle H' \tanh(bm^* + a + H)\right\rangle_1\right]\end{aligned}$$

where $\langle\cdot\rangle_s$ and Cov_s are w.r.t. $Y \sim \mathcal{N}(0, \sigma^2 s + 1/\beta)$, $s = 0, 1$, and

$$H' = -\frac{\sigma^2}{2(1+q)} + \frac{q(q+2)Y^2}{2(1+q)^2}.$$