Neuronal population size and reliable information transmission

Lubomir Kostal
Institute of Physiology CAS, Prague, Czech Republic

Summary

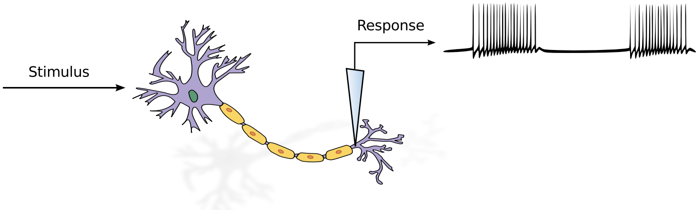
1. Classical 'information' measures: *additivity* (log *p*)
   Fisher (1925), Hartley (1928), Shannon (1948), Savage (1954)
2. Asymptotic *vs.* achievable: finite-size effects might be important!
3. Achievable (*operational*) info may *not* be additive in i.i.d. setup

- Application: 'Critical' size of neural population for reliable information transmission

$\rightarrow$ Ryota Kobayashi (University of Tokyo)
$\rightarrow$ LK & RK, *Phys Rev E (Rapid)* (2019)

Motivation

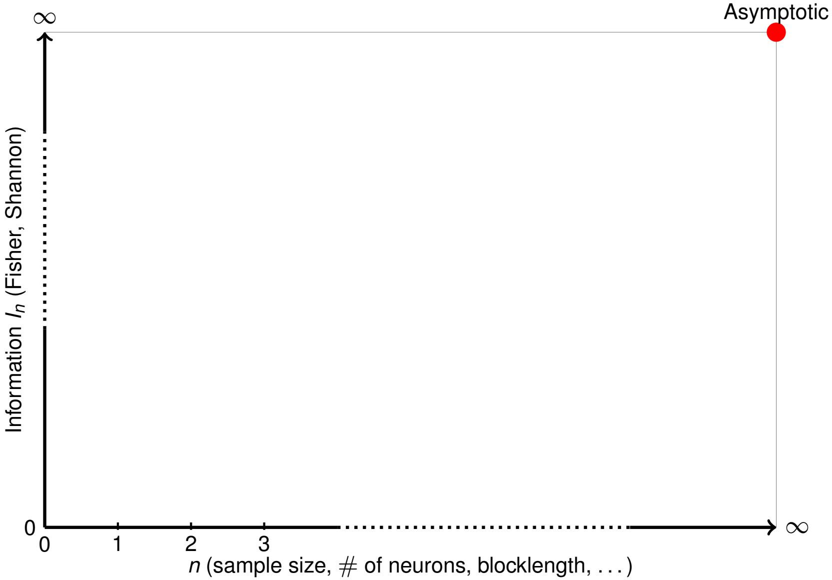
- Neural coding: How neurons (populations) encode and process information about their environment?



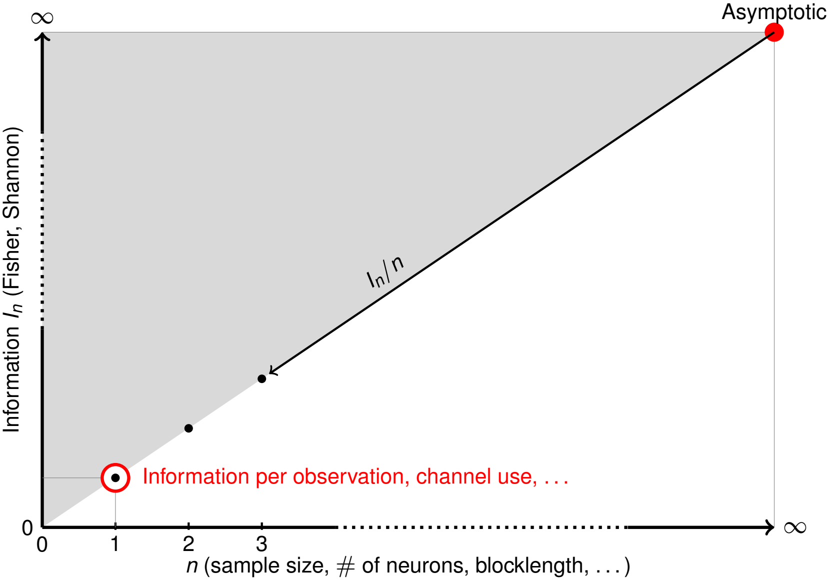
- *Indirect:* degree to which the response reflects the stimulus

1. "How much information?" (stimulus$\rightarrow$response)
   Mutual information (bits)
   MacKay & McCulloch (1952), Stein (1967), Laughlin (1981), Bialek *et al.*, ...
2. Coding precision: the accuracy of stimulus identification
   Fisher information (Cramér-Rao bound):
   Paradiso (1988), Stemmler (1996), Abbott & Dayan (1999), Greenwood *et al.*

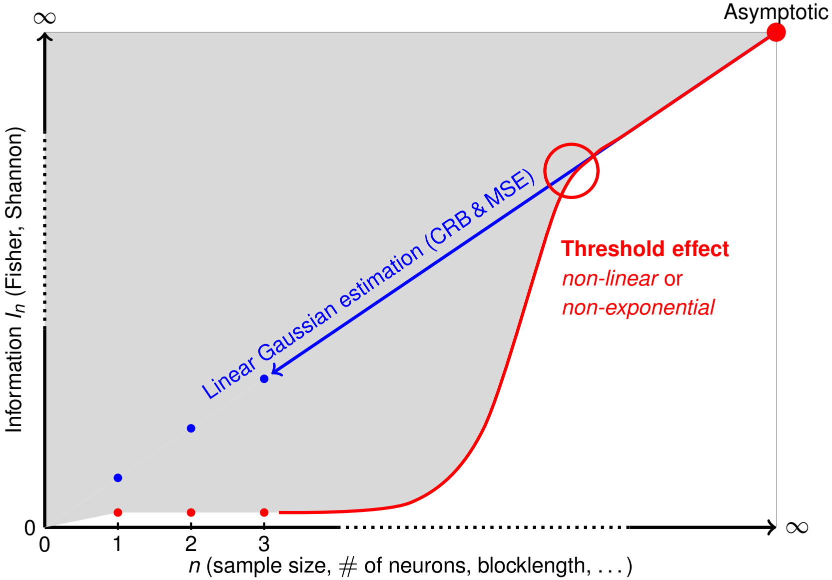
Asymptotic *vs.* non-asymptotic information (*heuristic*)



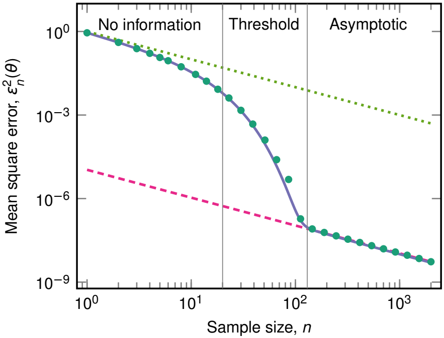
Asymptotic *vs.* non-asymptotic information (*heuristic*)



Asymptotic *vs.* non-asymptotic information (*heuristic*)



Example: threshold effect (toy model)



$$Y = \theta + Z,$$

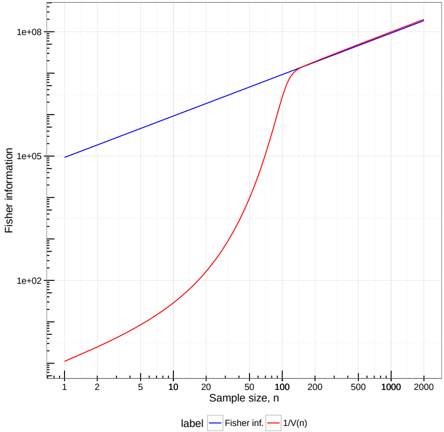
$$Z \sim (1-p)\mathcal{N}(0, \sigma_1^2) + p\mathcal{N}(0, \sigma_2^2),$$

$$\sigma_1 \gg \sigma_2, 0 < p < 1,$$

$$\varepsilon_n^2(\theta) \doteq \frac{(1-p)^n \sigma_1^2}{n} + \\ + \sum_{k=1}^{n} \binom{n}{k} p^k (1-p)^{n-k} \frac{\sigma_2^2}{k},$$

$$\theta = 0, p = 0.1, \sigma_1 = 1, \sigma_2 = 0.001$$

Example: threshold effect (toy model)



? Information theory ?

- Clustering, mixtures, ...
- Non-exponential parametric family, closed-form approx. to optimal MSE

Kostal *et al.*, *J. Neur. Eng.* (2015)

Information theory: methods

- Input (stim. intensity, feature) x, *duration* Δ, r.v. $X \sim \pi(x)$
- Response y, r.v. $Y \sim f(y|X = x)$ *(DT-MC, no feedback)*

- Mutual information and capacity (nat/s)

$$I(X;Y) = \frac{1}{\Delta}\,\mathbb{E}\left[\log\frac{f(Y|X)}{p(Y)}\right], \quad p(y) = \mathbb{E}[f(y|X)],$$

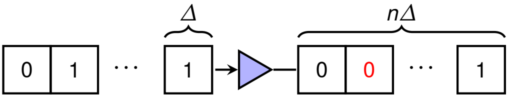
$$C = \sup_{\pi(x)} I(X;Y)$$

- $I(X;Y)$: maximum information that can be communicated *reliably* by neuronal 'model' $f(y|x)$ subject to the input statistics $\pi(x)$

- *Optimal* information decoding $\Rightarrow$ guiding principle (Efficient coding hypothesis Barlow, 1961)

Shannon's theorem

- 'Reliability' $\Rightarrow$ sequence *vs.* per-symbol decoding, 'errors' *(BSC)*



- Information rate (nat/s) assuming m (known) input n-sequences

$$\mathbf{x} = \{x_1, x_2, \ldots, x_n\} \Rightarrow R = \frac{\log m}{n\Delta}$$

- Shannon's theorem (channel coding) $\doteq$ if $R < C$ then $\exists$ a *set* of $\mathbf{x}$'s such that Pr of $\hat{\mathbf{x}} \neq \mathbf{x}$ is *arbitrarily small* ($\Rightarrow$ n increasing!)

- Signal *estimation* vs. *detection*: up to $m \approx e^{\Delta nC}$ 'patterns' decoded reliably for *n large* enough *(NN classifiers)*

Asymptotics *vs.* achievable information rates

- In fact: the (average) probability of decoding error P_e: phase transition at $R = C$ in the 'thermodynamic' limit $n \to \infty$
  (Gallager: $P_e = 0$ for $R < C$, Wolfowitz: $P_e = 1$ for $R > C$)
- Finite-size effects (n): relationship $R \leftrightarrow P_e$?
  Perhaps $R > C$ for some 'reasonable' P_e?
- *Maximal* asymptotic *vs.* achievable rates?

$$R_n \approx nC \quad \propto \log m \qquad (n \to \infty)$$

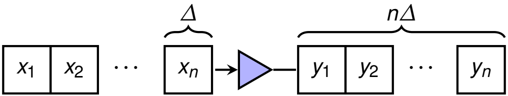
$$R_n = ? \qquad \text{(generaly a function of } P_e, n\text{)}$$

Shannon (1959), Gallager (1962–73), ..., Verdu, Polyanski (2010)

- Non-asymptotics: new relevant parameters
- Price to pay: delays, "complexity": $O(N^2), \ldots$
  (Punekar *et al.*, 2013: non-binary LDPC $\sim O(N)$)

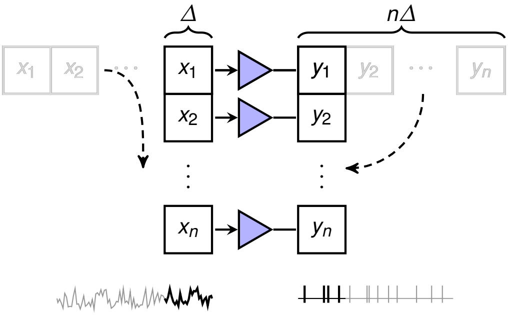
Simple population model

- Single-comp (HH): $I_{\text{syn}}(t) = g_e(t)(E_e - V) + g_i(t)(E_i - V)$
- $g_{e,i}(t) \sim$ OU process (Miura *et al.*, 2007), $\Delta = 50\,\text{ms}$



Simple population model

- Single-comp (HH): $I_{\text{syn}}(t) = g_e(t)(E_e - V) + g_i(t)(E_i - V)$
- $g_{e,i}(t) \sim$ OU process (Miura *et al.*, 2007), $\Delta = 50\,\text{ms}$



$x = \langle g_e \rangle_\Delta \quad y = \#\text{AP}/\Delta$

Useful preliminaries

- Y: spike-count in a time window $\Rightarrow \exists$ max. $\Rightarrow$ discrete & finite
- Let $Y \sim f(y|X = x)$: max. K points of support
- Witsenhausen, 1980 ($\Leftarrow$ Dubin's theorem): capacity is achieved by discrete $\pi(X)$ supported at most K points (finite dimensionality, almost no assumptions on X!)
- Extendable to other convex optimization problems:
- 'Model' *vs.* DMC: applicable bounds on R_n
- Numerical methods: cutting-plane (linear programming), ...

Decoding: maximum likelihood

- Set of m stimulus 'patterns': $\mathbf{x}^{(j)} = \{x_1^{(j)}, \ldots, x_n^{(j)}\}, j = 1, \ldots, m$
- For $\mathbf{X} = \mathbf{x}$ we observe $\mathbf{Y} = \mathbf{y}$, given by $f(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{n} f(y_i|x_i)$
- ML decoder *(optimality?, cf. ME decoding)*:

$$\mathbf{x}^{(d)} : \quad d = \arg\max_j f(\mathbf{y}|\mathbf{x}^{(j)}),$$

- Average probability of decoding error

$$P_e = \sum_{j=1}^{m} \Pr(\mathbf{x} = \mathbf{x}^{(j)}) \int_{\mathcal{E}(j)} f(\mathbf{y}|\mathbf{x}^{(j)})\, \mathrm{d}\mathbf{y}$$

$\mathcal{E}(j)$: set of $\mathbf{y}$ such that ML *fails* for $\mathbf{x}^{(j)}$

- How to obtain the inputs $\mathbf{x}$? *(Assume* $\Pr(\mathbf{x} = \mathbf{x}^{(j)}) = 1/m$.*)*

Lower bound on info rate (*achievable* & general)

- *Ensemble*: generate patterns *i.i.d.* according to some $X \sim \pi(x)$
- Prob. of particular set of m inputs, each of length n:

$$\prod_{j=1}^{m} \pi(\mathbf{x}^{(j)}) = \prod_{j=1}^{m} \pi(x_1^{(j)}) \cdots \pi(x_n^{(j)}) \tag{1}$$

- Use the random-coding bound (Gallager, 1968) & invert
- Optimize over $\pi(x)$ to get a tighter result *(convex)*:

$$R_n \geq \frac{n}{\Delta} E_r^{-1}\left(-\frac{\log P_e}{n} \right),$$

$$E_r(\Delta R) = \max_{0 \leq \rho \leq 1} \left[\max_{\pi(x)} E_0\big(\rho, \pi(x)\big) - \rho \Delta R \right],$$

$$E_0\big(\rho, \pi(x)\big) = -\log \int \left(\int f(y|x)^{1/(1+\rho)} \pi(x)\, \mathrm{d}x \right)^{1+\rho} \mathrm{d}y$$

- Note: optimal $\pi^*(X)$: $R \neq I(\pi^*(X), Y)$; 'good' codes? *(not iid)*

Upper bound on info rate

- Technical assumptions, validity . . . (*BSC*, Polyanski, 2014)
- *Cf.*: sphere-packing and straight-line bounds (Gallager, 1968)
- Strassen, 1962; Tomamichel, 2013, assume $P_e \leq 1/2$

$$R_n \leq nC - \frac{1}{\Delta}\left[\sqrt{nV(P_e)}Q^{-1}(P_e) + \frac{\log n}{2}\right] + O(1)$$

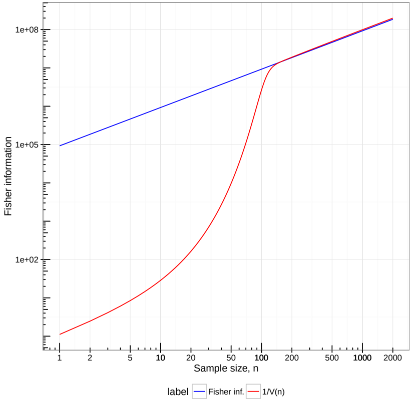
$$V(P_e) = \min_{\pi \in \mathcal{C}}\left[\mathbb{E}\left(\log\frac{f(Y|X)}{p(Y)}\right)^2 - \Delta^2 C^2\right]$$

$Q(z) = \int_z^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\,\mathrm{d}t$, *n.b.* (Cramér-Esseen, 1937, 1945,)
CLT *vs.* AEP (Feinstein, 1958; Wolfowitz, 1961, $nC + O(\sqrt{n})$)

- Note the const. term (*cf.* Feller, 1972; Tyurin, 2010)

Transition to asymptotic regime

- Heuristic: what we expect



- *'no information'* (small n), *'threshold'* (supra-linear in n) and *'asymptotic'* regimes

Transition to asymptotic regime

- *'no information'* (small n), *'threshold'* (supra-linear in n) and *'asymptotic'* regimes
- Critical population size n_c marks the transition towards the asymptotic regime

$$E_0\big(\rho, \pi(x)\big) = -\log \int \left(\int f(y|x)^{1/(1+\rho)} p(x)\,\mathrm{d}x \right)^{1+\rho} \mathrm{d}y,$$

$$R_c = \frac{1}{\Delta} \frac{\mathrm{d}}{\mathrm{d}\rho} \max_{\pi(x)} E_0\big(\rho, p(x)\big)\Big|_{\rho=1},$$

$$n_c = \lceil -(\log P_e)/E_r(\Delta R_c) \rceil$$

Gaussian approximation

- Gaussian approx. (AWGN: $\mathrm{Var}(X) \leq P$)

$$C_G = \frac{1}{\Delta} \ln \left(1 + \frac{P}{\sigma^2}\right)$$

- Effective SNR: $S = P/\sigma^2$
- Let $S = (e^{2\Delta C} - 1)$ (where C is the single neuron capacity), then *since* $C \propto \log(1 + SNR)$

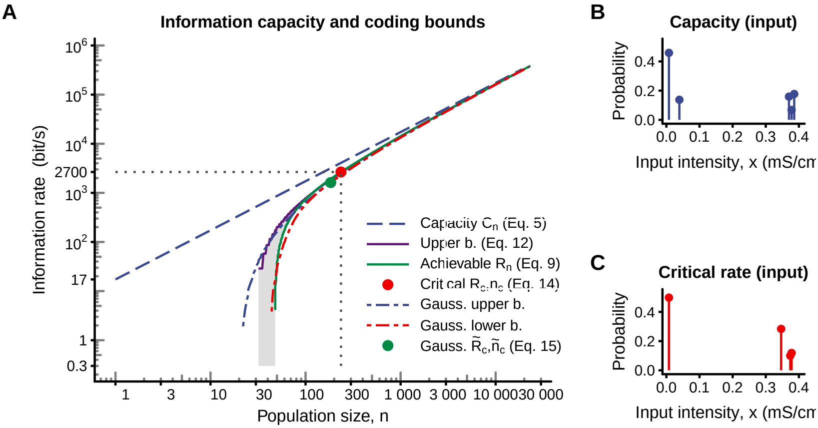
$$\tilde{R}_c = \frac{1}{2\Delta} \log \left(\frac{1}{2} + \frac{S}{4} + \frac{1}{2}\sqrt{1 + \frac{S^2}{4}}\right),$$

$$\tilde{n}_c = \Big\lceil -4\Big[2 + S - \sqrt{4 + S^2} - 4\log 2 + 2 + \log\left(2 - S + \sqrt{4 + S^2}\right)\Big]^{-1} \log P_e \Big\rceil$$

+ complete closed-form for the error exponents

Results

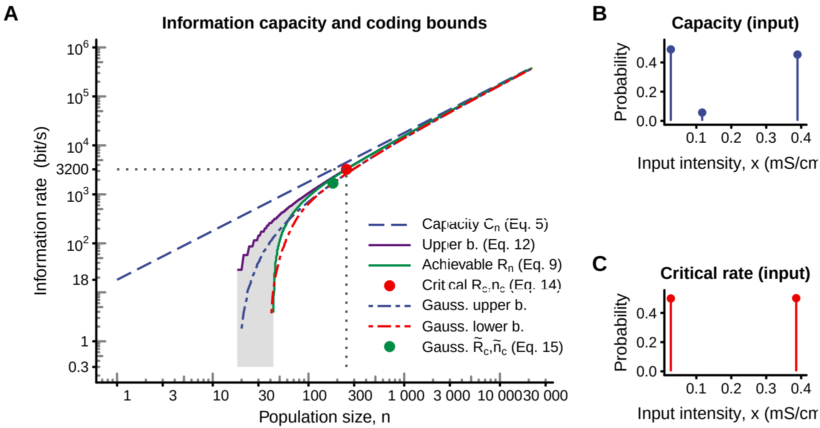
HH model + balanced input (Wehr and Zador, 2003; Berg et al., 2007)

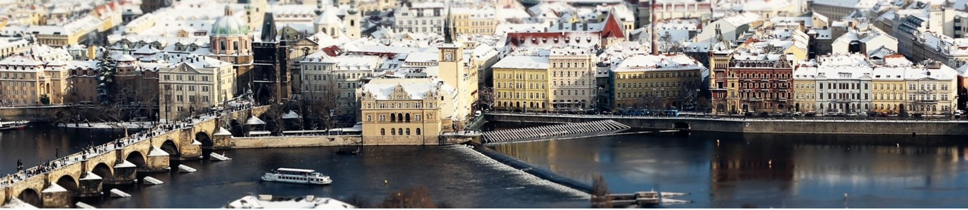


'Critical' rate R_c ($\approx$ bounds eq.) $P_e = 10^{-10}$ (rel. 'noise'), note $I(X_c; Y) \neq R_c$,

Results

Spike response model (Pfister, Toyoizumi, Barber, Gerstner, 2006)

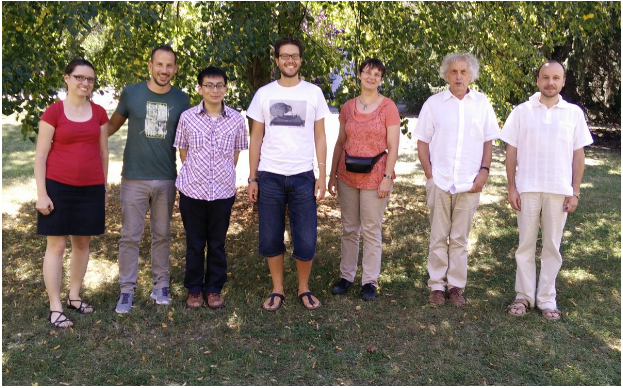




Summary & Outlooks

- Test robustness of bounds w.r.t. models, data, . . . (✔)
- Decoding: assoc. memory in NNs?
  McEliece *et. al*, 1987: Hopfield, $m \approx n/(4 \log n)$
  Jankowski *et. al*, 1996: non-binary, $m \approx n$
  Karbasi *et. al*, 2014; Hillar & Tran, 2018: conv. NN: $m \approx e^{cn}$
- Extensions
  - Rate R_n: upper bound (✘) *vs.* achievability (✔)
  - ‘Pattern’ ensemble optimization, expurgation, convexity?
  - Non-asymptotic optimality: uncoded transmission

Postdoc position: Computational Neuroscience Group

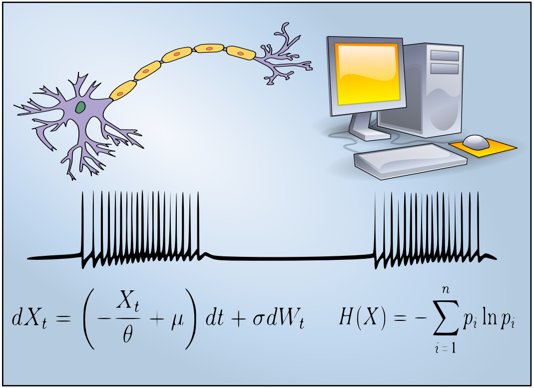


Institute of Physiology, Prague

kostal@biomed.cas.cz

http://comput.biomed.cas.cz

Postdoc position: Computational Neuroscience Group



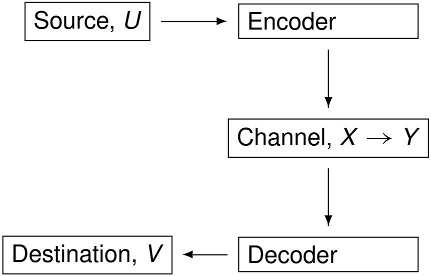
Institute of Physiology, Prague

kostal@biomed.cas.cz

http://comput.biomed.cas.cz

Mutual information: discussion

- Information theory: fundamental limits on the efficacy of: *i) representation* and *ii) reliable communication*



- Representation: compression (*source entropy,* $H(U)$)
- Reliability: probability of channel decoding error, P_e

Mutual information: discussion

- Two types of results: achievability and converse theorems

  1. Converse: if $H(U) > I(X; Y)$, *arbitrarily* small P_e is not possible, no matter the communication setup

     More broadly: Arbitrarily reliable information transfer greater than $I(W; Z)$ is *impossible* between any two random variables W, Z, no matter what "mechanism" connects them.

Mutual information: discussion

- Two types of results: achievability and converse theorems

  1. Converse: if $H(U) > I(X;Y)$, *arbitrarily* small P_e is not possible, no matter the communication setup

     More broadly: Arbitrarily reliable information transfer greater than $I(W;Z)$ is *impossible* between any two random variables W, Z, no matter what "mechanism" connects them.

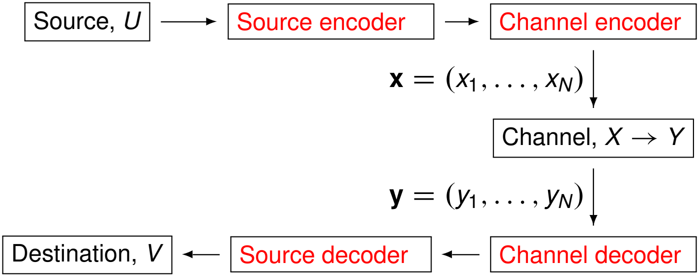
  2. Achievability: if $H(U) < I(X;Y)$, *arbitrarily* small P_e is possible under the separation setup (discussion)

- Note: *i)* arbitrarily reliable, *ii)* neither 1. nor 2. says *what is the actual information transfer* in the system if we calculate the mutual information only

- Interpretation of results, is the *converse* satisfactory? How "difficult" is *achievability*? Additional constraint to consider?

Information theory: setup

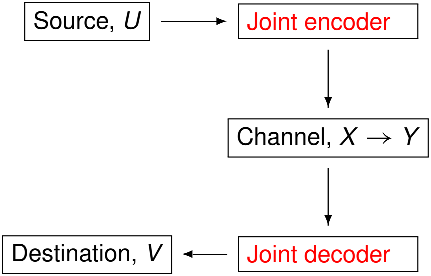
- Fundamental limits on the efficacy of: *i) representation* and *ii) reliable communication* of information



- Separation: traditional, flexible (applications)
- The bounds “asymptoticaly” achievable by separation cannot be improved by any other approach.

Information theory: setup

- Fundamental limits on the efficacy of: *i) representation* and *ii) reliable communication* of information



- Even though the *fundamental* limits cannot be improved, there are benefits (coding complexity, networks, . . .)
- JSCC: no global theory, *ad hoc* approaches

Source-channel matching: rates *vs.* measures

Source $U \to V$ (1 symb/s): $U \sim N(0, \sigma_U^2), \quad \mathbb{E}(U - V)^2 \leq D$
Channel $X \to Y$ (1 symb/s): $Y = X + Z,\ Z \sim N(0, \sigma_Z^2), \quad \mathbb{E}\, X^2 \leq P$

Separation *(traditional)*

a) lossy *compression* of U with distortion D: min. $R(D)$ bit/s
b) reliable *transfer* of information: max. $C(P)$ bit/s

Optimum: $R(D) = C(P)$ and thus

$$D = \frac{\sigma_U^2 \sigma_Z^2}{P + \sigma_U^2}$$

Block coding, complexity of decoding, . . .
Achievability of $R(D)$ and $C(P)$? Only *asymptoticaly* . . .
The mapping $U \to V$ is “deterministic”.

Source-channel matching: rates *vs.* measures

Source $U \to V$ (1 symb/s): $U \sim N(0, \sigma_U^2), \quad \mathbb{E}(U - V)^2 \leq D$
Channel $X \to Y$ (1 symb/s): $Y = X + Z,\ Z \sim N(0, \sigma_Z^2), \quad \mathbb{E}\, X^2 \leq P$

Joint source-channel "coding"

By scaling the inputs and outputs (*symbol-per-symbol*):

$$X = \sqrt{\frac{P}{\sigma_U^2}} U, \quad V = \sqrt{\frac{\sigma_U^2}{P}} \frac{P}{P + \sigma_Z^2} Y, \quad D = \mathbb{E}(U - V)^2 = \frac{\sigma_U^2 \sigma_Z^2}{P + \sigma_U^2}$$

The optimality (separation) is achieved *without coding*!
The mapping $U \to V$ is *stochastic*.

Cramér-Rao bound and Fisher information

- Electrophysiological experiment: stimulus, $\theta \to$ response, r
- Repeated trials (single neuron $\times$ population): response variability
- Stimulus-response model: $R \sim f(r;\theta)$ (θ continuously varying)
- *How precisely can we estimate the fixed θ from the observed r?*
- The estimator $\hat{\theta}(R)$ with mean $m(\theta) = \mathbb{E}_\theta\,\hat{\theta}(R)$

Cramér-Rao bound: $$\operatorname{Var}\hat{\theta}(R) \geq \frac{m'(\theta)^2}{J(\theta)}$$

Fisher information: $$J(\theta) = \int \left[\frac{\partial \log f(r;\theta)}{\partial\theta}\right]^2 f(r;\theta)\,\mathrm{d}r$$

- Conditions on $f(r;\theta)$? $J(\theta)$ easy to calculate, but $m(\theta)$?

Asymptotic theory

- *Problem*: CR bound achievability and bias $b(\theta) = m(\theta) - \theta$
- Restrict to mean squared error $\mathrm{MSE}(\theta)$ of unbiased estimators

$$\mathrm{MSE}(\theta) \geq \frac{1}{J(\theta)} \qquad \text{since } \mathrm{MSE}(\theta) = \mathrm{Var}\,\hat{\theta}(R) + b^2(\theta)$$

- Assume i.i.d. case: $f(r_1, \ldots, r_n; \theta) = \prod_{i=1}^n f(r_i; \theta)$
- As $n \to \infty$ there exists $\hat{\theta}_n$: $\sqrt{n}(\hat{\theta}_n - \theta) \to N\big(0, 1/J(\theta)\big)$

$$\mathrm{MSE}_n(\theta) \geq \frac{1}{nJ(\theta)} \qquad \text{tight for large } n$$

- More general $f(r_1, \ldots, r_n; \theta)$: CR bound *vs.* asymptotics of $\hat{\theta}_n$?
  LAN: Greenwood *et al.*, *Phys. Rev. E* **60** (1999)