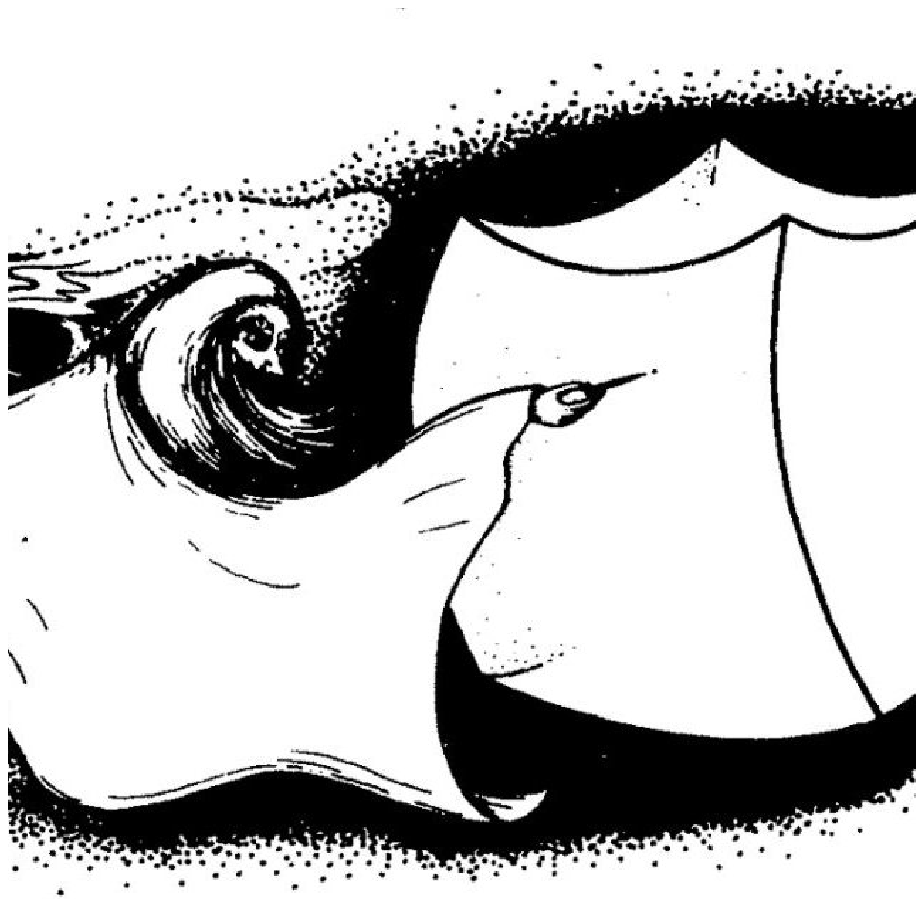
Information theory, entropy (and statistical mechanics)

Vojkan Jaksic
McGill University

PART II: RELATIVE ENTROPY



SETTING

$\mathcal{A} = \{a_1, \cdots, a_l\}$ finite alphabet (often $\mathcal{A} = \{0, 1\}^N$).

$\mathcal{P}(\mathcal{A})$–the set of all probability measures on $\mathcal{A}$.

The entropy of $P \in \mathcal{P}(\mathcal{A})$ is

$$S(P) = \sum -P(a) \log P(a).$$

This lecture is dedicated to relative entropy which involves pairs $(P, Q) \in \mathcal{P}(\mathcal{A}) \times \mathcal{P}(\mathcal{A})$.

RELATIVE ENTROPY

$$S(P,Q) = \sum_{a\in\mathcal{A}} P(a) \log \frac{P(a)}{Q(a)}$$

$0/0 = 0$. Kullback-Leibler divergence (1951).

If $Q = P_{\text{ch}}$, $P_{\text{ch}}(a) = 1/|\mathcal{A}|$,

$$S(P, P_{\text{ch}}) = S(P_{\text{ch}}) - S(P) = \log|\mathcal{A}| - S(P) \geq 0$$

with equality iff $P = P_{\text{ch}}$.

Relative entropy = ”information distance”. It is not a metric. $(P,Q) \mapsto S(P,Q)$ is not symmetric and the triangle inequality fails.

BASIC PROPERTIES

(1) $S(P,Q) \geq 0$ with equality iff $P = Q$.

(2) Pinker inequality:

$$S(P,Q) \geq \frac{1}{2}\left(\sum |P(a) - Q(a)|\right)^2$$

with equality iff $P = Q$.

(3) The map $(P,Q) \mapsto S(P,Q)$ is jointly convex:

$$S(\lambda P_1 + (1-\lambda)P_2, \lambda Q_1 + (1-\lambda)Q_2) \leq \lambda S(P_1, Q_1) + (1-\lambda)S(P_2, Q_2).$$

(4) For further properties and axiomatic characterizations see the Lecture Notes.

STOCHASTIC MONOTONICITY

$\mathcal{A}_1, \mathcal{A}_2$ two finite alphabets.

$M = [M(a,b)]_{(a,b)\in\mathcal{A}\times\mathcal{B}}$ stochastic matrix: $M(a,b) \geq 0$,

$$\sum_b M(a,b) = 1.$$

$M(a,b)$= transition probability $a \to b$.

Induced map $\mathbb{M} : \mathcal{P}(\mathcal{A}_1) \to \mathcal{P}(\mathcal{A}_2)$,

$$\mathbb{M}(P)(b) = \sum_{a\in\mathcal{A}_1} P(a)M(a,b).$$

Stochastic montotonicity

$$S(\mathbb{M}(P), \mathbb{M}(Q)) \leq S(P,Q).$$

FIRST APPLICATION: SHANNON THEOREM

$\{0,1\}^*$= the set of all finite length words with letters from $\{0,1\}$.

Code: one-one map

$$C : \mathcal{A} \to \{0,1\}^*.$$

$C(a)$ codeword, $C(\mathcal{A}) = \{C(a)\}$ the codebook.

If $P \in \mathcal{P}(\mathcal{A})$ is the source statistics, the expected code length is

$$\langle C \rangle_P = \sum_{a \in F} \ell(C(a)) P(a).$$

The goal is to minimize $\langle C \rangle_P$.

Prefix-free coding.

A word $u \in \{0,1\}^*$ is a prefix of a word $w \in \{0,1\}^*$ if $w = uv$ for some $u \in \{0,1\}^*$. A set $W \subset \{0,1\}^*$ is called prefix-free if no element of W is a prefix of some other element of W.

Kraft-McMillan inequality: If $W = \{w_1, \cdots, w_n\}$ is prefix-free, then

$$\sum_{j=1}^{n} 2^{-\ell(w_j)} \leq 1.$$

The converse also holds: if $1 \leq l_1 \leq \cdots \leq l_n$ is a sequence of integers satisfying

$$\sum_{j=1}^{n} 2^{-l_j} \leq 1,$$

then there exists a prefix-free set $W = \{w_1, \cdots, w_n\}$ in $\{0, 1\}^*$ such that $\ell(w_j) = l_j$.

A code C is called prefix-free if its codebook $C(\mathcal{A})$ is a prefix free subset of $\{0, 1\}^*$.

To each $P \in \mathcal{P}(\mathcal{A})$ we can associate a prefix-free code by taking

$$l(a) = -\lceil \log_2 P(a) \rceil.$$

Then $l(a) \geq 1$ and

$$\sum_{a \in \mathcal{A}} 2^{-l(a)} \leq \sum_{a \in \mathcal{A}} P(a) = 1,$$

and so there exists prefix-free code C such that

$$\ell(C(a)) = -\lceil \log_2 P(a) \rceil.$$

A code with these properties is not unique and any such code is called a *Shannon's code* for P. Note that for any Shannon code

$$\langle C \rangle_P \leq S(P) + 1.$$

A basic result of the prefix-free coding is:

The expected length of any prefix-free code $C : \mathcal{A} \mapsto \{0, 1\}^*$ satisfies

$$\langle C \rangle_P \geq S(P).$$

The basic asymptotic result discussed in the first lecture is an easy consequence of these results and asymptotic irrelevance of the prefix-free assumption. Quick reminder.

A code sequence $(C_N)_{N\geq 1}$ is the collection of one-one maps

$$C_N : \mathcal{A}^N \to \{0,1\}^*.$$

(P_N) sequence of marginals of an ergodic P on one-sided shift.

$$\langle C_N \rangle_{P_N} = \sum_{x_1^N \in \mathcal{A}^N} \ell(C_N(x_1^N)) P_N(x_1^N).$$

Theorem.

$$\liminf_{N\to\infty} \frac{\langle C_N \rangle_{P_N}}{N} \geq s(P)$$

and there are **optimal codes** for which

$$\lim_{N\to\infty} \frac{\langle C_N \rangle_{P_N}}{N} = s(P)$$

Proof: Apply previous discussion to $\mathcal{A}^N$ instead of $\mathcal{A}$ and then use Elias construction (transforming a code to prefix-free one without affecting the asymptotic).

Proof of the relation $\langle C \rangle_P \geq S(P)$.

We introduce the Kraft-MacMillan pressure

$$\mathsf{P}_{\mathsf{KM}} = \log_2 \left(\sum_{a \in \mathcal{A}} 2^{-\ell(C(a))} \right) \leq 0,$$

and the Kraft-McMillan probability distribution

$$P_{\mathsf{KM}}(a) = \frac{2^{-\ell(C(a)}}{\sum_{b \in \mathcal{A}} 2^{-\ell(C(b))}}.$$

Then (with logarithms in the base 2),

$$S(P, P_{\mathsf{KM}}) = \langle C \rangle_P - S(P) + \mathsf{P}_{\mathsf{KM}},$$

which can be written as

$$\langle C \rangle_P - S(P) = S(P, P_{\mathsf{KM}}) - \mathsf{P}_{\mathsf{KM}}.$$

Hence

$$\langle C \rangle_P \geq S(P)$$

follows from the sign of the relative entropy and the Kraft-McMillan inequality.

The identity gives much more and indicates the mechanism that leads to saturation of the Shannon bound in the asymptotic settings.

More preciaely, the Shannon bound is saturated,

$$\lim_{N\to\infty} \frac{\langle C_N \rangle_{P_N}}{N} = s(P),$$

iff

$$\lim_{N\to\infty} \frac{1}{N} S(P_N, P_{\mathsf{KM},N}) = 0$$

and

$$\lim_{N\to\infty} \frac{1}{N} \mathsf{P}_{\mathsf{KM},N} = 0.$$

Particularly interesting if the code is universal!

SECOND APPLICATION: GIBBS VARIATIONAL PRINCIPLE

$\mathcal{A}$= set of configurations of a physical system under consideration.

Example: Gas of molecules on lattice $\{1, \cdots, N\}$.

$$\mathcal{A} = \{(\omega_1, \cdots, \omega_N) \,|\, \omega_j \in \{0, 1\}\}$$

Molecule is present/absent at lattice site j corresponds to $\omega_j = 1/0$. Configurations: words of length N.

Hamiltonian (energy) map $H : \mathcal{A} \to \mathbb{R}$. $H(a)$= energy of the configuration a.

Physical states = elements of $\mathcal{P}(\mathcal{A})$.

$$\langle H \rangle_P = \sum_a H(a) P(a)$$

the expected value of energy in a state P.

A state of thermal equilibrium at inverse temperature β is described by the Gibbs Cannonical Ensemble

$$P_\beta(a) = \mathsf{e}^{-\beta H(a)} / Z(\beta)$$

$$Z(\beta) = \sum_{b \in \mathcal{A}} \mathsf{e}^{-\beta H(b)}.$$

Pressure $\mathsf{P}(\beta) = \log Z(\beta)$.

Gibbs Variational Principle:

$$\mathsf{P}(\beta) = \max_{P \in \mathcal{P}(\mathcal{A})} \left(S(P) - \beta \langle H \rangle_P \right)$$

with unique maximizer $P = P_\beta$.

Starting point of equilibrium statistical mechanics.

Proof. :

$$S(P, P_\beta) = \beta\langle H\rangle_P - S(P) + \mathsf{P}(\beta).$$

The result follows from $S(P, P_\beta) \geq 0$ which gives

$$\mathsf{P}(\beta) \geq S(P) - \beta\langle H\rangle_P$$

with equality iff $P = P_\beta$.

PARALLELS AND ORTHOGONALITY

Information theory (IT): the code length map $a \mapsto \ell(C(a))$.

Statistical mechanics (SM): Hamiltonian map $a \mapsto H(a)$.

In both cases one considers the expectation values $\langle C \rangle_P$ and $\langle H \rangle_P$.

Kraft-McMillan probability distribution parallels Gibbs Canonical Ensemble. Same for the respective pressures.

$$P_{\mathsf{KM}}(a) = \frac{2^{-\ell(C(a))}}{\sum_{b\in\mathcal{A}} 2^{-\ell(C(b))}}$$

$$P_\beta(a) = \frac{\mathrm{e}^{-\beta H(a)}}{\sum_{b\in\mathcal{A}} \mathrm{e}^{-\beta H(b)}}.$$

The starting points of both theories (Shannon theorem and the Gibbs variational principle) follow from the parallel relative entropy balance equations

$$S(P, P_{\mathsf{KM}}) = \langle C\rangle_P - S(P) + \mathsf{P}_{\mathsf{KM}},$$

$$S(P, P_\beta) = \beta\langle H\rangle_P - S(P) + \mathsf{P}(\beta).$$

Now to orthogonality:

In IT P is given, it is the statistics of the source. In SM one searches for P describing physical state of thermal equilibrium.

In SM Hamiltonian H is given. In IT one searches for codes that minimize the cost function $\langle C \rangle_P$.

SM comes with conservation of energy and one looks for thermal states such that

$$\langle H \rangle_{P_\beta} = e.$$

This defines $e \mapsto \beta(e)$ and the Gibbs Variational Principle gives that $P_{\beta(e)}$ is the unique maximizer or

$$\{S(P) \,|\, \langle H \rangle_P = e\}.$$

Setting

$$s(e) = S(P_{\beta(e)}), \qquad p(e) = \mathsf{P}(\beta(e))$$

one arrives at the basic thermodynamical equations

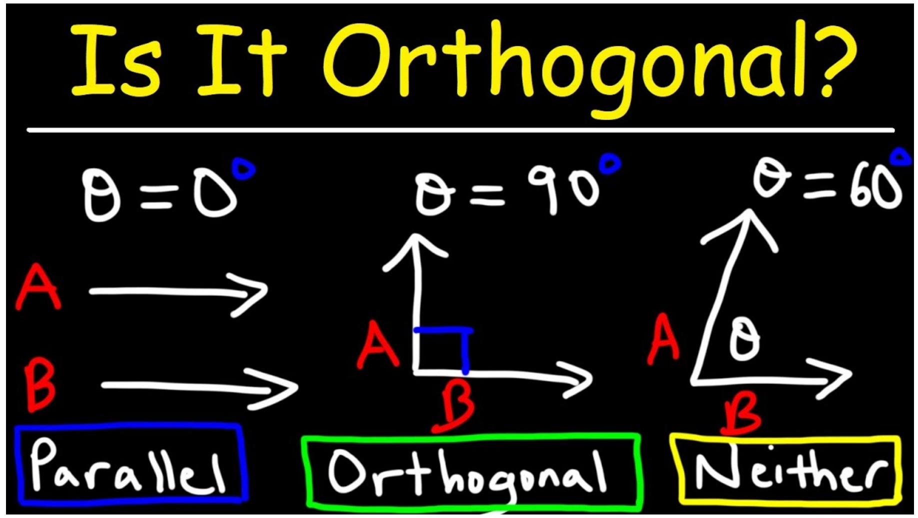
$$s(e) = e\beta(e) + p(e), \qquad \frac{\mathrm{d}s(e)}{\mathrm{d}e} = \beta(e).$$

In IT one minimizes the code cost while it is the pressure that is conserved through the bound

$$\mathsf{P}_{\mathsf{KM}} = \log_2 \left(\sum_{a \in \mathcal{A}} 2^{-\ell(C(a))} \right) \leq 0,$$

which is asymptotically saturated for optimal codes achieving Shannon's bound.

Universal codes lead to universal Hamiltonians with completely broken locality structure.



These observations lead to a particular research program partly sketched in the Toulouse Winter 2024 course. The further links with Boltzmann entropy and Sanov's theorem (Large Deviation Principle) are also discussed there.

THIRD APPLICATION: HYPOTHESIS TESTING AND STEIN LEMMA

We know that the underlying probabilistic experiment is with probability $1/2$ described by P and with probability $1/2$ by Q.

Hypothesis I: Q is correct. Hypothesis II: P is correct.

By performing an experiment we wish to decide with minimal error probability which Hypothesis is correct.

A *test* is $T \subset \mathcal{A}$. If the outcome is in T, we chose Hyp II. If the outcome is not in T, we choose Hyp I.

Error probabilities are $Q(T)$ (type-I error) and $P(T^c)$ (type-II error).

Two coins

Coin 1. $P(\mathsf{Head}) = P(\mathsf{Tail}) = 1/2$.

Coin 2. $Q(\mathsf{Head}) = 2/3$, $Q(\mathsf{Tail}) = 1/3$.

Test $T = \{\mathsf{Head}\}$. Type-I error = $1/3$, Type-II error =$1/2$.

Test $T = \{\mathsf{Tail}\}$. Type-I error =$2/3$, Type-II error=$1/2$.

Type-I error is minimized for $T = \mathsf{Head}$. Completely intuitive.

Back to the general setting.

For $\epsilon \in (0, 1)$, the Stein error exponent is

$$s(\epsilon) = \min\{Q(T) \mid P(T^c) < \epsilon\}.$$

The type-I error is minimized by allowing ϵ-window in the type-II error.

The errors and error exponents get better if the experiment is repeated N times. The outcomes are in $\mathcal{A}^N = \mathcal{A} \times \cdots \times \mathcal{A}$, and P, Q are replaced by $P_N = P \times \cdots \times P$ and Q_N. The Nth Stein error exponent is

$$s_N(\epsilon) = \min\{Q_N(T_N) \mid T_N \subseteq \mathcal{A}^N,\ P_N(T_N^c) < \epsilon\}$$

Stein Lemma:

$$\lim_{N\to\infty} \frac{1}{N} \log s_N(\epsilon) = -S(P,Q)$$

Symbolically,

$$s_N(\epsilon) \sim \mathrm{e}^{-NS(P,Q)}.$$

Basic (and very general result) + novel perspective on the first Shannon theorem (source coding).

General setting: $P \sim (P_N)$, $Q \sim (Q_N)$ two ergodic sources on (Ω, φ) such that the specific relative entropy

$$s(P,Q) = \lim_{N\to\infty} \frac{1}{N} S(P_N, Q_N)$$

exists.

Under very general conditions the Stein Lemma holds:

$$\lim_{N\to\infty} \frac{1}{N} \log s_N(\epsilon) = -s(P,Q).$$

BACK TO SOURCE CODING

$P \sim (P_N)$ ergodic source, $0 < \epsilon < 1$ "allowed coding error".

Coding pair (C_N, D_N). Coder $C_N : \mathcal{A}^N \to \{0, 1\}^M$. Decoder $D_N : \{0, 1\}^M \to \mathcal{A}^N$.
Compression coefficient = M/N.

The error probability of the coding pair (C_N, D_N) is

$$P_N \left\{ x_1^N \in \mathcal{A}^N \mid D_N \circ C_N(x_1^N) \neq x_1^N \right\}.$$

If this probability is $< \epsilon$, the pair (N, M) is called ϵ-good.

For given N, let $M(N)$ be smallest number such that the pair $(N, M(N))$ is ϵ-good.

$M(N)/N$ is the **best possible compression** subject to the allowed ϵ-error probability.

The optimal $M(N)$ is

$$M(N) = \min\{\lfloor \log_2 |T_N| \rfloor \mid T_N \subseteq \mathcal{A}^N,\ P_N(T_N^c) < \epsilon\}.$$

Taking Q to be the product of (uniform) measures P_{ch} on $\mathcal{A}$,

$$Q(T_N) = |T_N|/|\mathcal{A}|^N,$$

$$\begin{aligned}\lim_{N\to\infty} \frac{M(N)}{N} &= \log_2 |\mathcal{A}| + \lim_{N\to\infty} \frac{1}{N} s_N(\epsilon) \\ &= \log_2 |\mathcal{A}| - s(P, P_{\text{ch}}) = s(P).\end{aligned}$$

Stein Lemma can be viewed as the generalization of the Shannon source coding with completely different interpretation.

Source coding = hypothesis testing between $Q = \times P_{\mathrm{ch}}$ (the source of maximal specific entropy) and P.

Is statistical mechanics interpretation of Stein Lemma possible?

Yes, and it is linked with interpretation of a very important discoveries (early 1990's) in non-equilibrium statistical physics dealing with entropy production, second law of thermodynamics, and entropic fluctuation relations.

Evans-Cohen-Morriss, Evans-Searles, Gallavotti-Cohen, Lebowitz-Spohn...

FLUCTUATION RELATIONS AND ARROW OF TIME

Alphabet $\mathcal{A}$ is equipped with involution $\Theta : \mathcal{A} \to \mathcal{A}$. To be interpreted as time-reversal.

To $P \in \mathcal{P}(A)$ one associates P_Θ by $P_\Theta(a) = P(\Theta(a))$.

Relative entropy (relative information) function

$$I_{P,P_\Theta}(a) = \log \frac{P(a)}{P_\Theta(a)}.$$

$$\langle I_{P,P_\Theta} \rangle_P = \sum_a I_{P,P_\Theta}(a) P(a) = S(P, P_\Theta).$$

We denote by Q the probability distribution of the random variable $I_{P,P_\ominus}$ wrt P,

$$Q(s) = P\{a \,|\, I_{P,P_\ominus}(a) = s\}.$$

Fluctuation Relation: $Q(-s) \neq 0$ iff $Q(s) \neq 0$ and in this case

$$\frac{Q(-s)}{Q(s)} = \mathrm{e}^{-s}.$$

Fundamental universal relation that implies and refines the signature $S(P, P_\ominus) \geq 0$ since, with $\mathcal{S} = \{s \,|\, Q(s) > 0\}$,

$$\begin{aligned} S(P, P_\ominus) &= \sum_{s \in \mathcal{S}} sQ(s) = \sum_{s>0, s\in\mathcal{S}} s(Q(s) - Q(-s)) \\ &= \sum_{s>0, s\in\mathcal{S}} sQ(s)(1 - \mathrm{e}^{-s}) \geq 0. \end{aligned}$$

Proof of the Fluctuation Relation. Set

$$e(\alpha) = \sum_a \mathrm{e}^{-\alpha I_{P,P_\Theta}(a)} P(a)$$

$$e(\alpha) = \sum_a [P_\Theta(a)]^\alpha [P(a)]^{1-\alpha} = \sum_a [P_\Theta(\Theta(a))]^\alpha [P(\Theta(a))]^{1-\alpha}$$

$$= \sum_a [P_\Theta(a)]^{1-\alpha} [P(a)]^\alpha = e(1-\alpha).$$

Hence

$$\sum_{s\in\mathcal{S}} \mathrm{e}^{-\alpha s} Q(s) = \sum_{s\in\mathcal{S}} \mathrm{e}^{-(1-\alpha)s} Q(s),$$

It follows that for all $\alpha \in \mathbb{C}$,

$$\sum_{s\in\mathcal{S}} \mathrm{e}^{-\alpha s}(Q(s) - \mathrm{e}^s Q(-s))$$

and so

$$Q(s) = e^s Q(-s).$$

Back to one-sided shift (Ω, φ), ergodic $P \sim (P_N)_{N\geq 1}$.

Reversal $\Theta_N : \mathcal{A}^N \to \mathcal{A}^N$,

$$\Theta_N(x_1 x_2 \cdots x_N) = x_N x_{N-1} \cdots x_1.$$

$P_{\Theta_N} = P_N \circ \Theta_N$,

$$P_{\Theta_N}(x_1 \cdots x_N) = P_N(x_N x_{N-1} \cdots x_1).$$

Fluctuation Relation holds for pairs (P_N, P_{Θ_N})!

There exists unique ergodic source $\widehat{P}$ on (Ω, φ) such that

$$\widehat{P}_N = P_{\Theta_N}.$$

$\widehat{P}$ is *the reversal* of P.

The entropy production observables are ($x = x_1 x_2 \cdots \in \Omega$)

$$\sigma_N(x) = \sigma_N(x_1 x_2 \cdots x_N) = I_{P_N, P_{\Theta_N}}(x_1, \cdots, x_N)$$
$$= \log \frac{P_N(x_1 x_2 \cdots x_N)}{P_N(x_N x_{N-1} \cdots x_1)}.$$

Note that

$$\int_\Omega \sigma_N \mathrm{d}P = S(P_N, P_{\Theta_N}).$$

Under very mild regularity assumptions on P (subadditivity decoupling, in addition to ergodicity), for P-a.e. x,

$$\lim_{N\to\infty} \frac{1}{N}\sigma_N(x) = \lim_{N\to\infty} \frac{1}{N} S(P_N, P_{\Theta_N}) =: \mathrm{ep}$$

This limit is the entropy production of (Ω, φ, P), the measure of its irreversibility. The limit is automatically ≥ 0 (the Second Law).

Stein Lemma and hypothesis testing of arrow of time.

Hypothesis testing between P_N and P_{Θ_N}. Stein error exponent

$$s_N(\epsilon) = \min\{P_N(T_N) \,|\, T_N \subseteq \mathcal{A}^N,\ P_{\Theta_N}(T_N^c) < \epsilon\}$$

Stein Lemma connects to the entropy production:

$$\begin{aligned}\lim_{N\to\infty} \frac{1}{N} \log s_N(\epsilon) &= -\lim_{N\to\infty} \frac{1}{N} S(P_N, P_{\Theta_N}) \\ &= \underbrace{-\mathrm{ep} \leq 0}_{\text{Second Law}} \quad .\end{aligned}$$

Entropy production/the Second Law quantifies distinction/separation between the past and future.

Fluctuation Relations (tautological for finite N). They lead to the fine form of the Second Law.

Entropy production LLN

$$\lim_{N\to\infty} \frac{1}{N}\sigma_N(x) = \mathsf{ep} \qquad P - \text{a.s.}$$

Fine form concerns fluctuations in this convergence and validity of the Large Deviation Principle

$$P\{\sigma_N(x) \sim s\} \sim \mathsf{e}^{-NI(s)}$$

where I is the rate function (non-negative, convex, vanishing only at ep).

The real Fluctuation Relation follows from finite N relations and is

$$\underbrace{I(-s) = I(s) + s}_{\text{Fine form the Second Law}}$$

EXAMPLE: MARKOV SOURCE

Stoshastic matrix $M = [M(a,b)]_{(a,b)\in\mathcal{A}\times\mathcal{A}}$. $M(a,b) > 0$.

$\mathbf{p} = (p(a))_{a\in\mathcal{A}}$ the unique invariant probability vector,
$\mathbf{p}M = \mathbf{p}$, $p(a) > 0$.

Induced Markov chain source: unique P on (Ω, φ) with marginals

$$P_N(x_1 x_2 \cdots x_N) = p(x_1)M(x_1 x_2)M(x_2, x_3)\cdots M(x_{N-1}x_N).$$

Reversal $\widehat{P}$: also Markov chain induced by stochastic matrix

$$\widehat{M}(a,b) = \frac{p(b)}{p(a)}M(b,a).$$

Same invariant vector $\mathbf{p}$. P and $\widehat{P}$ are ergodic.

$$\sigma_N(x) = \log \frac{p(x_1)}{p(x_N)} + \sum_{j=1}^{N-1} \log \frac{M(x_j, x_{j+1})}{M(x_{j+1}, x_j)}.$$

Ergodic theorem:

$$\mathrm{ep} = \lim_{N\to\infty} \frac{1}{N}\sigma_N(x) = \sum_{(a,b)} p(a)M(a,b) \log \frac{M(a,b)}{M(b,a)}.$$

Very intutive formula!

$$R_a(b) = M(a,b), \widehat{R}_a(b) = \widehat{M}(a,b).$$

Rows of M and $\widehat{M}$, $R_a, \widehat{R}_a \in \mathcal{P}(\mathcal{A})$.

$$\mathrm{ep} = \sum_{a\in\mathcal{P}(\mathcal{A})} p(a) S(R_a, \widehat{R}_a).$$

This formula should be compared with the one for the specific entropy of the Markov process (first computed by Shannon)

$$s(P) = \lim_{N\to\infty} \frac{S(P_N)}{N} = -\sum_{(a,b)} p(a)M(a,b) \log M(a,b)$$

$$= \sum_{a\in\mathcal{P}(\mathcal{A})} p(a)S(R_a).$$

Note that ep ≥ 0 and ep $= 0$ iff $R_a = \widehat{R}_a$ for all a.

$$\text{ep} = 0 \qquad \text{iff} \qquad \underbrace{p(a)M(a,b) = p(b)M(b,a)}_{\text{Detailed Balance Condition}}$$

Far reaching generalizations, technical state of the art results: Cuneo N., VJ., Pillet C-A, Shirikyan A.: Large deviations and fluctuation theorem for selectively decoupled measures on shift spaces, Rev. Math. Phys. 31 (2019)

Fine form = standard LDP for Markov chains.
$r(\alpha)$ = spectral radius of the matrix $[M(a,b)^{1-\alpha}\widehat{M}(a,b)^{\alpha}]$,

$$e(\alpha) = \log r(\alpha).$$

Symmetry

$$e(\alpha) = e(1-\alpha).$$

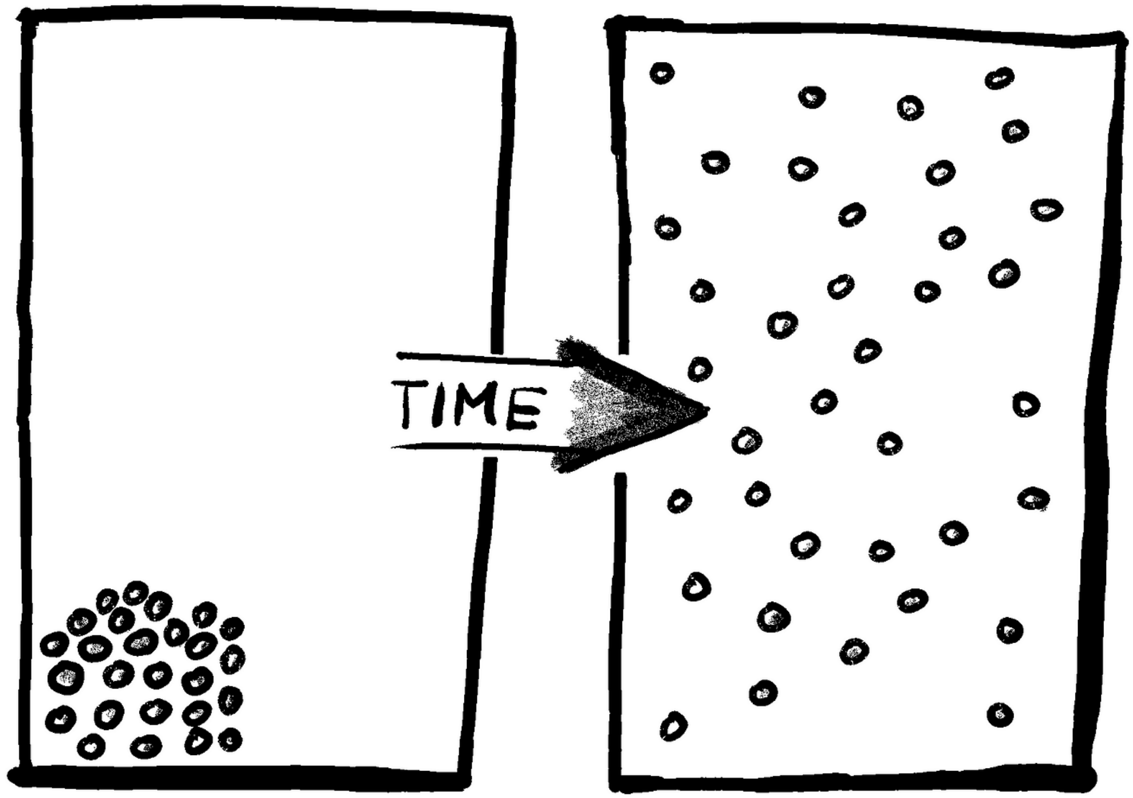
LDP for σ_N holds with the rate function

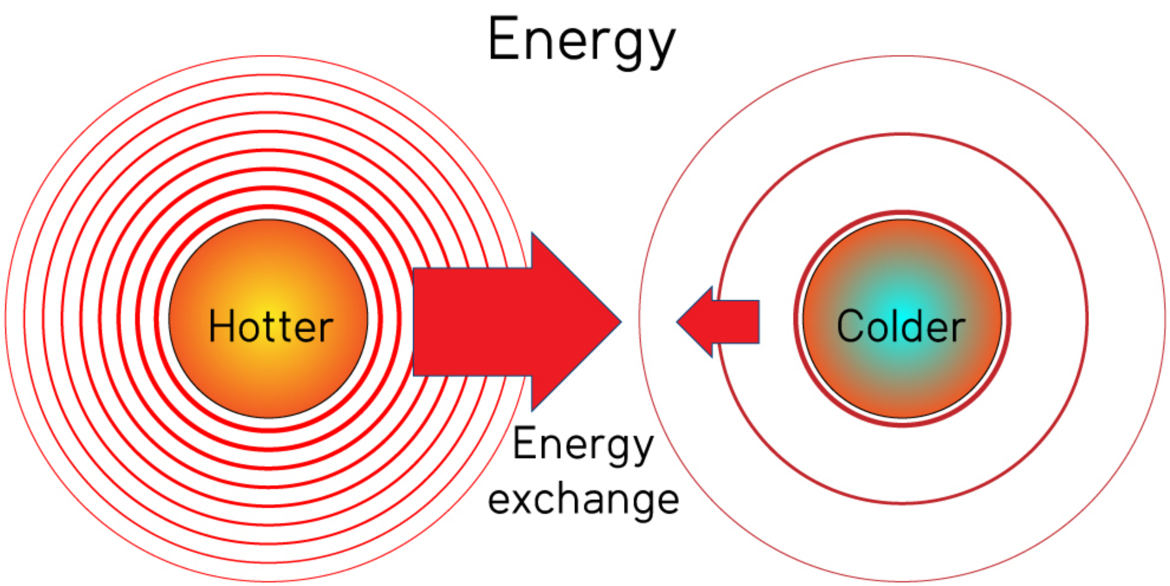
$$I(s) = \sup_{\alpha\in\mathbb{R}}(\alpha s - e(-\alpha)).$$

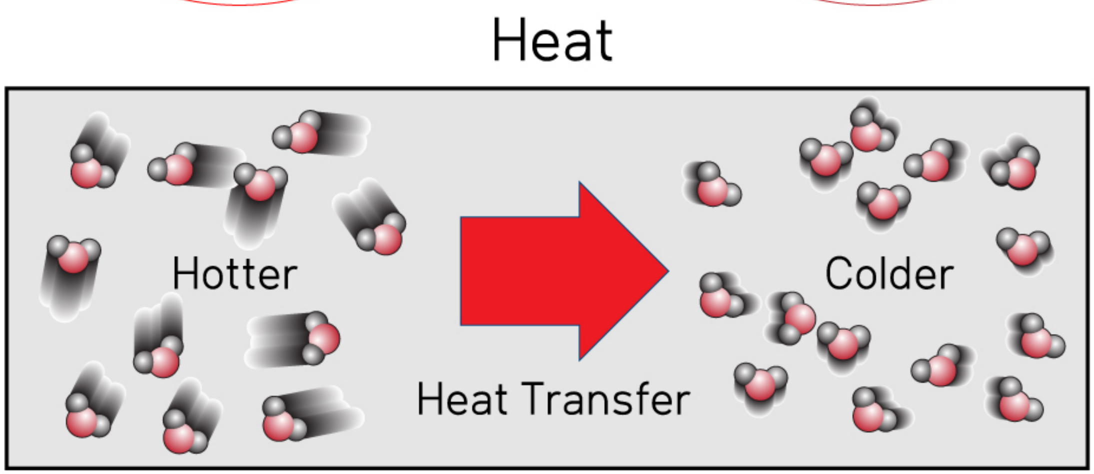
$$I(-s) = I(s) + s.$$

$e(\alpha)$ is linked to other (finer) error exponents (Chernoff, Hoeffding). That discussion involves Renyi's relative entropy.

Very general theory! Part of the theory of dynamical systems. LLN for σ_N gives the Second Law, LDP its fine form. Difficult mathematical problems. What about physics?

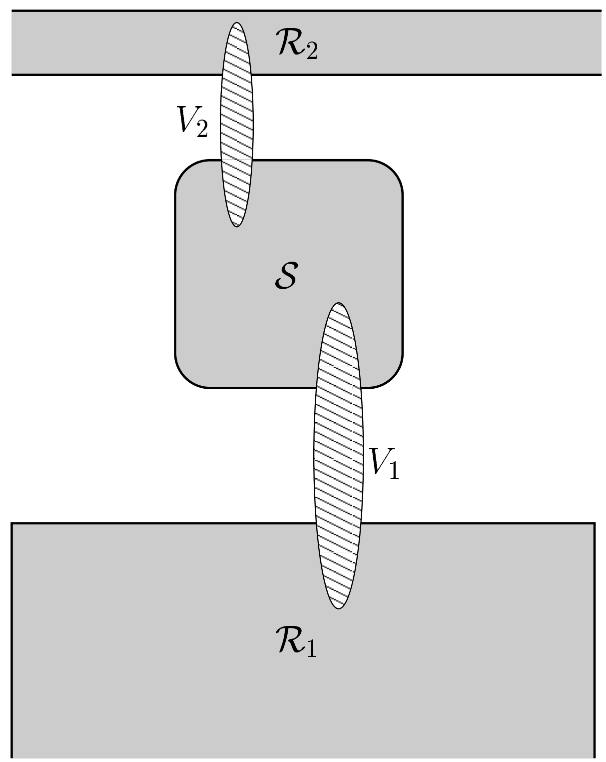

TIME


Energy
Hotter
Colder
Energy exchange


Heat
Hotter
Colder
Heat Transfer

sks.to/jg

OPEN SYSTEMS



Basic paradigm of non-equilibrium statistical mechanics. The reservoirs are in thermal equilibrium at inverse temperatures β_1, β_2. The temperature differential induces energy (heat) transfer from the hotter to the colder reservoir.

Hamiltonian setting of classical mechanics! The reservoirs are infinitely extended (to sustain constant energy fluxes). $\mathcal{S}$ is finite dimensional Hamiltonian system.

The formalism applies, and the Stein error exponent (hypothesis testing of the arrow of time) is linked to the thermodynamics by the basic relation

$$\text{ep} = \beta_1\Phi_1 + \beta_2\Phi_2,$$

where Φ_1, Φ_2, are heat fluxes ($\Phi_1 + \Phi_2 = 0$) out of reservoirs $\mathcal{R}_1$, $\mathcal{R}_2$.

$$\text{ep} \geq 0 \qquad \text{heat flows from hot to cold.}$$

$$\text{ep} > 0 \qquad \text{there is heat flowing from hot to cold!}$$

Rigorous results in Hamiltonian setting are scarce and technically difficult.

For additional information and references see

J.V., Pillet C-A., Shirikyan A.: Entropic fluctuations in thermally driven harmonic networks, J. Stat. Phys., 166 (2017), 926-1015

and forthcoming monographs:

1. Cuneo N., J.V.., Pillet C-A., Shirikyan A.: What is a Fluctuation Theorem? Springer.

2. Cuneo N., J.V., Nersesyan V., Pillet C-A., Shirikyan A.: Mathematical Theory of the Fluctuation Theorem. CRM monograph series.