The Multiple Faces of Entropy

Tristan Stérin
Supervision: Prof. Damien Woods
tristan.sterin@mu.ie
Maynooth University
Ireland

ERC No 772766, SFI 18/ERCS/5746

November 12, 2019

Personal Background

- Pure maths / theoretical computer science
- Machine Learning
- Finance
- Molecular Programming

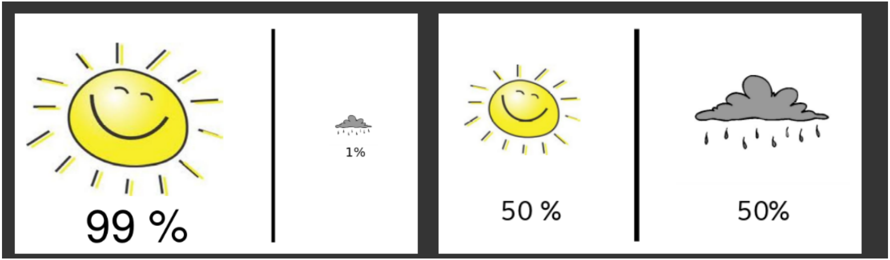
Personal Background

- Pure maths / theoretical computer science:
  - Combinatorics
  - Information Theory (Coding Theory)
- Statistics/Machine Learning:
  - Maximum Entropy Principle
  - Cross entropy and Softmax Layers
- Finance:
  - Risk Assessment
  - Portofolio Diversification
- Molecular Programming: thermodynamics, ΔS and ΔG.

Entropy is everywhere!!

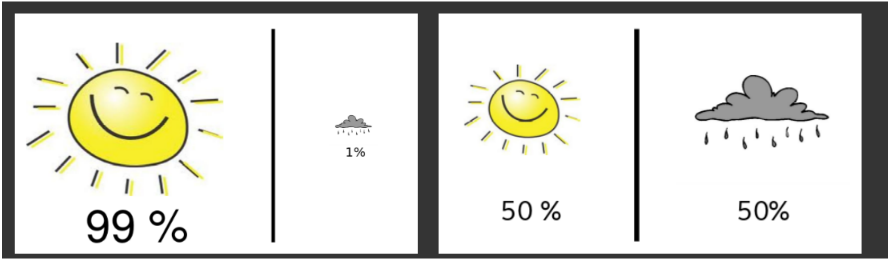
Measuring Unpredictability

Entropy is a property of a **probability distribution**: a probability distribution can have a low or a high entropy.



Which forecast leaves tomorrow's weather the most unpredictable ?

Measuring Unpredictability



Entropy can equivalently be seen as a:

- Measure of lack of information / unpredictability
- Measure of fairness
- Measure of impurity (cf. decision trees)

Goal: we want to mathematically construct a measure of unpredictability on probability distributions.

Khinchin Axioms for Entropy

Let Σ_n be the set of probability distributions with n outcomes:

$$\Sigma_n = \{(p_1, \ldots, p_n) \,|\, p_i \geq 0 \text{ and } \sum_{i=1}^{n} p_i = 1\}$$

And $\Sigma = \cup_n \Sigma_n$. We wish to construct $H : \Sigma \to \mathbb{R}$, our measure of unpredictability, such that:

1. The function H is symmetric in $(p_1, \ldots, p_n)$, for example: $H(p_1, p_2) = H(p_2, p_1)$.
2. The restriction of H to Σ_n is maximal for the uniform distribution $U_n = (1/n, 1/n, \ldots, 1/n)$: $\forall (p_1, \ldots, p_n) \in \Sigma_n, \quad H(p_1, \ldots, p_n) \leq H(1/n, \ldots, 1/n)$.
3. Zero probabilities don't count: $H(p_1, \ldots, p_n, 0, 0, \ldots, 0) = H(p_1, \ldots, p_n)$.

The Fourth Axiom

So far, our entropy function is defined for discrete probability distributions. We extend it to discrete random variables: $H(X) = H(p_X)$.

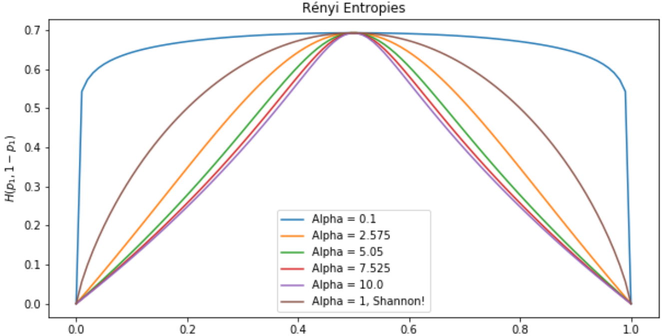
In order to get to the most common definitions of entropy we need a fourth axiom. There is a **weak** and a **strong** version of the axiom:

- **Weak**: Let X, Y be two discrete **independent** random variables, then: $H(X \times Y) = H(X) + H(Y)$.
- **Strong**: For any discrete random variables X, Y, we have: $H(X \times Y) = H(X) + \sum_x \mathrm{Pr}(X = x) H(Y|X = x)$.

The Fourth Axiom

The **weak** version of the fourth axiom leads to a class of functions, parametrised by $\alpha, b \geq 0$, called *Rényi entropies*:

$$H_\alpha(X) = \frac{1}{1-\alpha} \log_b \sum_{i=1}^{k} p_i^\alpha$$



Shannon's Entropy

The **strong** version of the fourth axiom gives the Rényi entropy H_1 ($\alpha = 1$), which is referred as **Shannon's entropy**:

$$H(X) = -\sum_{i=1}^{n} p_i \log_b(p_i)$$

Computer scientists will tend to choose $b = 2$ while physicists will tend to use $b = e$.

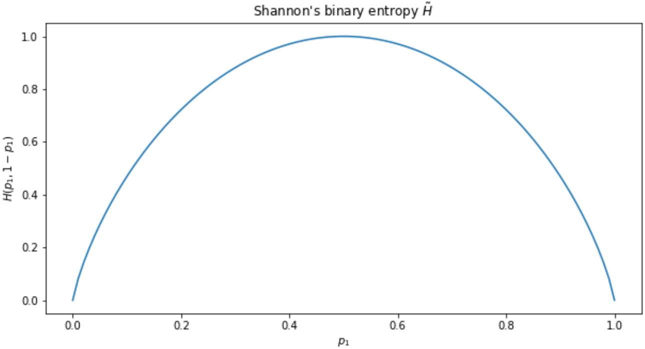
"My greatest concern was what to call it. I thought of calling it 'information', but the word was overly used, so I decided to call it 'uncertainty'. When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, 'You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, nobody knows what entropy really is, so in a debate you will always have the advantage."

Claude E. Shannon, 1961
Conversation with Myron Tribus, reported in [TM]

Shannon's Entropy: Basic Properties

$$0 \leq H(X) \leq H(U_n) = \log_b(n)$$

$$H(X_1 \times X_2 \times \cdots \times X_n) \leq \sum_{i=1}^{n} H(X_i) \text{ (subadditivity)}$$



Shannon's Entropy: Relation to KL Divergence

The **KL** divergence of two distributions $P = (p_1, \ldots, p_n)$ and $Q = (q_1, \ldots, q_n)$ is defined by:

$$D_{\mathsf{KL}}(P||Q) = -\sum_{i=1}^{n} p_i \log_b \frac{q_i}{p_i}$$

It has the following properties:

1. $0 \leq D_{\mathsf{KL}}(P||Q) < \infty$
2. $D_{\mathsf{KL}}(P||Q) = 0$ iff $P = Q$
3. $D_{\mathsf{KL}}(P||Q) \neq D_{\mathsf{KL}}(Q||P)$

Shannon's entropy reflects the "distance" to the uniform distribution $U_n = (1/n, \ldots, 1/n)$:

$$H(X) = \log_b(n) - D_{\mathsf{KL}}(p_X||U_n)$$

Shannon's Entropy: Continuous Generalization

In statistics or finance, a continuous version of Shannon's entropy is often used:

Let X be a continuous random variable with density f, the continuous entropy of X, or *differential entropy* of X is given by:

$$h(X) = -\int_{-\infty}^{\infty} f(x) \log_b f(x) dx$$

Warning: continuous entropy can be **negative** therefore it does not inherit of all the properties of Shannon's entropy. In that sense, KL divergence can be seen as more fundamental than entropy because it remains positive in the continuous domain:

$$D_{\mathsf{KL}}(f||g) = -\int_{-\infty}^{\infty} f(x) \log_b \frac{g(x)}{f(x)} \geq 0$$

Counting With Entropy

Let $[n] = \{1, \ldots, n\}$.
By a combinatorial argument, we can deduce that:

$$2^n = \sum_{k=0}^{n} \binom{n}{k}$$

Indeed, we know that $|\mathcal{P}([n])| = 2^n$ and $\mathcal{P}([n]) = \bigcup_{k=0}^{n} \mathcal{P}_k([n])$ (disjoint). Hence $2^n = |\mathcal{P}([n])| = \sum_{k=0}^{n} |\mathcal{P}_k([n])| = \sum_{k=0}^{n} \binom{n}{k}$.

Counting With Entropy I

Now, what can we tell about partial sums $\sum_{k=0}^{m} \binom{n}{k}$ with $0 \leq m \leq n$?
Entropy gives an answer! Let $\tilde{H}(\alpha)$ be *Shannon's binary entropy*, i.e. $\tilde{H}(\alpha) = H(\alpha, 1-\alpha)$ in base 2. Then, for $\alpha < 1/2$, for all n:

$$\sum_{k \leq \alpha n} \binom{n}{k} \leq 2^{\tilde{H}(\alpha) n}$$

Counting With Entropy II

For $\alpha < 1/2$, for all n:

$$\sum_{k \leq \alpha n} \binom{n}{k} \leq 2^{\tilde{H}(\alpha)n}$$

Proof. Let $\mathcal{C} = \bigcup_{k=0}^{\alpha n} \mathcal{P}_k([n])$, the set of subsets of $[n]$ of size at most αn. We have $|\mathcal{C}| = \sum_{k \leq \alpha n} \binom{n}{k}$. Let X be selected uniformly at random in $\mathcal{C}$. We have $H(X) = \log_2(|\mathcal{C}|)$. Hence, we just have to prove that $H(X) \leq \tilde{H}(\alpha n)$. We write $X = (X_1, X_2, \dots, X_n)$ with X_k indicating that $k \in X$. By subadditivity and symmetry:

$$H(X) \leq H(X_1) + \cdots + H(X_n) = nH(X_1)$$

We have: $H(X_1) = \tilde{H}(p)$ with $p = \Pr(1 \in X)$. We show $p \leq \alpha$.

Counting With Entropy III

We have:

$$\begin{aligned} P(1 \in X) &= \sum_{m=0}^{\alpha n} P(1 \in X \mid |X| = m) P(|X| = m) \\ &= \sum_{m=0}^{\alpha n} \frac{m}{n} P(|X| = m) \leq \sum_{m=0}^{\alpha n} \frac{\alpha n}{n} P(|X| = m) \\ &\leq \alpha \sum_{m=0}^{\alpha n} P(|X| = m) = \alpha \end{aligned}$$

Hence, $p \leq \alpha$, and, because $\alpha \leq 1/2$ we have $\tilde{H}(p) \leq \tilde{H}(\alpha)$ and the result:

$$\sum_{k \leq \alpha n} \binom{n}{k} \leq 2^{\tilde{H}(\alpha) n}$$

Counting With Entropy IV

Many more combinatorial applications, for instance:

- In geometry: Loomis-Whitney theorem about counting modulo projection
- In linear algebra: Brégman's theorem bounding matrix permanents
- In graph theory: Kahn-Lovász theorem about perfect matchings or also counting proper colorings of a regular graph

Coding Theory: Lossless Compression I

≪ Stately, plump Buck Mulligan came from the stairhead, bearing a bowl of lather on which a mirror and a razor lay crossed. A yellow dressinggown, ungirdled, was sustained gently behind him on the mild morning air. ≫ James Joyce, Ulysses.

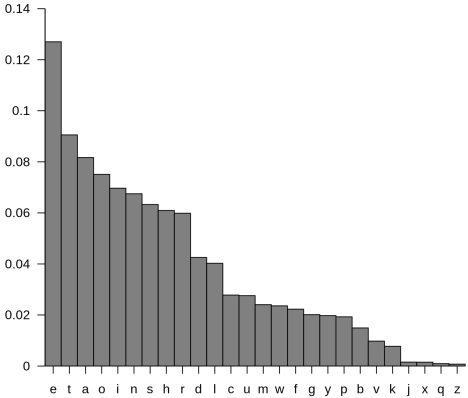
- There are **1,498,853** ascii characters in the book.
- Stored in the standard 8 bits (1 byte) per character: **1.5 M** bytes

Question: Can we do better?

Intuitively: Yes because of the distribution of letters in the English language.

Coding Theory: Lossless Compression II

Intuitively: Yes because of the distribution of letters in the English language. We could use a **variable length** code.



Coding Theory: Formalism I

We consider our source (e.g. *Ulysses*) to be randomly sampled from a random variable X with value in the alphabet $\mathcal{A} = \{\mathsf{a}, \mathsf{b}, \mathsf{c}, \dots, \mathsf{z}\}$. In the case of *Ulysses*, the probability distribution of X is the one of the English language.
We want to construct an encoding function $C : \mathcal{A} \to \{0,1\}^*$. For instance, in **ASCII**:

$$\begin{aligned} C(\mathsf{a}) &= 01100001 \\ C(\mathsf{b}) &= 01100010 \\ &\vdots \\ C(\mathsf{z}) &= 01111010 \end{aligned}$$

We naturally extend C from $\mathcal{A}$ to $\mathcal{A}^*$:
$C(\mathsf{ab}) = C(\mathsf{a})C(\mathsf{b}) = 0110000101100010$.

Coding Theory: Formalism II

Requirement. We want C to be **uniquely decodable**:

$$\forall w_1, w_2 \in \mathcal{A}^*,\ w_1 \neq w_2 \Rightarrow C(w_1) \neq C(w_2)$$

The following code is **not** uniquely decodable:

$$C(\mathsf{a}) = 0$$
$$C(\mathsf{b}) = 010$$
$$C(\mathsf{c}) = 01$$
$$C(\mathsf{d}) = 10$$

The code 010 could be either: b, ca or ad.

Coding Theory: Formalism III

Evaluation Metric. We evaluate the efficiency of a code C by average number of bits required to encode a character:

$$\mu(C) = \mathbb{E}[|C(X)|] = \sum_{x \in \mathcal{A}} |C(x)| \mathrm{Pr}(X = x)$$

A code C^* is **optimal** if it minimizes the function μ.

Coding Theory: Entropy Bounds

Theorem. An optimal (uniquely decodable) code satisfies:

$$H(X) \leq \mu(C^*) \leq H(X) + 1$$

Hence, entropy gives a bound on **how much a source can be compressed without loss**. Entropy corresponds to the average number of bits required to optimally encode data sampled from X.

Optimal codes can be constructed explicitly via **Huffman algorithm**. This algorithm is behind all lossless compression techniques (e.g. zip). Back to *Ulysses*:

- Stored in the standard 8 bit (1 byte) per character: **1.5 M**
- Stored with an *optimal uniquely decodable binary code*: **889K**

Compression ratio: x1.69

Statistics: Maximum Entropy Distributions

Standard probability distributions can be seen as maximum entropy distributions under some moment constraint.

- Uniform: $f(x) = \frac{1}{b-a}$. No constraints.
- Exponential: $f(x) = \lambda \text{exp}(-\lambda x)$. Constraint: $\mathbb{E}[X] = 1/\lambda$.
- Normal: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \text{exp}\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$. Constraints: $\mathbb{E}[X] = \mu$ and $\mathbb{E}[(X-\mu)^2] = \sigma^2$.

Statistics: Maximum Entropy Principle

The maximum entropy principle can be use to discriminate between several candidate distributions in a given statistical problem. This principle will select the *less specific*/*fairest* distribution given a set of constraints. In practice, this principle can be used for:

- Prior/posterior selection in Bayesian inference
- Maximum entropy models (e.g. logistic regression)
- Probability density estimation

Finance

- Measuring the risk of a portofolio: entropy of the distribution of returns
- Measuring the diversity of a portofolio: entropy of the distribution of investments

Machine-Learning: Cross entropy Objective I

Typical Situation. We want to classify pictures of **Cats** and **Dogs**. We are building a model m_α parametrized by α:

- **Input:** an image
- **Output:** a probability distribution $(p_{\text{Cat}}, p_{\text{Dog}})$ in Σ_2.

Our dataset $\mathcal{D}$ is a collection of pairs:

- (image of a cat, $(1, 0)$)
- (image of a dog, $(0, 1)$)

A common training objective for this situation is **Cross entropy**.

Machine-Learning: Cross entropy Objective II

The cross entropy of two distributions p, q is defined as:

$$H(p,q) = H(p) + D_{\mathsf{KL}}(p||q) = -\sum p_i \mathsf{log}(q_i)$$

Information theory interpretation:

- Entropy $H(p)$: average number of bits needed to optimally encode a stream of data sampled from p.
- Cross entropy $H(p,q)$: average number of bits needed to optimally encode a stream of data sampled from **unkown** p when we believe that the distribution is q.

Machine-Learning: Cross entropy Objective III

The cross entropy of two distributions p, q is defined as:

$$H(p,q) = H(p) + D_{\mathsf{KL}}(p||q) = -\sum p_i \mathsf{log}(q_i)$$

Cross entropy loss:

$$J(\alpha) = \frac{1}{|\mathcal{D}|} \sum_{\mathsf{img},p \in \mathcal{D}} H(p, m_\alpha(\mathsf{img}))$$

Optimization-wise equivalent to KL divergence, it is the **standard loss** in Computer Vision (AlexNet, VGG, Inception etc...).

Machine-Learning: Softmax Layers I

In Machine-Learning, transforming a vector of numbers into a probability distribution is very often done through a **Softmax Layer**. This layer (of a neural network for instance) transforms the vector $(E_1, \ldots, E_n) \in \mathbb{R}^n$ into $(p_1, \ldots, p_n) \in \Sigma_n$ by the rule:

$$p_i = \frac{1}{Z} e^{-\beta E_i}$$

With $Z = \sum e^{-\beta E_i}$ the normalization factor and β a parameter of the layer or a given constant.

But **why use this distribution** ?

Machine-Learning: Softmax Layers

In Machine-Learning, transforming a vector of numbers into a probability distribution is very often done through a **Softmax layer**. This layer (of a neural network for instance) transforms the vector $(E_1, \ldots, E_n) \in \mathbb{R}^n$ into $(p_1, \ldots, p_n) \in \Sigma_n$ by the rule:

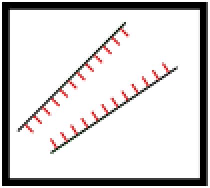
$$p_i = \frac{1}{Z} e^{-\beta E_i}$$

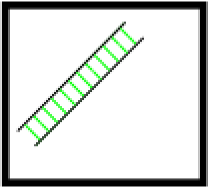
With $Z = \sum e^{-\beta E_i}$ the normalization factor and β a parameter of the layer or a given constant.

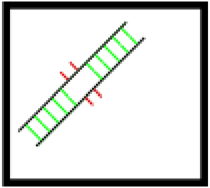
But **why use this distribution** ? Because statistical physics and the maximum entropy principle!!!!

Statistical Physics: The Boltzmann Distribution I

Example DNA Secondary Structures.



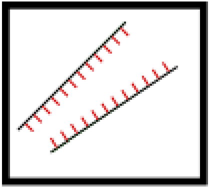




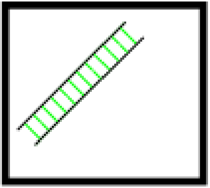
- Theoretically: $(n+1)!$ possible secondary structures for a strand of length n.
- In practice: way less because they are not all physically possible.

Statistical Physics: The Boltzmann Distribution II

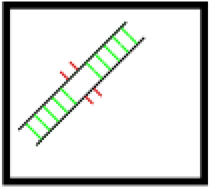
Thanks to **thermodynamics**, we can perform **calorimetric experiments** to measure the energy of each secondary structure: the more negative, the harder to melt.



E1
0 kcal/mol



E2
-23 kcal/mol



E3
-18 kcal/mol

Statistical Physics: The Boltzmann Distribution III

In a testube, we have billion of strands. What will be the secondary structure **distribution** ?

- Energetic argument: low energy configurations should be favored
- Entropic argument: a variety of different configurations should be covered

In statistical physics, the energy/entropy compromise is formalized by **Free Energy Minimization**:

$$\min_{p\in\Sigma_n}\mathbb{E}_p[\vec{E}] - k_B T H(p)$$

With $\vec{E} = (E_1, \dots, E_n)$ the measured energy vector, k_B the *Boltzmann constant* and T the temperature.

Statistical Physics: The Boltzmann Distribution IV

$$\min_{p \in \Sigma_n} \mathbb{E}_p[\vec{E}] - k_B T H(p)$$

The **Free Energy Minimization** has a unique solution, **the Boltzmann distribution**:

$$p_i = \frac{1}{Z} e^{-\frac{E_i}{k_B T}} = \frac{1}{Z} e^{-\beta E_i}$$

The normalization constant Z is seen as a function of $\beta = 1/(k_B T)$. It is called the **partition function**:

$$Z(\beta) = \sum e^{-\beta E_i}$$

Extreme cases:

- $T = 0$: the distribution is spiked at min E_i. Energy wins!
- $T = \infty$: the distribution is uniform. Entropy wins!

Thermodynamics: Clausius Entropy

"I propose to name the quantity S the entropy of the system, after the Greek word $\eta\ \tau\rho o\pi\eta$ (en tropein), the transformation. I have deliberately chosen the word entropy to be as similar as possible to the word energy: the two quantities to be named by these words are so closely related in physical significance that a certain similarity in their names appears to be appropriate."

Rudolf Clausius, 1865

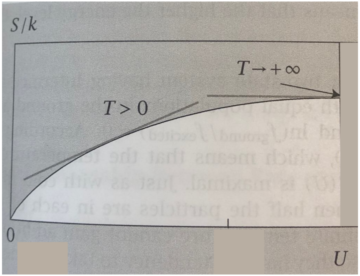
Thermodynamics: Clausius Entropy (theory)

$S = S(U, V, \vec{N})$ **Requirements.**

1. $S \in \mathcal{C}^1$, **concave**
2. S positively homogeneous of degree 1:

$$\forall \alpha > 0,\ S(\alpha U, \alpha V, \alpha \vec{N}) = \alpha S(U, V, \vec{N})$$

3. $\frac{1}{T} = \frac{\partial S}{\partial U} \geq 0$



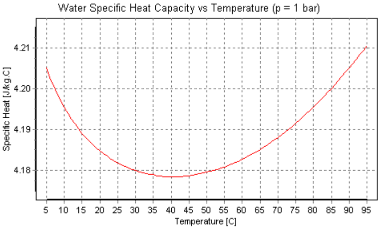
Source: Molecular Driving Forces, p.223

Thermodynamics: Clausius Entropy (practice)

Changes in Clausius entropy are physically **measurable**:

$$\Delta S = \int_{T_A}^{T_B} \frac{C_P(T)}{T} dT$$

Isobaric Heat Capacity. The function C_P is the amount of energy needed by the system to gain temperature at constant pressure. You can have an approximation of water's C_P by measuring the time needed by your kettle to boil 1L of water (use $Q = P\Delta t$).



Source: `https://www.engineersedge.com/physics/properties_of_water_14467.htm`

Thermodynamics: Clausius Entropy (intuition)

Gas phase are more entropic than liquid phases than solid phases. For water for instance:

$$\text{Solid} \xrightarrow{\Delta S = 22 \text{ J}/(\text{mol}K)} \text{Liquid} \xrightarrow{\Delta S = 118{,}89 \text{ J}/(\text{mol}K)} \text{Gas}$$

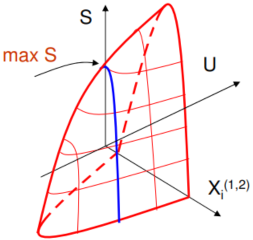
At a fixed volume, it takes more energy to heat a room than a pool than an iceberg by 1C.

Although, intuitively, we want to relate Clausius entropy to the **inner disorder of matter**, it is a **controversial subject**, see [1].

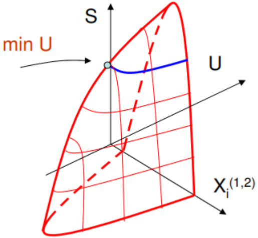
In statistical physics, **Boltzmann entropy**, $k_B \text{log}(W)$ is introduced. This quantity is closely related to Shannon's entropy and encompassed the same intuition. In some contexts it can be associated to Clausius entropy.

Thermodynamics: Second Law I

The second law of thermodynamic is a principle of **entropy maximization**.



Maximizing S at fixed U



Minimizing U at fixed S

Source: `https://ps.uci.edu/~cyu/p115B/LectureNotes/Lecture13.pdf`

Thermodynamics: Second Law II

Name	Variables	Reservoir	"Expression"	At Equilibrium
Entropy	$U, V, \vec{N}$	Constant U	S	Maximal
Energy	$S, V, \vec{N}$	Constant S	U	Minimal
Helmotz Free Energy	$T, V, \vec{N}$	Constant T	F = U - TS	Minimal
Enthalpy	$U, p, \vec{N}$	Constant P	H = U + pV	Minimal
Gibbs Free Energy	$T, p, \vec{N}$	Constant T, P	G = H - TS	Minimal

At equilibrium, chemical systems **minimize their Gibbs Free Energy**:

$$\Delta G = \Delta H - T\Delta S$$

This quantity is measurable since:

$$\Delta H = \int_{T_A}^{T_B} C_P(T)dT \text{ and } \Delta S = \int_{T_A}^{T_B} \frac{C_P(T)}{T}dT$$

Thermodynamics: Second Law III

For a chemical reaction $A \to B$ three possible cases:

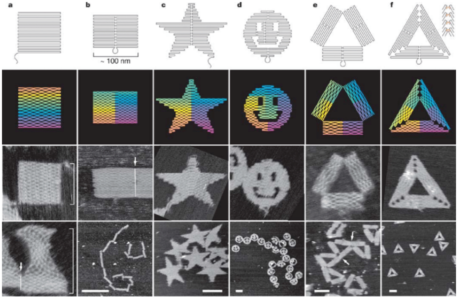
1. $\Delta G > 0$: the reaction $A \to B$ cannot happen, $B \to A$ is favored
2. $\Delta G = 0$: equilibrium
3. $\Delta G < 0$: the reaction can happen (maybe not kinetically favored)

Looking at two cases for ΔH and ΔS:

- $\Delta H > 0$ and $\Delta S < 0$: thermodynamically impossible reaction. An increase in enthalpy must result in an increase in entropy. You can't turn graphite into diamond.
- $\Delta H > 0$ and $\Delta S > 0$: **entropy driven reaction**. Depends on T. Example: $\mathrm{NaNO}_{3(\mathrm{s})} \to \mathrm{Na}+_{(\mathrm{aq})} + \mathrm{NO}3-_{(\mathrm{aq})}$

DNA Programming I

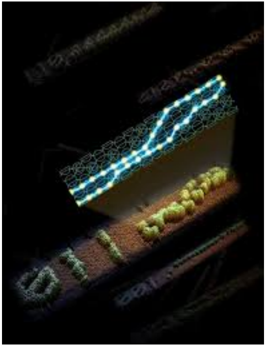
The goal of our field is to **engineer structures out of DNA**.



Folding DNA to create nanoscale shapes and patterns, P. Rothemund, Nature, 2006

DNA Programming II

Ideally, **structures that can compute**.



Diverse and robust molecular algorithms using reprogrammable DNA self-assembly, D. Woods, D. Doty et al., Nature, 2019

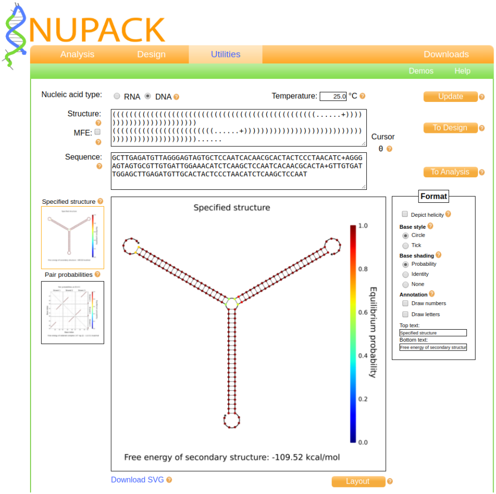
DNA Programming III

The **thermochemistry** of DNA is vital to our field.

Table 1. Nearest-neighbor parameters for DNA/DNA duplexes in 1 M NaCl.[13]

Nearest-neighbor sequence (5'-3'/3'-5')	$\Delta H°$ kJ/mol	$\Delta S°$ J/(mol·K)	$\Delta G°_{37}$ kJ/mol
AA/TT	−33.1	−92.9	−4.26
AT/TA	−30.1	−85.4	−3.67
TA/AT	−30.1	−89.1	−2.50
CA/GT	−35.6	−95.0	−6.12
GT/CA	−35.1	−93.7	−6.09
CT/GA	−32.6	−87.9	−5.40
GA/CT	−34.3	−92.9	−5.51
CG/GC	−44.4	−113.8	−9.07
GC/CG	−41.0	−102.1	−9.36
GG/CC	−33.5	−83.3	−7.66
Terminal A/T base pair	9.6	17.2	4.31
Terminal G/C base pair	0.4	−11.7	4.05

DNA Programming IV



DNA Programming V

Details

Sequence properties

Free energy:	-110.51	kcal/mol	
Base	Number	%	
A	45	27.8	
C	40	24.7	
G	35	21.6	
T	42	25.9	
Other	0	0.0	

Sequence/structure properties

Free energy:	-109.52	kcal/mol
Probability:	0.189	
Ensemble defect:	2.9	nt
Normalized ensemble defect:	1.8	%
Nucleotides:	162	nt

Conclusion

- Shannon's entropy measures the **lack of predictability** of the outcome of a random variable. Shannon's entropy can be applied in many fields. Each field shines a different light on the concept.
- Clausius entropy is related to the amount of heat necessary to change the temperature of a body. Changes in Clausius entropy are **measurable**. By the second law of thermodynamics, Clausius entropy is **maximized at equilibrium**.
- The formal link between Shannon's entropy and Clausius entropy is **obscure** and **controversial**.
- In the case of DNA Programming, Shannon's entropy and Clausius entropy work **hand in hand** in order to predict the probability of formation of DNA structures.

Conclusion

You can find these slides at the following URL:

`https://dna.hamilton.ie/tsterin/`

Thank you!!!

Questions :) ?

References I

P. G. Wright. "Entropy And Disorder". In: *Contemp. Phys.* 11 (1970), pp. 581–588.

Cosma Shalizi. "Note: Information Theory, Axiomatic Foundations, Connections to Statistics". In: *Manuscript* (2008). URL: https://www.stat.cmu.edu/~cshalizi/350/2008/lectures/06a/lecture-06a.pdf.

David Galvin. "Three tutorial lectures on entropy and counting". In: *Manuscript* (2014). URL: https://www3.nd.edu/~dgalvin1/pdf/Kalamazoo_entropy.pdf.

References II

Djalil Chafaï. "From Boltzmann to random matrices and beyond". In: *Annales de la Faculté des Sciences de Toulouse. Mathématiques.* 24.4 (2015), pp. 641–689. DOI: 10.5802/afst.1459. URL: https://hal.archives-ouvertes.fr/hal-00987177.

Rongxi Zhou, Ru Cai, and Guanqun Tong. "Applications of Entropy in Finance: A Review". In: *Entropy* 15 (Nov. 2013), pp. 4909–4931. DOI: 10.3390/e15114909.

Jonathan Fors. "Information Theory Lectures". In: (2013). URL: https://www.icg.isy.liu.se/courses/infotheory/lect3.pdf.