Foundations of Entropy III

MaxEnt and related stuff

Lecture series at the
School on Information, Noise, and Physics of Life
Nis 19.-30. September 2022

by Jan Korbel
all slides can be found at: **slides.com/jankorbel**

Activity III

You have 3 minutes to write down on a piece of paper:

Have you been using entropy in
your research/ your projects?

If yes, how?

My applications: statistical physics, information theory, econophysics, sociophysics, image processing...

"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."

John von Neuman's reply to Claude Shannon's question how to name newly discovered measure of missing information

Information Theory and Statistical Mechanics

E. T. JAYNES
Department of Physics, Stanford University, Stanford, California
(Received September 4, 1956; revised manuscript received March 4, 1957)

Information theory provides a constructive criterion for setting up probability distributions on the basis of partial knowledge, and leads to a type of statistical inference which is called the maximum-entropy estimate. It is the least biased estimate possible on the given information; i.e., it is maximally noncommittal with regard to missing information. If one considers statistical mechanics as a form of statistical inference rather than as a physical theory, it is found that the usual computational rules, starting with the determination of the partition function, are an immediate consequence of the maximum-entropy principle. In the resulting "subjective statistical mechanics," the usual rules are thus justified independently of any physical argument, and in particular independently of experimental verification; whether or not the results agree with experiment, they still represent the best estimates that could have been made on the basis of the information available.

It is concluded that statistical mechanics need not be regarded as a physical theory dependent for its validity on the truth of additional assumptions not contained in the laws of mechanics (such as ergodicity, metric transitivity, equal *a priori* probabilities, etc.). Furthermore, it is possible to maintain a sharp distinction between its physical and statistical aspects. The former consists only of the correct enumeration of the states of a system and their properties; the latter is a straightforward example of statistical inference.

Information entropy = thermodynamic entropy

Maximum entropy principle

In the case of inductive inference, the constraints, or prior information, are given in terms of linear expectation values; i.e., the constraints considered are of the form

$$I_k \equiv \langle \mathcal{I}_k \rangle = \sum_i \mathcal{I}_{k,i} p_i, \tag{2}$$

where $\{\mathcal{I}_{k,i}\}$ are possible realizations (alphabet) of the observable $\mathcal{I}_k$.

Theorem 1 (MEP).—Given the set of constraints $C = \{I_k\}_{k=1}^{\nu}$, the best estimate of the underlying (i.e., true) probability distribution $P = \{p_i\}_{i=1}^{n}$ is the one that maximizes the entropy functional $S(P)$ subject to the constraints; i.e., it maximizes the Lagrange functional

$$S(P) - \sum_{k=1}^{\nu} \lambda_k I_k. \tag{1}$$

Maximum entropy principle

General approach - method of Lagrange multipliers

Maximize $L(p) = S(p) - -\alpha \sum_i p_i - \sum_k \lambda_k \sum_i I_{i,k} p_i$

$$\frac{\partial L}{\partial p_i} = \frac{\partial S(p)}{\partial p_i} - \alpha - \sum_k \lambda_k I_{i,k} \stackrel{!}{=} 0$$

In case $\psi_i(P) = \frac{\partial S(p)}{\partial p_i}$ is invertible for p_i, we get that

$$p_i^\star = \psi_i^{(-1)} \left(\alpha + \sum_k \lambda_k I_{i,k} \right)$$

Legendre structure of thermodynamics - interpretation of L

$$L(p) = S(p) - \beta U(p) = \Psi(p) = -\beta F(p)$$

free entropy

MB, BE & FD MaxEnt

Maxwell-Boltzmann

$$S_{MB} = -\sum_{i=1}^{k} p_i \log \frac{p_i}{g_i}$$

$$p_i^\star = \frac{g_i}{Z} \exp(-\epsilon_i / T)$$

Bose-Einstein

$$S_{BE} = \sum_{i=1}^{k} \left[(\alpha_i + p_i) \log(\alpha_i + p_i) - \alpha_i \log \alpha_i - p_i \log p_i \right]$$

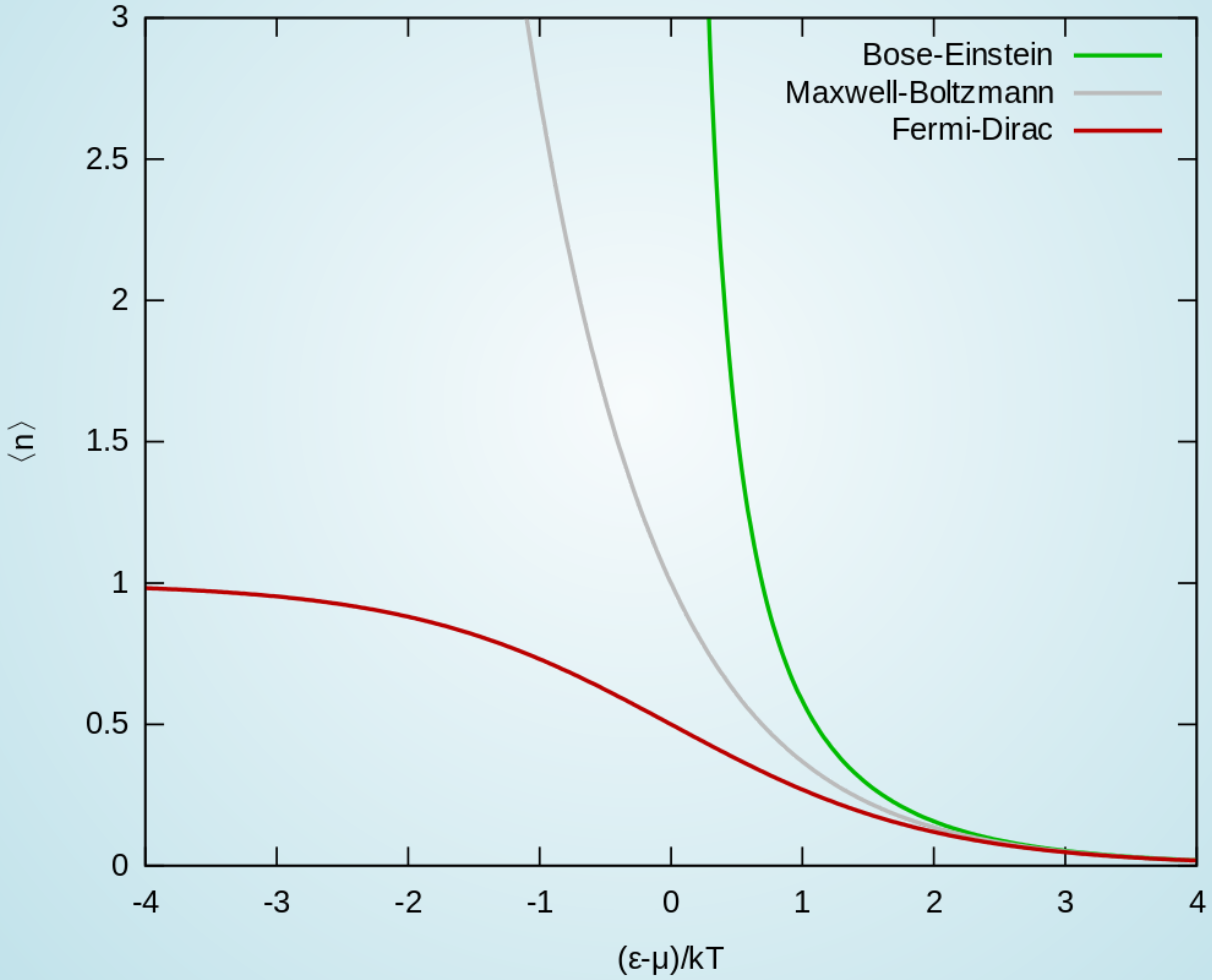
$$p_i^\star = \frac{\alpha_i}{Z} \frac{1}{\exp(\epsilon_i / T) - 1}$$

Fermi-Dirac

$$S_{FD} = \sum_{i=1}^{k} \left[-(\alpha_i - p_i) \log(\alpha_i - p_i) + \alpha_i \log \alpha_i - p_i \log p_i \right]$$

$$p_i^\star = \frac{\alpha_i}{Z} \frac{1}{\exp(\epsilon_i / T) + 1}$$

MB, BE & FD MaxEnt



Structure-forming systems

$$S(\wp) = -\sum_{ij} \wp_i^{(j)} (\log \wp_i^{(j)} - 1) - \sum_{ij} \wp_{ij} \log \frac{j!}{n^{j-1}}$$

where $j\wp_i^{(j)} = p_i^{(j)}$

Normalization: $\sum_{ij} j\wp_i^{(j)} = 1$ Energy: $\sum_{ij} \epsilon_i^{(j)} \wp_i^{(j)} = U$

MaxEnt distribution: $\wp_i^{(j)} = \frac{n^{j-1}}{j!} \exp(-\alpha j - \beta \epsilon_i^{(j)})$

The normalization condition gives $\sum_j j \mathcal{Z}_j e^{-\alpha j} = 1$

where $\mathcal{Z}_j = \frac{n^{j-1}}{j!} \sum_i \exp(-\beta \epsilon_i^{(j)})$ is the partial partition function

We get a polynomial equation in $e^{-\alpha}$

Average number of molecules $\mathcal{M} = \sum_{ij} \wp_i^{(j)}$

Free energy: $F = U - TS = -\frac{\alpha}{\beta} - \frac{\mathcal{M}}{\beta}$

MaxEnt of Tsallis entropy

$$S_q(p) = \frac{\sum_i p_i^q - 1}{1-q}$$

MaxEnt distribution is: $p_i^\star = \exp_q(\alpha + \beta\epsilon_i)$

Note that this is not equal in general to $q_i^\star = \frac{\exp_q(\beta\epsilon_i)}{\sum_i \exp_q(\beta\epsilon_i)}$

However, it is possible to use the identity

$$\exp_q(x+y) = \exp_q(x)\exp_q\left(\frac{y}{\exp_q(x))^{1-q}}\right)$$

The MaxEnt distribution of Tsallis entropy can be expressed as

$$p_i^\star(\beta) = \exp_q(\alpha + \beta\epsilon_i) = \exp_q(\alpha)\exp_q(\tilde{\beta}\epsilon_i) = q_i^\star(\tilde{\beta})$$

where $\tilde{\beta} = \frac{\beta}{\exp_q(\alpha)^{1-q}}$

(sometimes called self-referential temperature)

MaxEnt for path-dependent processes

and relative entropy

- What is the most probable histogram of a process $X(N, \theta)$?
  - θ - parameters, k histogram of $X(N, \theta)$
  - $P(k|\theta)$ is probability of finding a histogram
- Most probable histogram $k^\star = \arg\min_k P(k|\theta)$
- In many cases, the probability can be decomposed to

$$P(k|\theta) = W(k)G(k|\theta)$$

  - $W(k)$ - multiplicity of histogram
  - $G(k|\theta)$ - probability of a microstate belong to k
- $$\underbrace{\log P(k|\theta)}_{S_{rel}} = \underbrace{\log W(k)}_{S_{MEP}} + \underbrace{\log G(k|\theta)}_{S_{cross}}$$
  - S_{rel} - relative entropy (divergence)
  - S_{cross}- cross-entropy, depends on constraints given by θ

The role of constraints

The cross-entropy corresponds to the constraints

For the case of expected energy, it can be expressed through the cross entropy

$$S_{cross}(p|q) = -\sum_i p_i \log q_i$$

where q_i are prior probabilities. By taking $q_i^\star = \frac{1}{Z} e^{-\beta \epsilon_i}$ we get

$$\boxed{S_{cross}(p|q^\star) = \beta \sum_i p_i \epsilon_i + \ln Z}$$

However, for the case of path-dependent process, the natural constraints might not be of this form

Kullback-Leibler divergence

$$\boxed{D_{KL}(p||q) = -S(p) + S_{cross}(p,q)}$$

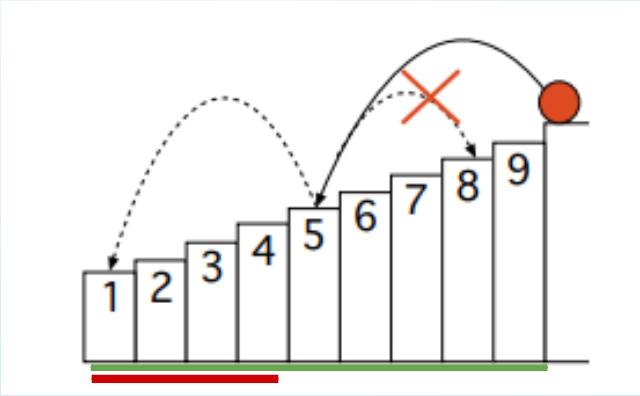
MaxEnt for SSR processes

From multiplicity of trajectory histograms, we have shown that the entropy of SSR is

$$S_{SSR}(p) = -N \sum_{j=2}^{n} \left[p_i \log\left(\frac{p_i}{p_1}\right) + (p_1 - p_i) \log\left(1 - \frac{p_i}{p_1}\right) \right]$$

Let us now consider that after each run (when the system reaches the ground state) we drive the ball to a random state with probability q_i

After each jump the effective space reduces



MaxEnt for SSR processes

One can see that the probability of sampling a histogram k_i is

$$G(k|q) = \prod_{i=1}^{n} \frac{q_i^{k_i}}{Q_{i-1}^{k_i}}$$

where $Q_i = \sum_{j=1}^{i} q_i$ and $Q_0 \equiv 1$.

Similarly, one can determine the probability of sampling a particular sequence x. Each visit to a state $i > 1$ in the sequence x contributes to the probability of the next visit to a state $j < i$ with a factor $1/Q_{i-1}$, whatever j gets sampled. Only if $i = 1$ do we not get such a renormalization factor, since the process restarts and all states i are valid targets with probability q_i. It follows that $G(k|q,N) = \prod_{i=1}^{W} q_i^{k_i} \prod_{j=2}^{W} Q_{i-1}^{-k_i}$, and the cross-entropy is found to be

$$S_{cross}(p|q) = -\sum_{i=1}^{n} p_i \log q_i - \sum_{i=2}^{n} p_i \log Q_{i-1}$$

By assuming in $q_i \propto e^{-\beta\epsilon_i}$ the cross-entropy is

$$S_{cross}(p|q) = \beta \sum_{i=1}^{n} p_i \epsilon_i + \beta \sum_{i=2}^{n} p_i f_i = \mathcal{E} + \mathcal{F}$$

where $f_i = \ln \sum_i e^{-\beta\epsilon_i}$

MaxEnt for Pólya urns

Probability of observing a histogram

$$p(\mathcal{K}) = \binom{N}{k_1, \ldots, k_c} p(\mathcal{I})$$

By carefully taking into account the initial number of balls in the urn n_i we end with

$$S_{Pólya}(p) = -\sum_{i=1}^{c} \log(p_i + 1/N)$$

$$S_{Pólya}(p|q) = -\sum_{i=1}^{c} \left[\frac{q_i}{\gamma} \log\left(p_i + \frac{1}{N}\right) - \log\left(1 + \frac{1}{N\gamma} \frac{q_i - \gamma}{p_i + \frac{1}{N}}\right) + \log q_i \right]$$

where $q_i = n_i/N, \gamma = \delta/N$

Long-run limit

By taking $N \to \infty$, we get

$$S_{Pólya}(p) = -\sum_{i=1}^{c} \log p_i$$

$$S_{Pólya}(p|q) = -\sum_{i=1}^{c} \left[\frac{q_i}{\gamma} \log p_i + \log q_i \right]$$

Maximizing $\psi(p|\theta)$ with respect to p on $\sum p_i = 1$, either leads to the solution

$$p_i = \frac{1}{\zeta}(q_i - \gamma), \tag{14}$$

for $0 < p_i < 1$, or, if this can not be satisfied, to boundary solutions $p_i = 0$. ζ is a normalization constant. There exist three scenarios:

(A) For $\gamma < \min(q)$, equation (14) is the max-ent solution for all i (no boundary-solutions). The limit $\gamma \to 0$ provides the correct multinomial limit $p_i \to q_i$.

(B) If $\max(q) > \gamma > \min(q)$, ψ gets maximal for those i with $q_i > \gamma$ and follows solution equation (14); those i where $q_i < \gamma$ are boundary-solutions, $p_i = 0$.

(C) For $\gamma > \max(q)$ all p_i are boundary-solutions, meaning that one winner i takes it all, $p_i = 1$, while all other states have vanishing probability.

Related extremization principles

As we already found out, the MaxEnt principle can be seen as a special case of the **principle of minimum relative entropy**

$$\boxed{p^\star = \arg\min_p D(p||q)}$$

In many cases, the divergence can be expressed as

$$D(p||q) = -S(p) + S_{cross}(p, q)$$

Priors q can be obtained from theoretical models or measurements

Posteriors p can be from parametric family or from a special class of probability distributions

Relative entropy is well defined for both discrete and continuous distributions

It connects information theory, thermodynamics and geometry

Maximization for trajectory probabilities - Maximum caliber

Let us now consider the whole trajectory $\boldsymbol{x}(t)$ with probability $p(\boldsymbol{x}(t))$

We define the term **caliber**, which is the KL-divergence of the path probability

$$S_{cal}(p|q) = \int \mathcal{D}\boldsymbol{x}(t) p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x}(t))}{q(\boldsymbol{x}(t))}$$

N.B.: Entropy production can be written in terms of caliber as

$$\Sigma_t = S_{cal}[p(\boldsymbol{x}(t))|\tilde{p}(\tilde{\boldsymbol{x}}(t))]$$

Review on MaxEnt & MaxCal

REVIEWS OF MODERN PHYSICS, VOLUME 85, JULY–SEPTEMBER 2013

Principles of maximum entropy and maximum caliber in statistical physics

Steve Pressé*

Department of Physics, Indiana University-Purdue University Indianapolis, Indianapolis, Indiana 46202, USA

Kingshuk Ghosh

Department of Physics and Astronomy, University of Denver, Denver, Colorado 80208, USA

Julian Lee

Department of Bioinformatics, Soongsil University, Seoul 156-743, Korea

Ken A. Dill

Laufer Center for Physical and Quantitative Biology and Departments of Physics and Chemistry, Stony Brook University, New York, New York 11794, USA

(published 16 July 2013)

The variational principles called *maximum entropy* (MaxEnt) and *maximum caliber* (MaxCal) are reviewed. MaxEnt originated in the statistical physics of Boltzmann and Gibbs, as a theoretical tool for predicting the equilibrium states of thermal systems. Later, entropy maximization was also applied to matters of information, signal transmission, and image reconstruction. Recently, since the work of Shore and Johnson, MaxEnt has been regarded as a principle that is broader than either physics or information alone. MaxEnt is a procedure that ensures that inferences drawn from stochastic data satisfy basic self-consistency requirements. The different historical justifications for the entropy $S = -\sum_i p_i \log p_i$ and its corresponding variational principles are reviewed. As an illustration of the broadening purview of maximum entropy principles, maximum caliber, which is path entropy maximization applied to the trajectories of dynamical systems, is also reviewed. Examples are given in which maximum caliber is used to interpret dynamical fluctuations in biology and on the nanoscale, in single-molecule and few-particle systems such as molecular motors, chemical reactions, biological feedback circuits, and diffusion in microfluidics devices.

DOI: 10.1103/RevModPhys.85.1115 PACS numbers: 82.20.Pm, 05.40.−a, 89.70.Cf, 02.50.Tt

MaxCal and Markov processes

For a discrete time process, the path entropy is

$$H(T) = -\sum_{i_0, i_1, \ldots, i_T} p_{i_0 i_1, \ldots, i_T} \log p_{i_0 i_1, \ldots, i_T}. \tag{82}$$

Now, we impose pairwise constraints for each step $m \to n$ over the time period $[0, T]$, i.e.,

$$\langle N_{m\to n} \rangle = \sum_{i_0, \ldots, i_T} p_{i_0 \ldots, i_T} N_{m\to n}(i_0, \ldots, i_T), \tag{83}$$

where $N_{m\to n}(i_0, \ldots, i_T) \equiv \sum_{k=0}^{T-1} \delta_{i_k, m} \delta_{i_{k+1}, n}$ counts the number of $m \to n$ transitions. We verify that $\sum_{m,n} N_{m\to n} = T$.

We then maximize the path entropy, Eq. (72), with the constraints given by Eq. (83) using λ_{mn} as the Lagrange multiplier to constrain $\langle N_{m\to n} \rangle$. This yields

$$p_{i_0, \ldots, i_T} = \prod_{k=0}^{T-1} p_{i_k \to i_{k+1}} \propto e^{-\sum_{m,n} \lambda_{mn} \sum_{k=0}^{T-1} \delta_{i_k, m} \delta_{i_{k+1}, n}}, \tag{84}$$

where, from the second proportionality, we have $p_{i_k \to i_{k+1}} \propto e^{-\lambda_{i_k i_{k+1}}}$ and the probability $p_{i_k \to i_{k+1}}$ is understood as the conditional probability $p(i_{k+1}|i_k)$. Thus, under the constraints imposed by Eq. (83), the joint probability distribution $p_{i_0 \cdots i_T}$ given by Eq. (84) is a first-order Markov process. That is, it can be rewritten as the product of transition probabilities which describe the probability of being in a state at some time $k + 1$ as depending only on the state at time k.

Other extremal principles in ThD

Prigogine's principle of minimum entropy production

Principle of maximum entropy production (e.g., for living systems)

Further reading

Extremal principles in non-equilibrium thermodynamics

From Wikipedia, the free encyclopedia

The Azimuth Project

Extremal principles in non-equilibrium thermodynamics

MaxEnt as an inference tool

Maximum entropy principle consists of two steps:

(I) Finding a distribution (*MaxEnt* distribution) that maximizes entropy under given constraints.
(II) Plugging the distribution into the entropic functional and calculating physical quantities as thermodynamic potentials, temperature, or response coefficients (specific heat, compressibility, etc.).

The first step is a statistical inference procedure.

The second step gives us the connection to thermodynamics.

(i) For each MaxEnt distribution, there exists the whole class of entropies and constraints leading to generally different thermodynamics.
(ii) It is possible to establish transformation relations of Lagrange parameters (and subsequently the thermodynamic quantities) for classes of entropies and constraints giving the same MaxEnt distribution.

Entropy 23 (2021) 96

3. Calibration Invariance of MaxEnt Distribution with Entropy Transformation

The simplest transformation of Lagrange functional that leaves the MaxEnt distribution invariant is to consider an arbitrary increasing function of entropy, i.e., we replace $S(P)$ by $c(S(P))$, where $c'(x) > 0$. Let us note that this transform preserves the uniqueness of the MEP because it is easy to show that if $S(P)$ is Schur-concave, $c(S(P))$ is also Schur-concave [42] which is a sufficient condition for uniqueness of the MaxEnt distribution.

In this case, the Lagrange equations are adjusted as follows,

$$c'(S(P))\frac{\partial S(P)}{\partial p_i} - \alpha_c \frac{\partial f_0(P)}{\partial p_i} - \beta_c \frac{\partial \mathcal{E}(P)}{\partial p_i} = 0 \tag{15}$$

leading to

$$\alpha_c = c'(S(P))\langle \nabla_P S(P)\rangle - \beta_c \langle \nabla_P \mathcal{E}(P)\rangle \tag{16}$$

and

$$\beta_c = c'(S(P))\frac{\Delta_i(\nabla_P S(P))}{\Delta_i(\nabla_P \mathcal{E}(P))} \tag{17}$$

so we get that the function c causes *rescaling* of α and β, so

$$\alpha_c = c'(S(P))\,\alpha \tag{18}$$
$$\beta_c = c'(S(P))\,\beta \tag{19}$$

while its ratio remains unchanged, i.e., $\alpha_c/\beta_c = \alpha/\beta$. Actually, the set of increasing functions conform a group of Lagrange multipliers, because it is easy to show that the Lagrange parameters related to the entropy $c_1(c_2(S(P))$

$$\beta_{c_1 \circ c_2} = c_1'(c_2(S(P)) \cdot c_2'(S(P))\,\beta = c_1'(c_2(S(P))\beta_{c_2} \tag{20}$$

which can be described as the group operation $(c_1 \circ c_2) \mapsto c_1'(c_2) \cdot c_2'$.

Exercise: what is the relation between Lagrange multipliers between **Tsallis entropy** $S_q = \frac{1}{1-q}\left(\sum_i p_i^q - 1\right)$ and **Rényi entropy** $R_q = \frac{1}{1-q}\ln \sum_i p_i^q$?

Rényi entropy and Tsallis entropy: *Two most famous examples of generalized entropies are Rényi entropy* $R_q(P) = \frac{1}{1-q}\ln\left(\sum_i p_i^q\right)$ *and Tsallis entropy* $S_q(P) = \frac{1}{1-q}\left(\sum_i p_i^q - 1\right)$. *Their relation can be expressed as*

$$R_q(P) = c_q(S_q(P)) = \frac{1}{1-q}\ln\left[(1-q)S_q(P) + 1\right] \tag{23}$$

and therefore we obtain that

$$c_q'(S_q(P)) = \frac{1}{1+(1-q)S_q} = \frac{1}{\sum_i p_i^q}. \tag{24}$$

The difference between free entropy and α can be obtained as

$$\psi_R - \alpha_R = \frac{1}{\sum_i p_i^q}(\psi_S - \alpha_S) + \left(R_q(P) - \frac{S_q(P)}{\sum_i p_i^q}\right). \tag{25}$$

Summary