BEYOND SHANNON

Generalized entropies and rational inquiry

Ancona

Bayes by the Sea

September 2018

Vincenzo CRUPI

Center for Logic, Language, and Cognition

Department of Philosophy and Education

University of Turin

vincenzo.crupi@unito.it

www.vincenzocrupi.com

COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 42 (2018) 1410–1456

ISSN: 1551-6709 online
DOI: 10.1111/cogs.12613

Generalized Information Theory Meets Human Cognition: Introducing a Unified Framework to Model Uncertainty and Information Search

Vincenzo Crupi,[a] Jonathan D. Nelson,[b,c] Björn Meder,[c] Gustavo Cevolani,[d] Katya Tentori[e]

[a]*Center for Logic, Language, and Cognition, Department of Philosophy and Education, University of Turin*
[b]*School of Psychology, University of Surrey*
[c]*Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development*
[d]*IMT School for Advanced Studies, Lucca*
[e]*Center for Mind/Brain Sciences, University of Trento*

Received 27 January 2017; received in revised form 5 March 2018; accepted 6 March 2018

Abstract

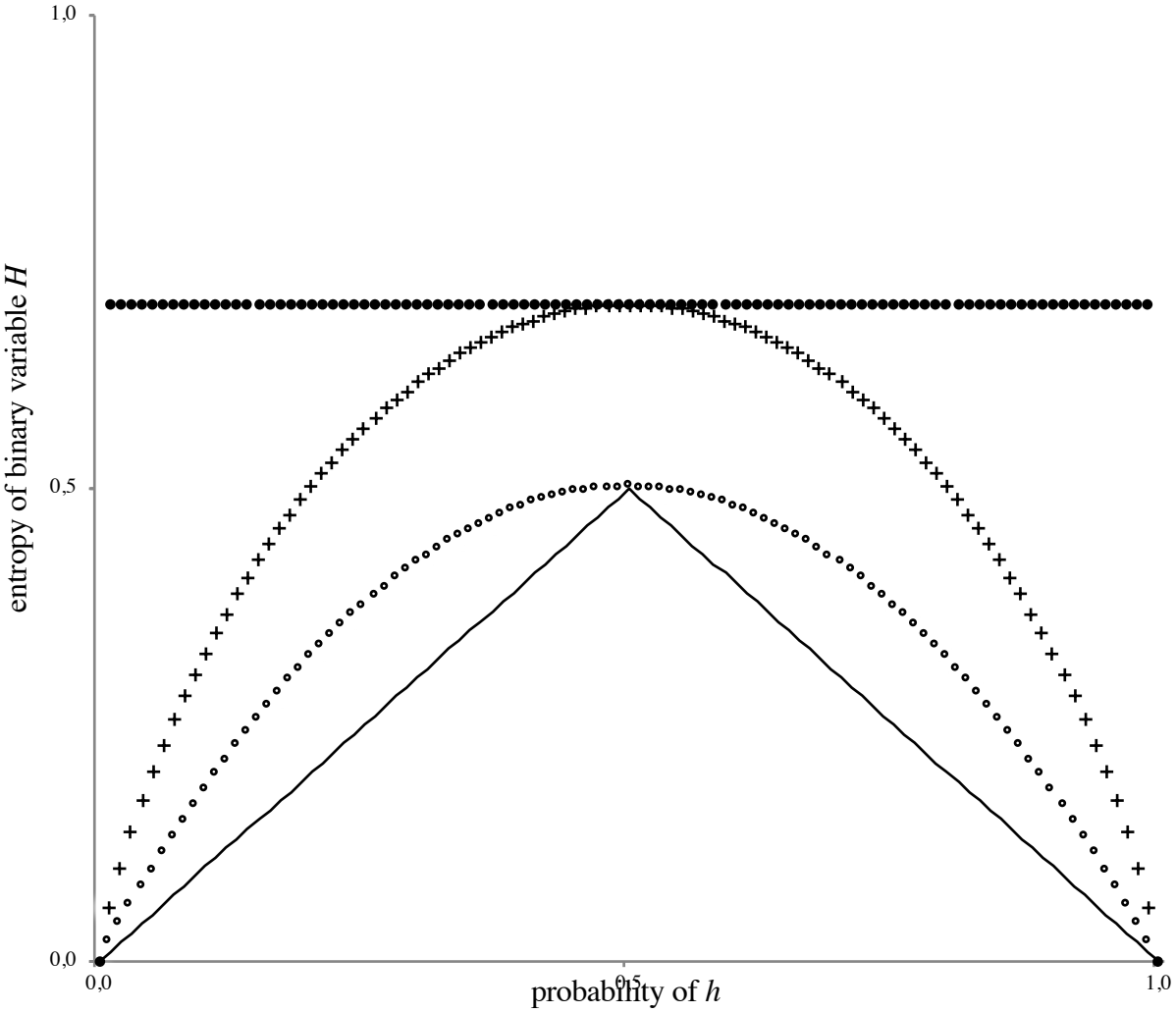
Searching for information is critical in many situations. In medicine, for instance, careful choice of a diagnostic test can help narrow down the range of plausible diseases that the patient might have. In a probabilistic framework, test selection is often modeled by assuming that people's goal is to reduce uncertainty about possible states of the world. In cognitive science, psychology, and medical decision making, Shannon entropy is the most prominent and most widely used model to formalize probabilistic uncertainty and the reduction thereof. However, a variety of alternative entropy metrics (Hartley, Quadratic, Tsallis, Rényi, and more) are popular in the social and the natural sciences, computer science, and philosophy of science. Particular entropy measures have been predominant in particular research areas, and it is often an open issue whether these divergences emerge from different theoretical and practical goals or are merely due to historical accident. Cutting across disciplinary boundaries, we show that several entropy and entropy reduction measures arise as special cases in a unified formalism, the Sharma–Mittal framework. Using mathematical results, computer simulations, and analyses of published behavioral data, we discuss four key questions: How do various entropy models relate to each other? What insights can be obtained by considering diverse entropy models within a unified framework? What is the psychological plausibility of different entropy models? What new questions and insights for research on human information acquisition follow? Our work provides several new pathways for theoretical and empirical research, reconciling apparently conflicting approaches and empirical findings within a comprehensive and unified information-theoretic formalism.

Keywords: Entropy; Uncertainty; Value of information; Information search; Probabilistic models

key rival models from the literature

	entropy measure	expected entropy reduction	key references
quadratic entropy	$1-\sum_{h\in H}P(h)^2$	$\sum_{e\in E}P(e)\left[\sum_{h\in H}P(h\mid e)^2-\sum_{h\in H}P(h)^2\right]$	Gini (1912) Niiniluoto and Toumela (1973) Horwich (1982) Crupi and Tentori (2014)
Hartley entropy	$ln\sum_{h\in H}P(h)^0$	$\sum_{e\in E}P(e)\left[ln\sum_{h\in H}P(h)^0-ln\sum_{h\in H}P(h\mid e)^0\right]$	Hartley (1928) Aczél, Forte, and Ng (1974)
Shannon entropy	$\sum_{h\in H}P(h)\,ln\left(\frac{1}{P(h)}\right)$	$\sum_{e\in E}P(e)\left[\sum_{h\in H}P(h\mid e)\,ln\left(P(h\mid e)\right)-\sum_{h\in H}P(h)\,ln\left(P(h)\right)\right]$	Shannon (1948) Carnap and Bar-Hillel (1953) Oaksford and Chater (1994, 2003) Nelson (2005, 2008) Nelson *et al.* (2010)
error entropy	$1-\max_{h\in H}\left[P(h)\right]$	$\sum_{e\in E}P(e)\left[\max_{h\in H}\left[P(h\mid e)\right]-\max_{h\in H}\left[P(h)\right]\right]$	Fano (1961) Baron (1985) Baron, Beattie, and Hershey (1988) Crupi, Tentori, and Lombardi(2009) Meder and Nelson (2012) Rusconi *et al.* (2014)

key rival models from the literature



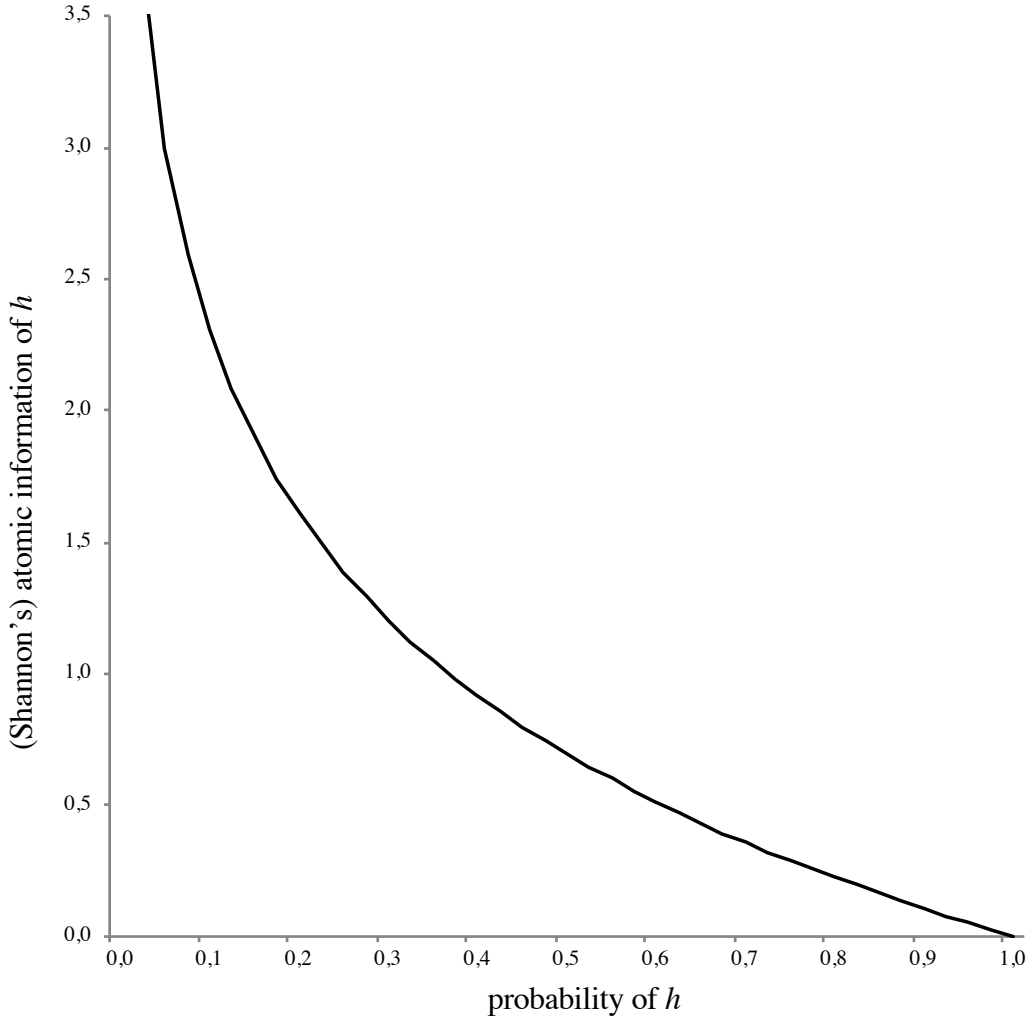
A graphical illustration of quadratic (○), Hartley (•), Shannon (+), and error entropy (solid line) as distinct measures of uncertainty over a binary hypothesis set $H = \{h, \neg h\}$ as a function of the probability of h. (Note that Hartley entropy "jumps" to 0 for extreme values of $P(h)$.)

towards a unified framework

Rényi (1961): entropy
as a (parametric)
generalized mean
of Shannon's original
atomic information values
[i.e., $\ln(1/P(x))$]

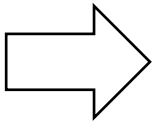
towards a unified framework

Rényi (1961): entropy as a (parametric) generalized mean of Shannon's original atomic information values [i.e., $ln(1/P(x))$]



towards a unified framework

Rényi (1961): entropy as a (parametric) generalized mean of Shannon's original atomic information values [i.e., $ln(1/P(x))$]



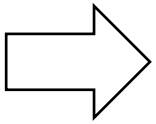
Rényi entropies (r = order):

$$ent_P^{(r)}(H) = \frac{1}{1-r}\, ln\left[\sum_{h \in H} P(h)^r\right]$$

Hartley and Shannon entropy are special cases (for $r = 0, 1$, respectively), but not quadratic and error

towards a unified framework

Rényi (1961): entropy as a (parametric) generalized mean of Shannon's original atomic information values [i.e., $ln(1/P(x))$]



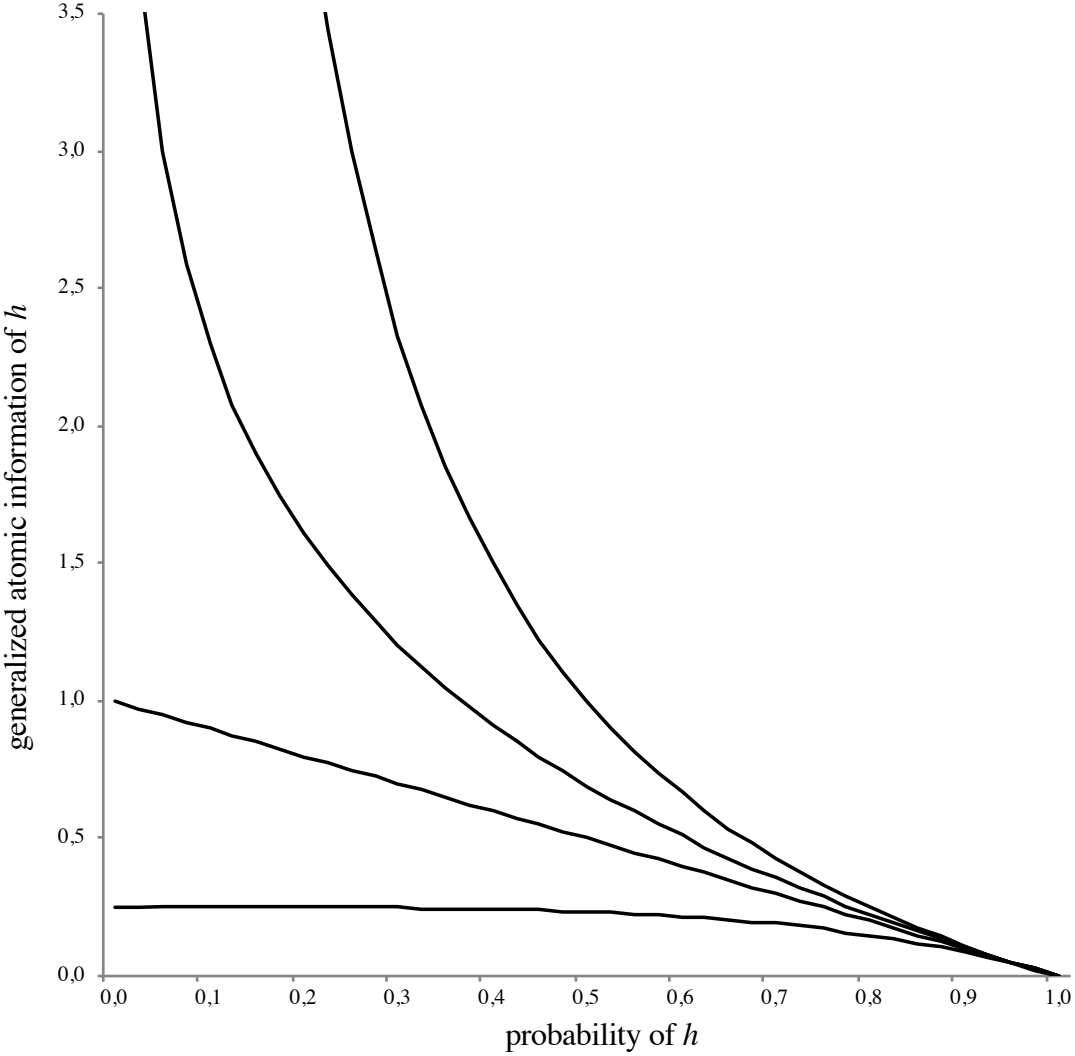
Rényi entropies (r = order):

$$ent_P^{(r)}(H) = \frac{1}{1-r}\ ln\left[\sum_{h\in H} P(h)^r\right]$$

Hartley and Shannon entropy are special cases (for $r = 0, 1$, respectively), but not quadratic and error

Tsallis (1988): standard weighted average of a parametric generalization of Shannon's original atomic information values [→ Tsallis logarithm]

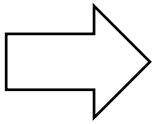
towards a unified framework



A graphical illustration of the generalized atomic information function $ln_t[1/P(h)]$ based on Tsallis's generalized logarithm for four different values of the parameter t (0, 1, 2, and 5, respectively, for the curves from top to bottom).

towards a unified framework

Rényi (1961): entropy as a (parametric) generalized mean of Shannon's original atomic information values [i.e., $ln(1/P(x))$]

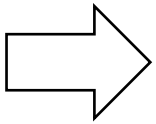


Rényi entropies (r = order):

$$ent_P^{(r)}(H) = \frac{1}{1-r}\ ln\left[\sum_{h\in H} P(h)^r\right]$$

Hartley and Shannon entropy are special cases (for $r = 0, 1$, respectively), but not quadratic and error

Tsallis (1988): standard weighted average of a parametric generalization of Shannon's original atomic information values [→ Tsallis logarithm]



Tsallis entropies (t = degree):

$$ent_P^{(t)}(H) = \frac{1}{t-1}\left[1-\sum_{h\in H} P(h)^t\right]$$

Shannon and quadratic entropy are special cases (for $t = 1, 2$, respectively), but not Hartley and error

towards a unified framework

Rényi (1961): entropy as a (parametric) generalized mean of Shannon's original atomic information values [i.e., $ln(1/P(x))$]

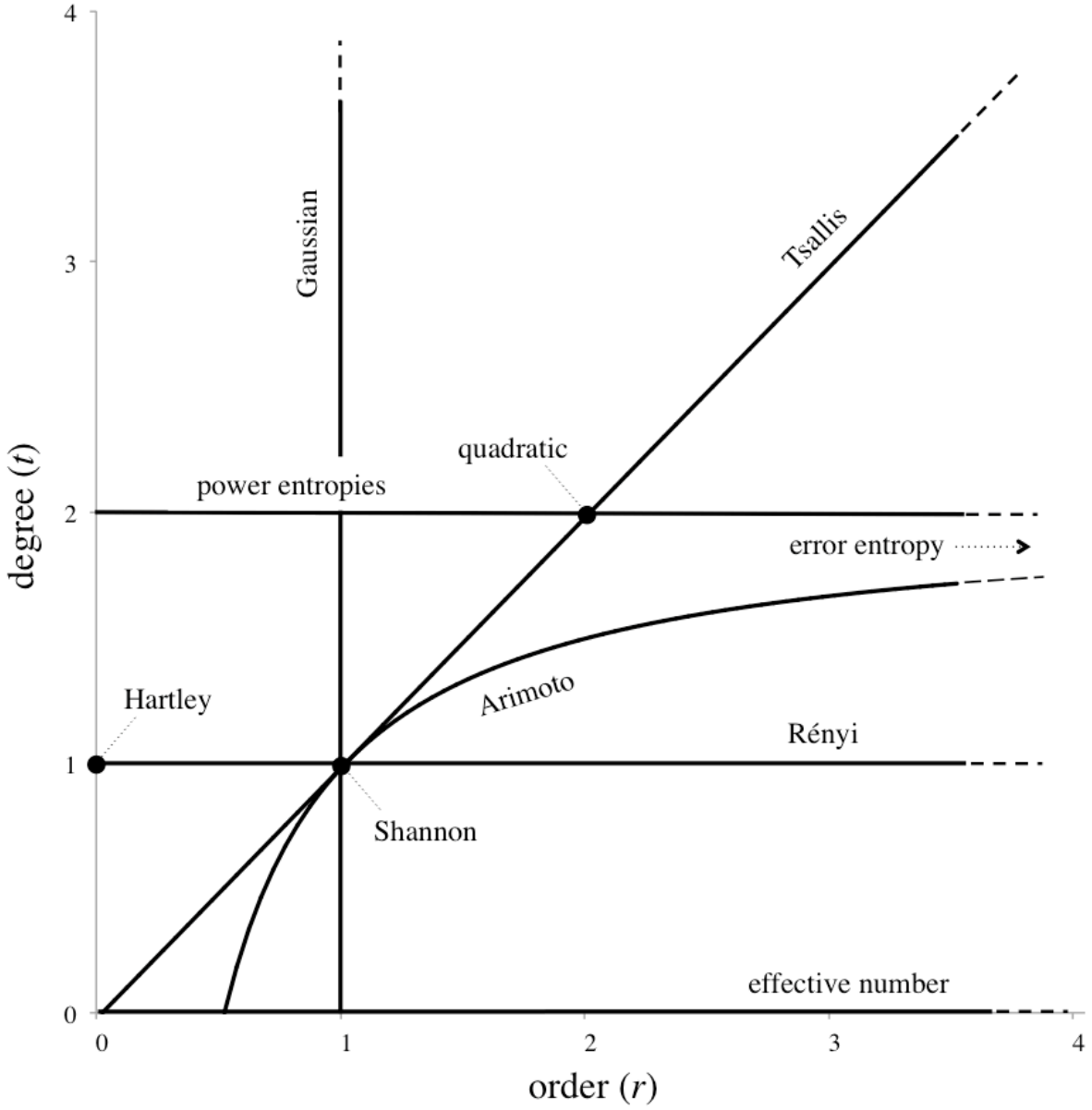
Tsallis (1988): standard weighted average of a parametric generalization of Shannon's original atomic information values [→ Tsallis logarithm]

Sharma-Mittal entropies: biparametric family of measures of order r and degree t

$$ent_P^{(r,t)}(H) = \frac{1}{t-1}\left[1-\left(\sum_{h\in H} P(h)^r\right)^{\frac{t-1}{r-1}}\right]$$

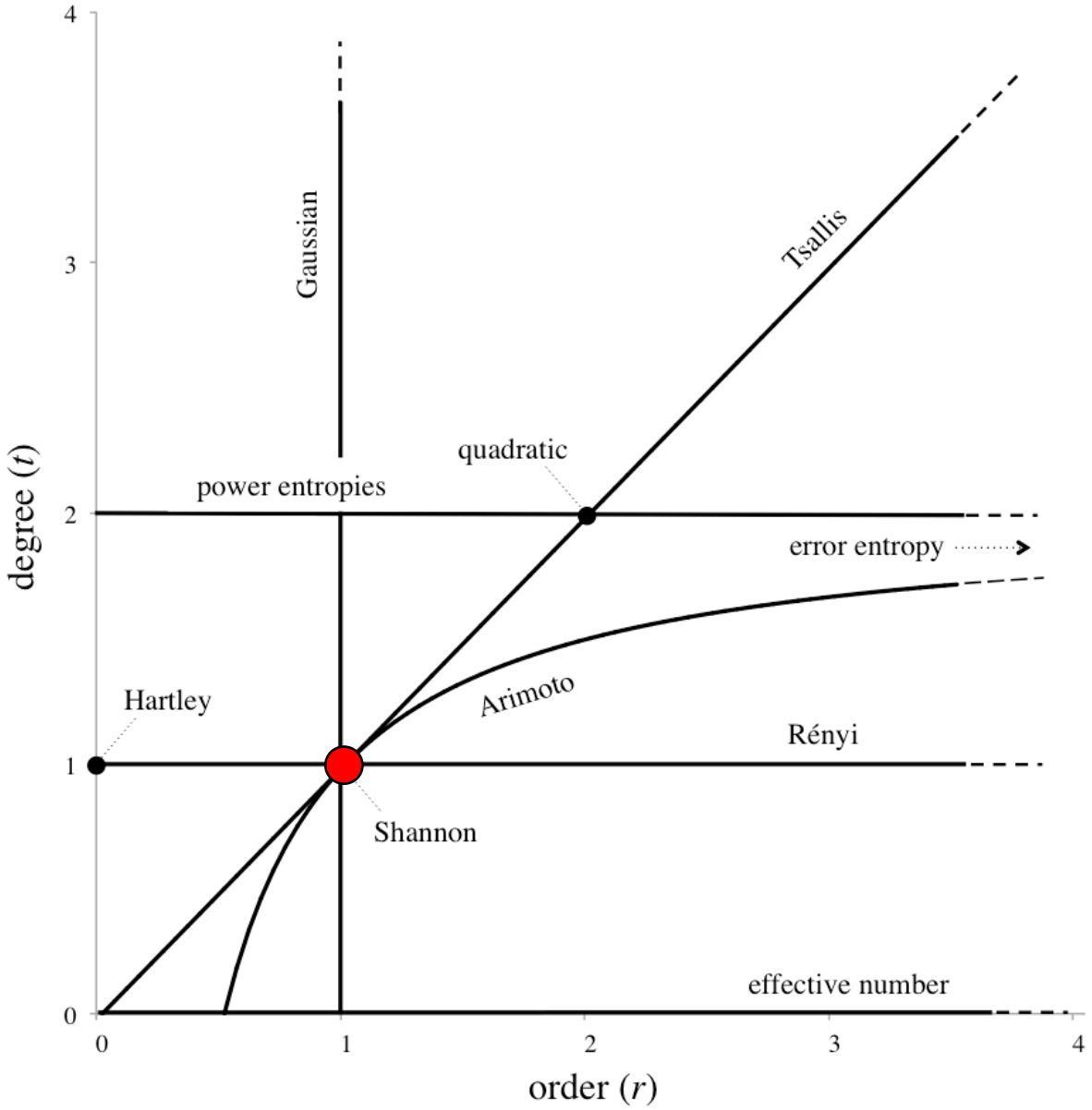
$$ent_P^{(r,t)}(H) = ln_t e_r\left[\sum_{h\in H} P(h)\ ln_r\left(\frac{1}{P(h)}\right)\right]$$

towards a unified framework



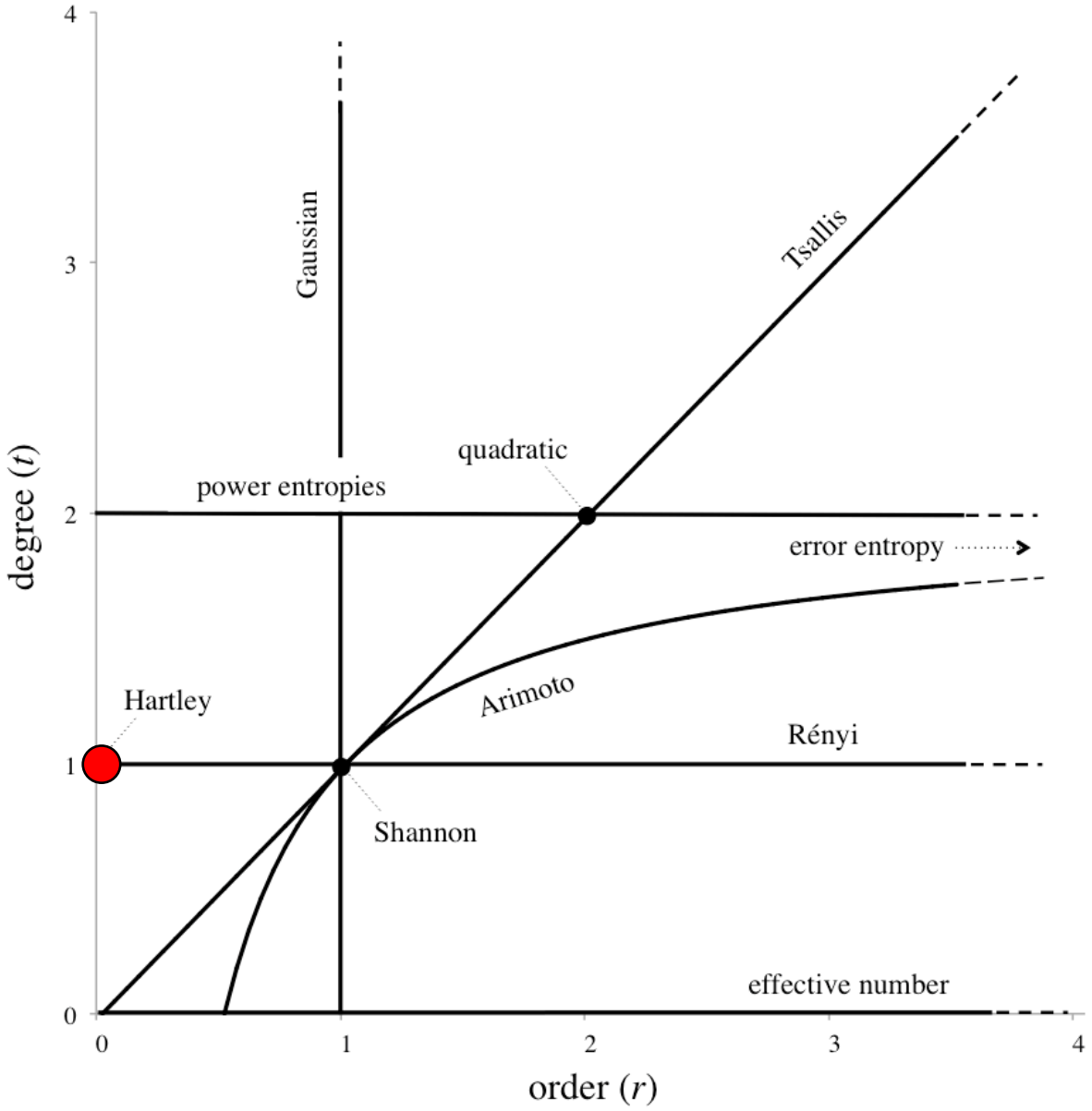
The Sharma-Mittal family of entropy measures represented in a Cartesian quadrant with values of the degree parameter *t* and of the order parameter *r* lying on the *y* and *x* axes, respectively. Each point in the quadrant corresponds to a specific entropy measure, each line corresponds to a distinct one-parameter generalized entropy function, and various special cases are highlighted.

towards a unified framework



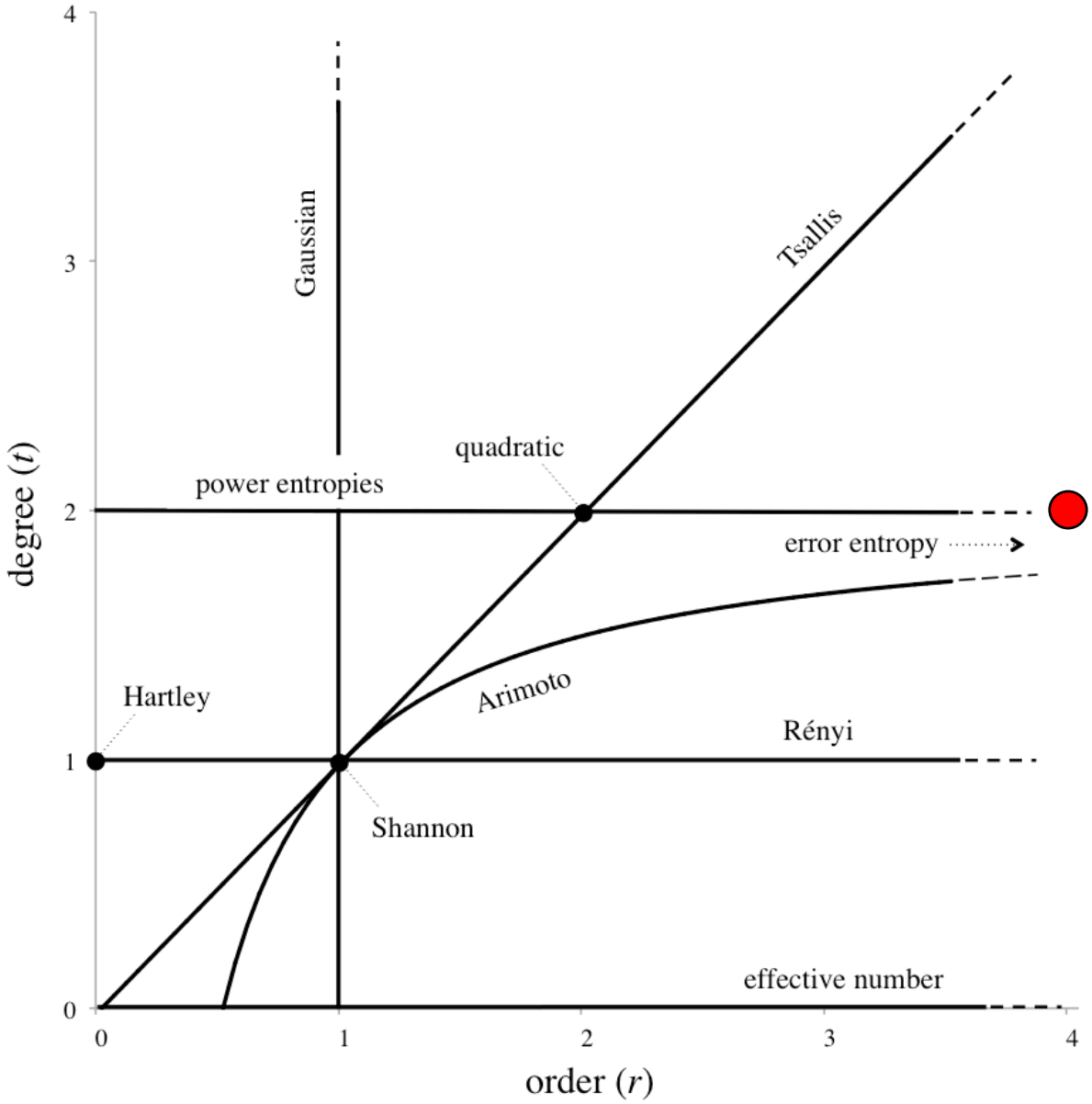
The Sharma-Mittal family of entropy measures represented in a Cartesian quadrant with values of the degree parameter *t* and of the order parameter *r* lying on the *y* and *x* axes, respectively. Each point in the quadrant corresponds to a specific entropy measure, each line corresponds to a distinct one-parameter generalized entropy function, and various special cases are highlighted.

towards a unified framework



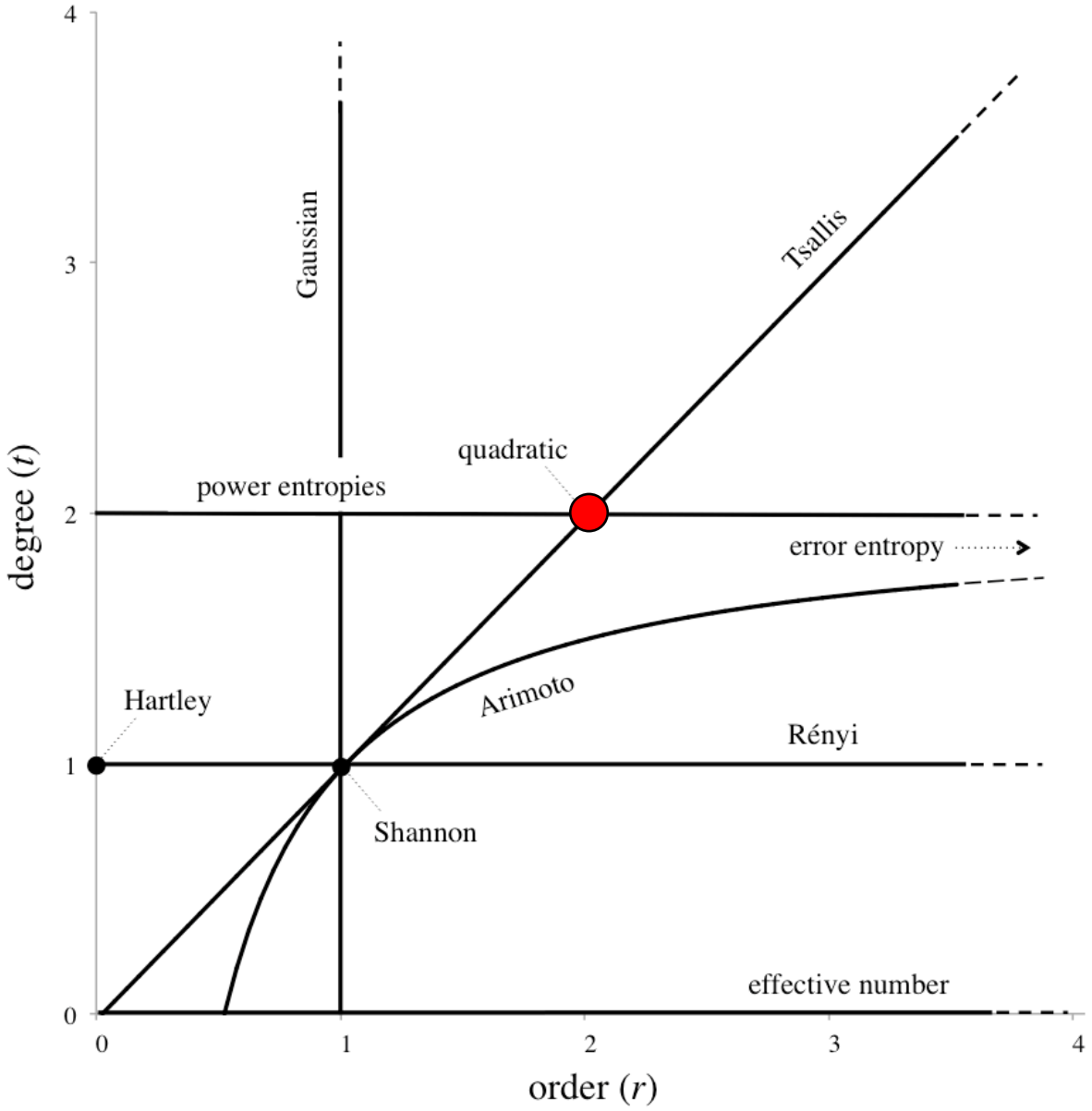
The Sharma-Mittal family of entropy measures represented in a Cartesian quadrant with values of the degree parameter *t* and of the order parameter *r* lying on the *y* and *x* axes, respectively. Each point in the quadrant corresponds to a specific entropy measure, each line corresponds to a distinct one-parameter generalized entropy function, and various special cases are highlighted.

towards a unified framework



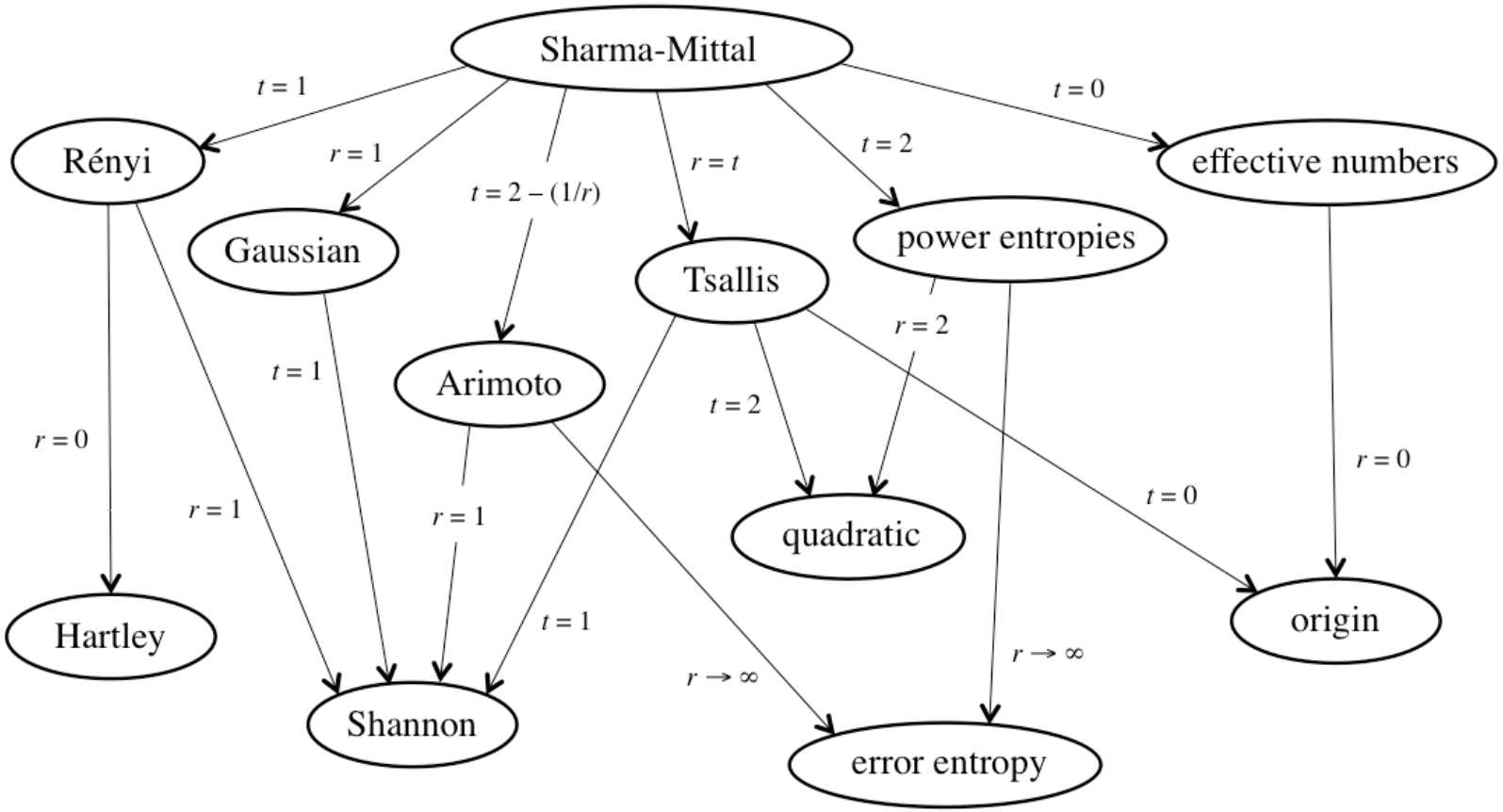
The Sharma-Mittal family of entropy measures represented in a Cartesian quadrant with values of the degree parameter *t* and of the order parameter *r* lying on the *y* and *x* axes, respectively. Each point in the quadrant corresponds to a specific entropy measure, each line corresponds to a distinct one-parameter generalized entropy function, and various special cases are highlighted.

towards a unified framework



The Sharma-Mittal family of entropy measures represented in a Cartesian quadrant with values of the degree parameter *t* and of the order parameter *r* lying on the *y* and *x* axes, respectively. Each point in the quadrant corresponds to a specific entropy measure, each line corresponds to a distinct one-parameter generalized entropy function, and various special cases are highlighted.

towards a unified framework



A map of how different entropy measures are derived within the Sharma-Mittal framework for different settings of the order (*r*) and degree (*t*) parameters.

two odd implications of Shannon entropy

(i) irrelevant combination

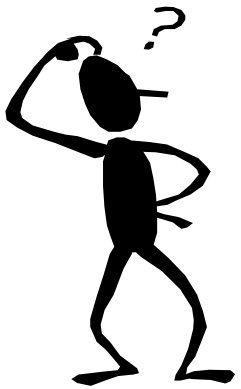
ex.: A card was drawn from a well-shuffled stardard deck and kept hidden. H = {2, 3, …, king, ace} is the set of (thirteen) hypotheses concerning the value of the card drawn, while K = {hearts, diamonds, clubs, spades} is the set of (four) hypotheses concerning its suit. So knowing the true value of the combined variable $H \times K$ (value *plus* suit) amounts to knowing exactly which card has been drawn out of the whole deck of fifty-two. (We assume H and K probabilistically independent.) You can ask about H.

two odd implications of Shannon entropy

(i) irrelevant combination

ex.: A card was drawn from a well-shuffled stardard deck and kept hidden. H = {2, 3, …, king, ace} is the set of (thirteen) hypotheses concerning the value of the card drawn, while K = {hearts, diamonds, clubs, spades} is the set of (four) hypotheses concerning its suit. So knowing the true value of the combined variable $H \times K$ (value *plus* suit) amounts to knowing exactly which card has been drawn out of the whole deck of fifty-two. (We assume H and K probabilistically independent.) You can ask about H.

Shannon: $R_P^S(H,H) = ln(13) = R_P^S(H \times K,H)$



two odd implications of Shannon entropy

(i) irrelevant combination

ex.: A card was drawn from a well-shuffled stardard deck and kept hidden. $H = \{2, 3, \ldots, \text{king}, \text{ace}\}$ is the set of (thirteen) hypotheses concerning the value of the card drawn, while $K = \{\text{hearts}, \text{diamonds}, \text{clubs}, \text{spades}\}$ is the set of (four) hypotheses concerning its suit. So knowing the true value of the combined variable $H \times K$ (value *plus* suit) amounts to knowing exactly which card has been drawn out of the whole deck of fifty-two. (We assume H and K probabilistically independent.) You can ask about H.

Shannon: $R_P^S(H,H) = \ln(13) = R_P^S(H \times K, H)$

quadratic: $R_P^Q(H,H) = 0.92 > 0.23 = R_P^Q(H \times K, H)$

two odd implications of Shannon entropy

(i) irrelevant combination

(ii) commutativity

ex.: A card was drawn from a well-shuffled stardard deck and kept hidden.
H = {red, black}, while K = {hearts, diamonds, clubs, spades}

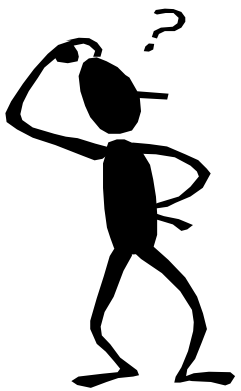
two odd implications of Shannon entropy

(i) irrelevant combination

(ii) commutativity

ex.: A card was drawn from a well-shuffled stardard deck and kept hidden.
H = {red, black}, while K = {hearts, diamonds, clubs, spades}

Shannon: $R_P^S(H,K) = ln(2) = R_P^S(K,H)$



two odd implications of Shannon entropy

(i) irrelevant combination

(ii) commutativity

ex.: A card was drawn from a well-shuffled stardard deck and kept hidden. H = {red, black}, while K = {hearts, diamonds, clubs, spades}

Shannon: $R_P^S(H,K) = ln(2) = R_P^S(K,H)$

quadratic: $R_P^Q(H,K) = 0.75 > 0.19 = R_P^Q(K,H)$

entropy, additivity, and concavity

(i) additivity *of what*?

for all and only Rényi entropies (i.e., SM entropies of degree $t = 1$):

$$ent_P(H \times K) = ent_P(H) + ent_P(K) \qquad \text{if } H \perp_P K$$

while for degree $t = 2$ (including quadratic):

$$ent_P(H \times K) = ent_P(H) + ent_P(K) - ent_P(H)ent_P(K) \qquad \text{if } H \perp_P K$$

but for ALL SM entropies (of any degree t and order r):

$$R_P(H, E \times F) = R_P(H, E) + R_P(H, F) \qquad \text{if } E \perp_P F \mid H$$

entropy, additivity, and concavity

(i) additivity *of what*?

(ii) order *vs.* concavity?

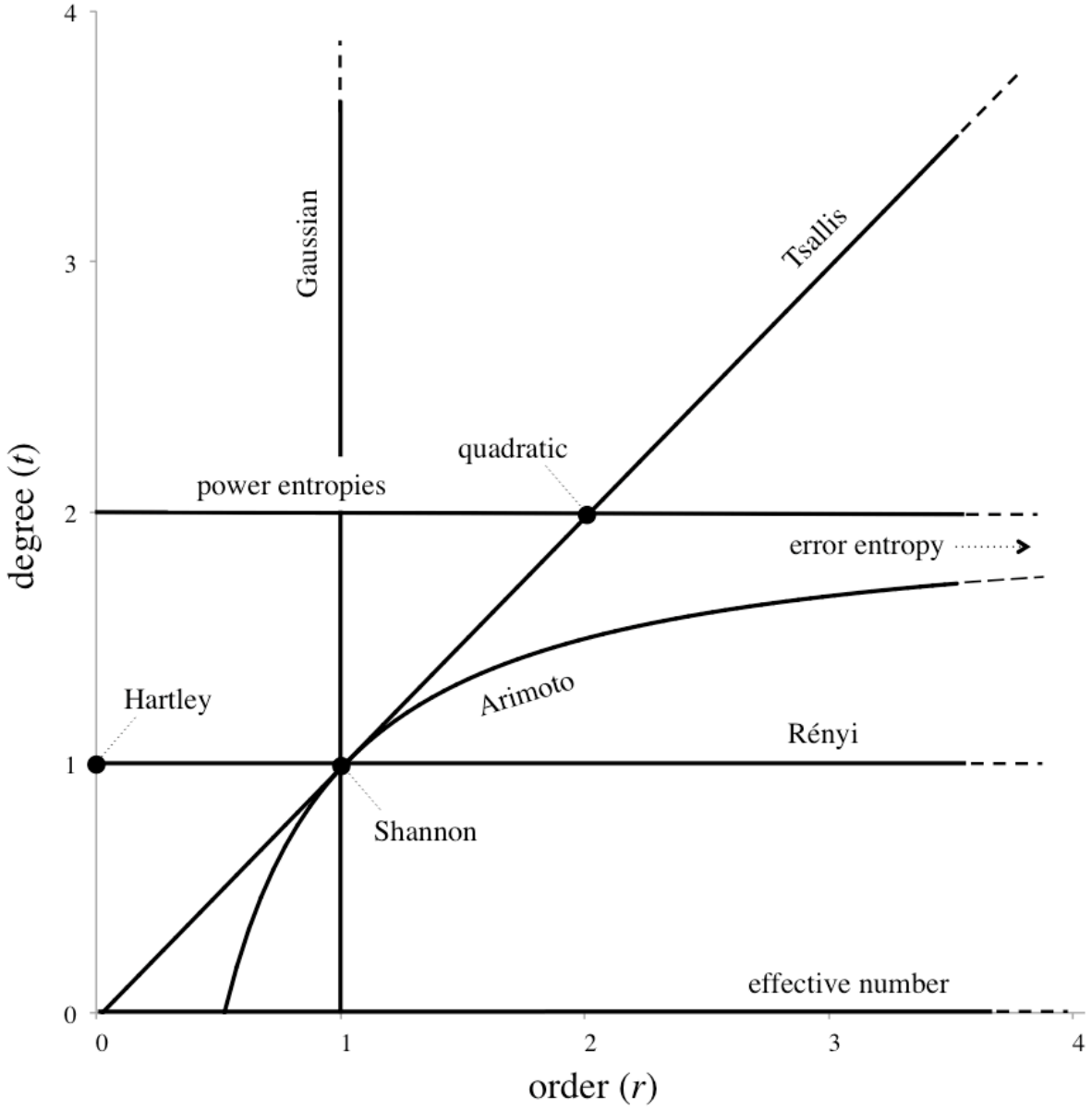
order (*r*) is an index of the *imbalance* of the entropy function:
it indicates how much the entropy measure discounts minor hypotheses

order-0 measures → the actual probability distribution is neglected:
live hypothesis are just counted, as if they were all equally important

order-∞ measures → only the most likely hypothesis matters,
and all minor hypotheses are disregarded altogether

intermediate values of *r* → more likely hypotheses count more,
but less likely hypotheses do retain some weight: the higher (lower) *r* is,
the more (less) it so happens that more likely hypotheses are regarded
and less likely hypotheses are discounted

entropy, additivity, and concavity



The Sharma-Mittal family of entropy measures represented in a Cartesian quadrant with values of the degree parameter *t* and of the order parameter *r* lying on the *y* and *x* axes, respectively. Each point in the quadrant corresponds to a specific entropy measure, each line corresponds to a distinct one-parameter generalized entropy function, and various special cases are highlighted.

entropy, additivity, and concavity

(i) additivity *of what*?

(ii) order *vs.* concavity?

but most Rényi entropies ($r > 1$) are non-concave,
thus allowing for *informationally detrimental* experiments!

ex:

prior $P(H) = \{0.63, 0.185, 0.185\}$

$P(\text{e}) = 0.44$ $\quad P(H|\text{e}) = \{1, 0, 0\}$
$P(\text{not-e}) = 0.56$ $\quad P(H|\text{not-e}) = \{0.33\ldots, 0.33\ldots, 0.33\ldots\}$

$P(f) = 0.9$ $\quad P(H|f) = \{0.63, 0.185, 0.185\}$
$P(\text{not-}f) = 0.1$ $\quad P(H|\text{not-}f) = \{0.63, 0.185, 0.185\}$

which test is most useful, $E = \{\text{e}, \text{not-e}\}$ or $F = \{f, \text{not-}f\}$?

entropy, additivity, and concavity

(i) additivity *of what*?

(ii) order *vs.* concavity?

but most Rényi entropies ($r > 1$) are non-concave,
thus allowing for *informationally detrimental* experiments!

order WITH concavity → *power entropies* (quadratic for order $r = 2$)

$$ent_P^{(r)}(H) = 1 - \left(\sum_{h \in H} P(h)^r \right)^{\frac{1}{r-1}}$$

entropy, surprise, and evidential support

Shannon and quadratic entropy of H amount
to the *expected surprise* of finding out the true element of H,
with surprise $= ln_t[1/P(h)]$, for $t = 1$ and $t = 2$, respectively

suprise $\rightarrow$ probababilistic evidential support as *reduction of surprise*:

$$C_P(h, e) = ln_t\left(\frac{1}{P(h)}\right) - ln_t\left(\frac{1}{P(h \mid e)}\right)$$

$t = 1 \rightarrow$ log probability ratio measure

$t = 2 \rightarrow$ probability *difference* measure

entropy, inaccuracy, and divergence

Shannon and quadratic entropy of H also amount to the *expected inaccuracy* of P, with

$$Inacc(P, w_i) = t \cdot ln_t\left(\frac{1}{P(h_i)}\right) - ln_t\left(\frac{1}{\sum_{h \in H} P(h)^t}\right)$$

for $t = 1$ and $t = 2$, respectively

$t = 1 \rightarrow$ logarithmic score

$t = 2 \rightarrow$ Brier score

if *divergence* = reduction (difference) of expected inaccuracy
from a posterior perspective (e.g., Roche & Shogenji),
then $t = 1 \rightarrow$ KL-divergence, while $t = 2 \rightarrow$ squared Euclidean distance
and their expectations are identical to the expected reduction of Shannon and quadratic entropy, respectively

conclusions: with quadratic instead of Shannon…

- *irrelevant combination* and *communtativity* problems are solved…
- additive behaviour of (independent) experiments with regards to a target hypothesis space is retained…
- a better behaved (because uniformly concave) continuum of entropy measures of varying order is obtained…
- a connection with a much better behaved probabilistic measure of evidential support is established…
- and the advantages of a close link with sound (in)accuracy and divergence measures are also preserved