Proceedings of the 2007 IEEE Symposium on
Foundations of Computational Intelligence (FOCI 2007)

# Nonextensive Variational Principles for Rate Distortion Theory and the Information Bottleneck Method

R. C. Venkatesan
Systems Research Corporation
Aundh, Pune 411007, India
ravi@systemsresearchcorp.com

## Abstract

*Variational principles for the rate distortion (RD) theory, and the Information Bottleneck (IB) method, are formulated within the ambit of the generalized nonextensive statistics of Tsallis. Numerical schemes for the nonextensive RD theory and IB method are derived. The physical implications of using nonextensive statistics vis-á-vis Boltzmann-Gibbs statistics are exemplified with the aid of numerical simulations.*

## 1. Introduction

The generalized (nonextensive) statistics of Tsallis [1,2] has recently been the focus of much attention in statistical physics, and allied disciplines[1]. Nonextensive statistics has found applications in a wide spectrum of disciplines ranging from condensed matter physics to financial mathematics. A continually updated bibliography of works related to nonextensive statistics may be found at http://tsallis.cat.cbpf.br/biblio.htm.

Nonextensive statistics generalizes the extensive Boltzmann-Gibbs statistics, and has much utility in complex systems. Some of features possessed by complex systems, which invite the use of nonextensive statistics are long range correlations, fluctuations, ergodicity, chirality and fractal behavior, amongst others. By definition, the Tsallis entropy is defined in terms of discrete variables as

$$S_q\left(x\right) = \frac{1 - \sum\limits_x p^q\left(x\right)}{1-q}, where, \sum_x p\left(x\right) = 1. \quad (1)$$

The constant $q$ is referred to as the *nonextensivity parameter*. Given two independent variables $x$ and $y$, one of the fundamental consequences of nonextensivity is demonstrated by the *pseudo-additivity* relation

$$S_q\left(xy\right) = S_q\left(x\right) + S_q\left(y\right) + \left(1-q\right) S_q\left(x\right) S_q\left(y\right). \quad (2)$$

Here, (1) and (2) imply that extensive statistics is recovered as $q \to 1$. Taking the limit $q \to 1$ in (1) and evoking l'Hospital's rule, $S_q\left(x\right) \to S\left(x\right)$, the Shannon entropy.

The jointly convex generalized Kullback-Leibler divergence (K-Ld) is of the form [3]

$$I_q\left(p\left(x\right) \| q\left(x\right)\right) = \sum_x p\left(x\right) \frac{\left(\frac{p(x)}{r(x)}\right)^{q-1} - 1}{q-1}. \quad (3)$$

In the limit $q \to 1$, the extensive K-Ld is readily recovered. Akin to the Tsallis entropy, the generalized K-Ld obeys the *pseudo-additivity* relation [3].

Rate distortion (RD) theory constitutes one of the cornerstones of contemporary information theory [4, 5]. RD theory has found applications in diverse disciplines, which include data compression and clustering. Deterministic annealing (DA) [6, 7] and the information bottleneck (IB) method [8] are two influential paradigms in machine learning, that are closely related to RD theory. The representation of RD theory in the form of a variational principle, expressed within the Boltzmann-Gibbs-Shannon framework, has been established (see Chapter 13 of [4]).

The seminal paper on source coding within the framework of nonextensive statistics by Landsberg and Vedral [9], has provided the impetus for a number of investigations into the use of nonextensive information theory within the context of coding problems. The works of Yamano [10,11] represent a sample of some of the prominent efforts in this regard.

The objective of this paper is to re-formulate RD theory and the IB method within the ambit of nonextensive statistics. Further, numerical algorithms derived on the basis of generalized statistics are presented for the nonextensive RD

[1] In this paper the terms generalized and nonextensive are used interchangeably

theory and IB method. For the RD theory, the numerical algorithm modifies the well known Blahut-Arimoto scheme [12].

One of the noteworthy results of the IB method is the self-consistent derivation of the distortion measure from the joint statistics of the source distribution $X$ and the relevance variable $Y$ [8]. Section 4 of this paper tacitly demonstrates that this result carries over into the nonextensive regime. Finally, the distinctions between extensive *vis-á-vis* nonextensive statistics are highlighted with the aid of simple, but qualitatively revealing, numerical simulations.

## 2. Select Theory in Generalized Statistics

### 2.1 Selected results from q-algebra

Generalized statistical mechanics often utilizes results from *q-algebra*, in order to derive equations that resemble their counterparts obtained from Boltzmann-Gibbs statistics. Herein, selected results utilized in this paper are described. The Tsallis entropy (1), and, the generalized K-Ld (3) may be written as

$$S_q\left(p\left(x\right)\right) = -\sum_x p\left(x\right)^q \ln_q p\left(x\right), \tag{4}$$

and,

$$I_q\left(p\left(x\right)\left\|r(x)\right.\right) = -\sum_x p\left(x\right)\ln_q\frac{r(x)}{p(x)}, \tag{5}$$

respectively. Here, the *q-deformed* logarithm is

$$ln_q\left(x\right) = \frac{x^{1-q}-1}{1-q}. \tag{6}$$

In this paper, $0 < q < 1$. Similarly, the *q-deformed* exponential is described by

$$e_q^x = \begin{cases} \left[1+(1-q)\,x\right]^{1/(1-q)} & ; 1+(1-q)\,x \geq 0 \\ 0; otherwise & \end{cases} \tag{7}$$

It is important to note that the results of *q-algebra* are not constructions specifically manufactured to express results obtained from generalized statistics, in a form analogous to their counterparts in extensive statistics. The results of *q-algebra* have a long lineage, and the existence of certain results (specifically the *q-deformed* exponential) were known to Leibniz [13]. Some of the salient results of *q-algebra*, that are employed in this paper are

$$\begin{array}{l} \ln_q\left(xy\right) = \ln_q\left(x\right)\oplus_q \ln_q\left(x\right) \\ = \ln_q\left(x\right) + \ln_q\left(y\right) + \left(1-q\right)\ln_q\left(x\right)\ln_q\left(y\right), \\ \ln_q\left({}^{x}\!/_{y}\right) = \ln_q\left(x\right)\ominus_q \ln_q\left(x\right) \\ where, \ominus_q y = \frac{-y}{1+(1-q)y)}. \end{array} \tag{8}$$

### 2.2 Constraints

Generalized statistics has utilized a number of constraints to define expectations. The original Tsallis (OT) constraints of the form $\langle A\rangle = \sum\limits_i p_i A_i$ [1], were convenient owing to their similarity to the maximum entropy constraints. These were abandoned because of difficulties encountered in obtaining an acceptable form for the partition function. The OT constraints were subsequently replaced by the Curado-Tsallis (C-T) [14] constraints $\langle A\rangle_q = \sum\limits_i p_i^q A_i$. The C-T constraints were later replaced by the normalized Tsallis-Mendes-Plastino (T-M-P) constraints [15] $\langle A\rangle_q = \sum\limits_i \frac{p_i^q}{\aleph_q(x)} A_i; \aleph_q\left(x\right) = \sum\limits_i p_i^q$. The dependence of the expectation value on the normalization *pdf* $\aleph_q\left(x\right)$, rendered the T-M-P constraints to be *self-referential*.

A recent work by Ferri, Martinez, and Plastino [16] has described a methodology to "rescue" the OT constraints, and, has linked the OT, C-T, and, T-M-P constraints. The present paper utilizes a procedure that is closely related to the work of Ferri, Martinez, and Plastino [16]. This has enabled the nonextensive variational principles for the RD theory and the IB method to be cast in a manner that closely parallels the extensive case.

## 3 Nonextensive RD Variational Principle

Let $X \in \Xi$ be a discrete random variable, distributed according to the marginal *pdf* $p\left(x\right)$. Let $\tilde{X} \in \tilde{\Xi}$ be another discrete random variable, distributed according to the marginal *pdf* $p\left(\tilde{x}\right)$. Here, $\tilde{X}$ is a compressed representation (*quantized codebook*) of $X$, defined through a (possibly stochastic) mapping between each value $x \in \Xi$ to a representative value $\tilde{x} \in \tilde{\Xi}$. This mapping is characterized by the conditional probability $p\left(\tilde{x}\left|x\right.\right)$, which induces a soft partitioning (assignment) of the $X$ discrete random variables. Each $x \in \Xi$ relates to all $\tilde{x} \in \tilde{\Xi}$ through a normalized conditional *pdf* conditional probability $p\left(\tilde{x}\left|x\right.\right)$.

Consider the nonextensive RD Lagrangian

$$L_{RD}^q\left[\tilde{x}\left|x\right.\right] = I_q\left(X;\tilde{X}\right) + \beta\left\langle d\left(x,\tilde{x}\right)\right\rangle_{p(x,\tilde{x})}, \tag{9}$$

subject to the normalization of the conditional probability $p\left(\tilde{x}\left|x\right.\right)$. The distortion measure is denoted by $d(x,\tilde{x})$, which is taken to be the Euclidean square distance for most problems in science and engineering [7]. Here, $I_q(X;\tilde{X})$ is the generalized mutual information defined by

$$\begin{aligned} I_q\left(X;\tilde{X}\right) &= -\sum_{x,\tilde{x}} p\left(x,\tilde{x}\right)\ln_q\left(\tfrac{p(x)p(\tilde{x})}{p(x)p(\tilde{x}|x)}\right) \\ &= \sum_{x,\tilde{x}} \frac{p(x)p(\tilde{x}|x)\left(\frac{p(\tilde{x}|x)}{p(\tilde{x})}\right)^{q-1}-1}{q-1}. \end{aligned} \tag{10}$$

Defining $R_q\left(D\right) = \min\limits_{p(\tilde{x}|x):\langle d(x,\tilde{x})\rangle_{p(x,\tilde{x})}\le D} I_q\left(X;\tilde{X}\right)$,

$$\begin{aligned} &\delta R_q\left(D\right) = \delta I_q\left(X;\tilde{X}\right) + \beta\delta\left\langle d\left(x,\tilde{x}\right)\right\rangle_{p(x,\tilde{x})} = 0 \\ &\Rightarrow \tfrac{\delta I_q\left(X;\tilde{X}\right)}{\delta\langle d(x,\tilde{x})\rangle_{p(x,\tilde{x})}} = -\beta. \end{aligned} \tag{11}$$

The joint *pdf* $p\left(x,\tilde{x}\right)$ is taken to be normalized, thus, $\sum\limits_{x,\tilde{x}} p\left(x,\tilde{x}\right) = 1$. Expanding $p\left(x,\tilde{x}\right) = p\left(x\right)p\left(\tilde{x}\left|x\right.\right)$, (11) acquires the form

$$\begin{aligned} &L_{RD}^q\left[\tilde{x}|\,x\right] = \\ &\sum_{x,\tilde{x}} \frac{p(x)p(\tilde{x}|x)\left(\frac{p(\tilde{x}|x)}{p(\tilde{x})}\right)^{q-1}-1}{(q-1)} + \beta\sum_{x,\tilde{x}} d\left(x,\tilde{x}\right)p\left(x\right)p\left(\tilde{x}|\,x\right) + \\ &+\sum_{x}\lambda\left(x\right)\sum_{\tilde{x}} p\left(\tilde{x}|\,x\right). \end{aligned} \tag{12}$$

Defining $p\left(\tilde{x}\right) = \sum\limits_{x} p\left(x\right)p\left(\tilde{x}\left|x\right.\right)$, and noting that $\frac{\delta p(\tilde{x})}{\delta p(\tilde{x}|x)} = p\left(x\right)$, the variational derivative of (12) yields

$$\begin{aligned} &\tfrac{\delta L_{RD}^q[\tilde{x}|x]}{\delta p(\tilde{x}|x)} = \\ &= p\left(x\right)\left[\tfrac{q}{q-1}\left(\tfrac{p(\tilde{x}|x)}{p(\tilde{x})}\right)^{q-1} + \beta d\left(x,\tilde{x}\right) + \tilde{\lambda}\left(x\right)\right] = 0. \end{aligned} \tag{13}$$

It is important to note that conditional probabilities first acquired prominence in nonextensive statistics owing to their utility in characterizing *quantum entanglement*. This was accomplished in the seminal work of Abe and Rajagopal [17]. In (13) the scaled Lagrange multiplier $\tilde{\lambda}\left(x\right) = \frac{\lambda(x)}{p(x)} - p\left(x\right)^{(1-q)}$, evaluated for each value of $x$, enforces the normalization of the conditional probability . *Note that for both the RD theory and the IB method, the normalization Lagrange multiplier acts as a "reservoir" into which terms independent of $\tilde{x}$ may be absorbed.* Expanding (13), yields

$$p\left(\tilde{x}\left|x\right.\right) = p\left(\tilde{x}\right)\left[\frac{(1-q)}{q}\left\{\tilde{\lambda}\left(x\right) + \beta d\left(x,\tilde{x}\right)\right\}\right]^{1/(q-1)}. \tag{14}$$

The scaled normalization Lagrange multiplier is obtained as follows. Multiplying the terms in the square bracket in (13) by the conditional probability $p\left(\tilde{x}|\,x\right)$, and summing over $\tilde{x}$ yields

$$\begin{aligned} &\tfrac{q}{q-1}\sum_{\tilde{x}} p\left(\tilde{x}\right)\left(\tfrac{p(\tilde{x}|x)}{p(\tilde{x})}\right)^{q} + \\ &+\underbrace{\beta\sum_{\tilde{x}} d\left(x,\tilde{x}\right)p\left(\tilde{x}|\,x\right)}_{\beta\langle d(x,\tilde{x})\rangle_{p(\tilde{x}|X=x)}} + \tilde{\lambda}\left(x\right)\sum_{\tilde{x}} p\left(\tilde{x}|\,x\right) = 0. \end{aligned} \tag{15}$$

Evoking the normalization condition for the conditional *pdf* with the source sample fixed at $X = x$, i.e. $\sum\limits_{\tilde{x}} p\left(\tilde{x}|\,X = x\right) = 1$, yields

$$\tilde{\lambda}\left(x\right) = \frac{q}{(1-q)}\aleph_q\left(x\right) - \beta\left\langle d\left(x,\tilde{x}\right)\right\rangle_{p(\tilde{x}|X=x)}. \tag{16}$$

Here, $\aleph_q\left(x\right) = \sum\limits_{\tilde{x}} p\left(\tilde{x}\right)\left(\frac{p(\tilde{x}|x)}{p(\tilde{x})}\right)^{q}$. The conditional *pdf* $p\left(\tilde{x}|\,x\right)$ acquires the form

$$p\left(\tilde{x}|\,x\right) = p\left(\tilde{x}\right)\left\{\aleph_q\left(x\right) + \frac{(1-q)}{q}\beta\Delta d\left(x,\tilde{x}\right)\right\}^{1/(q-1)}, \tag{17}$$

where, $\Delta d\left(x,\tilde{x}\right) = \left\{d\left(x,\tilde{x}\right) - \left\langle d\left(x,\tilde{x}\right)\right\rangle_{p(\tilde{x}|X=x)}\right\}$. Rearranging the terms in (17) yields

$$\begin{aligned} p\left(\tilde{x}|\,x\right) &= \frac{p(\tilde{x})\left\{1-\frac{(q-1)\beta d(x,\tilde{x})}{q\aleph_q(x)+(q-1)\beta\langle d(x,\tilde{x})\rangle_{p(\tilde{x}c|X=x)}}\right\}^{1/(q-1)}}{\left[q\aleph_q(x)+(q-1)\beta\langle d(x,\tilde{x})\rangle_{p(\tilde{x}|X=x)}\right]^{1/(1-q)}}, \\ &= \frac{p(\tilde{x})\{1-(q-1)\beta^* d(x,\tilde{x})\}^{1/(q-1)}}{\Im_{RD}^{1/(1-q)}}, \end{aligned} \tag{18}$$

where the *effective inverse temperature* $\beta^* = \frac{\beta}{q\aleph_q(x)+(q-1)\beta\langle d(x,\tilde{x})\rangle_{p(\tilde{x}|X=x)}} = \frac{\beta}{\Im_{RD}}$. Transforming $q \rightarrow 2-q^*$ in the numerator and evoking (7), (18) acquires the familiar form

$$p\left(\tilde{x}\left|x\right.\right) = \frac{p\left(\tilde{x}\right)\exp_{q^*}\left(-\beta^* d\left(x,\tilde{x}\right)\right)}{\tilde{Z}(x,\beta^*)}. \tag{19}$$

Note that $\tilde{Z}\left(x,\beta^*\right) = \Im_{RD}^{1/(1-q)}$, the partition function evaluated at each point of the source distribution.

To provide the nonextensive RD theory with a statistical physics connotation [5-7], the *effective nonextensive RD Helmholtz free energy* is

$$F_{RD}^q\left(\beta^*\right) = \frac{-1}{\beta^*}\left\langle \ln_q\tilde{Z}\left(x,\beta^*\right)\right\rangle_{p(x)} = \frac{1}{\beta^*}\left\langle \frac{\Im_{RD}-1}{q-1}\right\rangle_{p(x)}. \tag{20}$$

Solution of (19) may be viewed from two distinct perspectives, i.e. the *canonical perspective* and the *parametric perspective*. Owing to the *self-referential* nature of the *effective inverse temperature* $\beta^*$, the analysis and solution of (19) within the context of the *canonical perspective* is a formidable undertaking. For practical applications, the *parametric perspective* is utilized by evaluating the conditional *pdf* $p\left(\left.\tilde{x}\right| x\right)$, employing the nonextensive Blahut-Arimoto algorithm, for *a-priori* specified $\beta^* \in [0, \infty]$. *Note that within the context of the parametric perspective, the self-referential nature of* $\beta^*$ *vanishes.* The *inverse temperature* $\beta$ and the *effective inverse temperature* $\beta^*$ relate as

$$\beta = \frac{q\aleph_q\left(x\right)\beta^*}{\left[1 - \beta^*\left(q-1\right)\left\langle d\left(x, \tilde{x}\right)\right\rangle_{p\left(\left.\tilde{x}\right|X=x\right)}\right]}. \tag{21}$$

The Blahut-Arimoto algorithm for nonextensive RD theory is described in Algorithm 1.

**Algorithm 1** Nonextensive Blahut-Arimoto

```
Input
1. Source distribution $p\left(x\right) \in X$.
2. Set of representatives of quantized codebook given by $p\left(\tilde{x}\right) \in \tilde{X}$ values.
3. Parameter (effective inverse temperature) $\beta^* \in [0, \infty]$.
4. Convergence parameter $\varepsilon$.
5. Distortion measure $d(x, \tilde{x})$.
Output
Value of $R_q(D)$ where its slope equals $-\beta = \frac{-q\aleph_q(x)\beta^*}{\left[1-\beta^*(q-1)\langle d(x,\tilde{x})\rangle_{p(\tilde{x}|X=x)}\right]}$.
Initialization
Initialize $R_q^o \leftarrow \infty$ and randomly initialize $p(\tilde{x})$.
While True
• $p^{m+1}\left(\tilde{x}\left|x\right.\right) \leftarrow \frac{p^{(m)}(\tilde{x})\exp_{q^*}(-\beta^*\Delta d(x,\tilde{x}))}{\tilde{Z}^{(m+1)}(x,\beta^*)}$
• $p^{m+1}\left(\tilde{x}\right) \leftarrow \sum_x p\left(x\right)p^{m+1}\left(\tilde{x}\left|x\right.\right)$
$R_q^{(m+1)}\left(D\right) = I_q\left(p\left(x\right)p^{(m+1)}\left(\left.\tilde{x}\right| x\right)||p\left(x\right)p^{(m+1)}\left(\tilde{x}\right)\right)$.
If $\left(R_q^{(m)}\left(D\right) - R_q^{(m+1)}\left(D\right)\right) \leq \varepsilon$
Break
```

## 4. Nonextensive IB Variational Principle

The IB method of Tishby, Pereira, and Bialek [8] extends RD theory by introducing a principle for extracting relevant structure from data. This is accomplished by modeling structure extraction as data compression, followed by a quantification of the information preserved by the extracted structure with regards to a specific relevance variable. A thorough discussion of the IB method and some of its extensions is provided in the doctoral dissertation of Slonim [18].

Here, $X \in \mathcal{X}$ is a discrete random variable, distributed according to the marginal *pdf* $p\left(x\right)$ - the source distribution. Further, $\tilde{X} \in \tilde{\mathcal{X}}$ is another discrete random variable, distributed according to the marginal *pdf* $p\left(\tilde{x}\right)$. Here, $\tilde{X}$ is the *bottleneck representation* (analogous to the *quantized codebook* in RD theory). The relevance variable is represented by the discrete random variables $Y \in \mathcal{Y}$[2]. Inclusion of the relevance variable leads to the further introduction of two conditional *pdf*'s $p\left(\left.y\right| x\right)$, and, $p\left(\left.y\right| \tilde{x}\right)$.

The crux of the IB method is to "squeeze" the information between the source distribution $X$ and space of relevant variables $Y$ through a compact bottleneck representation $\tilde{X}$. This process is described by the Markov relation $\tilde{X} \leftrightarrow X \leftrightarrow Y$.

The nonextensive IB Lagrangian is

$$L_{IB}^q = I_q\left(X; \tilde{X}\right) - \beta I_q\left(\tilde{X}; Y\right) - \sum_x \lambda\left(x\right)\sum_{\tilde{x}} p\left(\left.\tilde{x}\right| x\right), \tag{22}$$

subject to the normalization of the conditional probability, i.e. $\sum_{\tilde{x}} p\left(\tilde{x}\left|x\right.\right) = 1$. Here, $I_q\left(\bullet;\bullet\right)$ denotes the generic generalized mutual information. The IB Lagrangian represents a tradeoff between the compression information $I_q\left(X; \tilde{X}\right)$, and, the relevance information $I_q\left(\tilde{X}; Y\right)$. Minimizing (22) with respect to the conditional probability $p\left(\left.\tilde{x}\right| x\right)$ for each $x$ and $\tilde{x}$, yields

$$\begin{aligned}
&\tfrac{qp(x)}{q-1}\left(\tfrac{p(\tilde{x}|x)}{p(\tilde{x})}\right)^{q-1} + \\
&-\beta\left\{\tfrac{q}{q-1}\sum_y p\left(y\right)p^q\left(\left.x\right| y\right)\left(\tfrac{p(\tilde{x}|x)}{p(\tilde{x})}\right)^{q-1}\right\} - \lambda^{(1)}\left(x\right) = 0, \\
&where, \\
&-\lambda^{(1)}\left(x\right) = -\lambda\left(x\right) - p^{(2-q)}\left(x\right) + \beta\sum_y p\left(y\right)\tfrac{p^q(x|y)}{p^{(q-1)}(x)}
\end{aligned} \tag{23}$$

The Markov consistency condition $p\left(\left.\tilde{x}\right| x\right) = \frac{p(\tilde{x}|y)}{p(x|y)}$ with Bayes rule $\frac{p(\tilde{x}|y)}{p(\tilde{x})} = \frac{p(y|\tilde{x})}{p(y)}$, yields

$$p\left(\left.\tilde{x}\right| x\right) = \frac{p\left(\tilde{x}\right)p\left(\left.y\right| \tilde{x}\right)}{p\left(y\right)p\left(\left.x\right| y\right)}. \tag{24}$$

Substituting (24) into (23), followed by algebraic manipulations to introduce the generalized K-Ld with the aid of (8), yields

$$\begin{aligned}
&\left[\tfrac{q}{q-1}\left(\tfrac{p(\tilde{x}|x)}{p(\tilde{x})}\right)^{q-1} + \beta\left\{q\sum_y p\left(\left.y\right| x\right)\ln_q\left(\tfrac{p(y|x)}{p(y|\tilde{x})}\right)\right\}\right] + \\
&-\lambda^{(2)}\left(x\right) = 0, \\
&where, \\
&-\lambda^{(2)}\left(x\right) = \oplus_q q\beta\sum_y p\left(\left.y\right| x\right)\ln_q\left(\tfrac{p(y)}{p(y|x)}\right) - \lambda^{(1)}\left(x\right)\big/p\left(x\right) + \\
&+\tfrac{q\beta}{1-q}\sum_y p\left(\left.y\right| x\right).
\end{aligned} \tag{25}$$

[2] Calligraphic fonts denote sets

Solving (25) for $p\left(\left.\tilde{x}\right|x\right)$ yields

$$\begin{array}{l} p\left(\left.\tilde{x}\right|x\right) = \\ p\left(\tilde{x}\right)\left[\frac{q-1}{q}\left\{-q\beta\sum\limits_{y} p\left(\left.y\right|x\right)\ln_q\left(\frac{p(\left.y\right|x)}{p(\left.y\right|\tilde{x})}\right)+\lambda^{(2)}\left(x\right)\right\}\right]^{\frac{1}{(q-1)}} . \end{array} \tag{26}$$

Multiplying (25) by $p\left(\left.\tilde{x}\right|x\right)$, and, summing over $\tilde{x}$ yields

$$\begin{array}{l} \lambda^{(2)}\left(x\right) = \\ = \frac{q}{q-1}\aleph_q\left(x\right)+q\beta\left\langle\sum\limits_{y} p\left(\left.y\right|x\right)\ln_q\left(\frac{p(\left.y\right|x)}{p(\left.y\right|\tilde{x})}\right)\right\rangle_{p(\left.\tilde{x}\right|X=x)} . \end{array} \tag{27}$$

Here, $\aleph_q\left(x\right) = \sum\limits_{\tilde{x}} p\left(\tilde{x}\right)\left(\frac{p(\left.\tilde{x}\right|x)}{p(\tilde{x})}\right)^q$. Substituting (27) into (26), setting $q \to 2-q^*$ and evoking (7), yields an expression in the form of the *q-deformed* exponential

$$\begin{array}{l} p\left(\left.\tilde{x}\right|x\right) = p\left(\tilde{x}\right)\frac{\exp_{q^*}\left\{-\beta^*_{IB}\sum\limits_{y} p(\left.y\right|x)\ln_q\left(\frac{p(\left.y\right|x)}{p(\left.y\right|\tilde{x})}\right)\right\}}{\Im_{IB}^{\frac{1}{(1-q)}}} \\ = p\left(\tilde{x}\right)\frac{\exp_{q^*}\{-\beta^*_{IB}I_q[p(\left.y\right|x)\|p(\left.y\right|\tilde{x})]\}}{\Im_{IB}^{\frac{1}{(1-q)}}} . \end{array} \tag{28}$$

In (28), the *effective trade-off parameter*, and, the partition function evaluated at each instance of the source distribution are

$$\begin{array}{l} \beta^*_{IB} = \frac{\beta}{\Im_{IB}}, \\ and, \\ \tilde{Z}\left(x,\beta^*_{IB}\right) = \Im_{IB}^{\frac{1}{(1-q)}} = \\ = \sum\limits_{\tilde{x}} p\left(\tilde{x}\right)\exp_{q^*}\left(-\beta^*_{IB}I_q\left[p\left(\left.y\right|x\right)\|\,p\left(\left.y\right|\tilde{x}\right)\right]\right), \\ where, \\ \Im_{IB} = \aleph_q\left(x\right)+ \\ +\left(q-1\right)\beta\left\langle\sum\limits_{y} p\left(\left.y\right|x\right)\ln_q\left(\frac{p(\left.y\right|x)}{p(\left.y\right|\tilde{x})}\right)\right\rangle_{p(\left.\tilde{x}\right|X=x)} , \end{array} \tag{29}$$

respectively. The *effective trade-off parameter* for the IB method relates to the *trade-off parameter* as

$$\beta = \frac{\beta^*_{IB}\aleph_q\left(x\right)}{\left[1-\left(q-1\right)\beta^*_{IB}\left\langle\sum\limits_{y} p\left(\left.y\right|x\right)\ln_q\left(\frac{p(\left.y\right|x)}{p(\left.y\right|\tilde{x})}\right)\right\rangle_{p(\left.\tilde{x}\right|X=x)}\right]} . \tag{30}$$

From (28), one obtains the iterative relation

$$\begin{array}{l} p^{(m+1)}\left(\left.\tilde{x}\right|x\right) \leftarrow \frac{p^{(m)}(\tilde{x})}{\tilde{Z}^{m+1}\left(x,\beta^*_{IB}\right)}\times \\ \times\exp_{q^*}\left\{-\beta^*_{IB}I_q\left[p\left(\left.y\right|x\right)\left\|p^{(m)}\left(\left.y\right|\tilde{x}\right)\right.\right]\right\}, \\ \forall\tilde{x}\in\tilde{\mathcal{X}}, x\subset\mathcal{X} \end{array} \tag{31}$$

*Here, (31) represents the primary distinction between the nonextensive IB method and the extensive IB method.* Comparing (31) to the equivalent expression in the method IB method [8], readily reveals that in the nonextensive case the exponential term is replaced by the *q-deformed* exponential and the K-Ld by the generalized K-Ld. The Markov relation $\tilde{X} \leftrightarrow X \leftrightarrow Y$ yields relations common to both nonextensive and extensive IB methods [18]

$$\begin{array}{l} p^{(m)}\left(\tilde{x}\right) = \sum\limits_{x} p\left(x\right)p^{(m)}\left(\left.\tilde{x}\right|x\right), \\ and, \\ p^{(m)}\left(\left.y\right|\tilde{x}\right) = \frac{1}{p^{(m)}(\tilde{x})}\sum\limits_{x} p^{(m)}\left(\left.\tilde{x}\right|x\right)p\left(x,y\right) . \end{array} \tag{32}$$

The gist of the simulation is to *a-priori* vary $\beta^*_{IB}\in[0,\infty]$, followed by an *a-posteriori* solution of (31) and (32) in a manner akin to the EM algorithm [19]. Algorithm 2 demonstrates the implementation of an iterative generalized IB method.

**Algorithm 2** Nonextensive Iterative Information Bottleneck Method

```
Input
1. Joint distribution $p(x,y)$.
2. Effective trade-off parameter $\beta^*_{IB}\in[0,\infty]$.
4. Cardinality $|\tilde{\mathcal{X}}| = M$, convergence parameter $\varepsilon$.
Output
A (typically "soft") partition $\tilde{X}$ of $\mathcal{X}$ into $M$ clusters.
Initialization
Randomly initialize $p(\tilde{x}|x)$ and find the corresponding $p(\tilde{x})$ and $p(y|\tilde{x})$.
While True
  • $p^{(m+1)}\left(\left.\tilde{x}\right|x\right) \leftarrow \frac{p^{(m)}(\tilde{x})}{\tilde{Z}^{m+1}\left(x,\beta^*_{IB}\right)}\times$
    $\times\exp_{q^*}\left\{-\beta^*_{IB}I_q\left[p\left(\left.y\right|x\right)\left\|p^{(m)}\left(\left.y\right|\tilde{x}\right)\right.\right]\right\}$,
    $\forall\tilde{x}\in\tilde{\mathcal{X}}, x\subset\mathcal{X}$
  • $p^{(m+1)}\left(\tilde{x}\right) \leftarrow \sum_{x} p\left(x\right)p^{(m+1)}\left(\left.\tilde{x}\right|x\right), \forall\tilde{x}\in\tilde{\mathcal{X}}$
  • $p^{(m+1)}\left(\left.y\right|\tilde{x}\right) = \frac{1}{p^{(m+1)}(\tilde{x})}\sum_{x} p^{(m+1)}\left(\left.\tilde{x}\right|x\right)p\left(x,y\right)$,
    $\forall\tilde{x}\in\tilde{\mathcal{X}}, y\subset\mathcal{Y}$
  If $\forall x \in \mathcal{X}, J-S_q(Nonextensive Jensen-Shannon)\left[p^{(m+1)}\left(\left.\tilde{x}\right|x\right), p^{(m)}\left(\left.\tilde{x}\right|x\right)\right]\leq\varepsilon$
  Break
```

## 5 Numerical Simulations and Physical Interpretations

The qualitative distinctions between nonextensive statistics and extensive statistics is demonstrated with the aid of the respective RD models. To this end, a sample of 1000 two-dimensional data points is drawn from three spherical

Gaussian distributions with means at $(2, 3.5), (0, 0), (0, 2)$ (the *quantized codebook*). The priors and standard deviations are $0.3, 0.4, 0.3$, and, $0.2, 0.5, 1.0$, respectively.

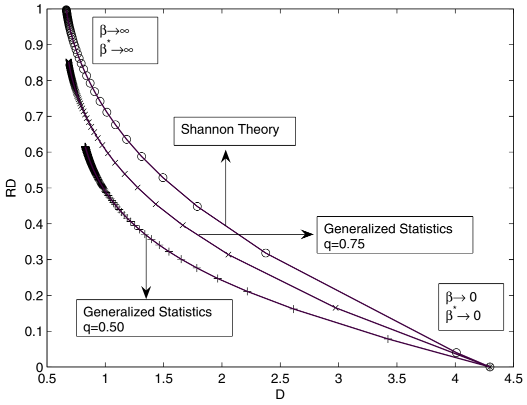

**Figure 1. Rate distortion curves for nonextensive and extensive statistics**

Fig.1 depicts the extensive and nonextensive RD curves, with the constituent discrete points overlaid upon them. Each curve has been generated for $\beta \in [.1, 2.5]$ (extensive case), and $\beta^* \in [.1, 100]$ (nonextensive cases), respectively. *It is observed that the nonextensive theory for the nonextensivity parameter q in the range* $0 < q < 1$ *uniformly exhibit a lower threshold for the minimum achievable compression-information in the distortion-compression plane, as compared to the extensive case. Note that for the nonextensive cases, the slope of the tangent drawn at any point on the RD curve is the negative of the inverse temperature* $-\beta$*, and not,* $-\beta^*$*.*

At the commencement, $\beta \to 0, \beta^* \to 0$, the Blahut-Arimoto algorithm solves for the compression phase. As $\beta$ and $\beta^*$ increase, the data points undergo soft clustering around the cluster centers. The hard clustering regime signifies regions where $\beta \to \infty, \beta^* \to \infty$.

As is observed in Fig. 1, the hard clustering regions correspond to portions of the RD curves where the discrete points are tightly packed. It is observed that the nonextensive RD models undergo compression and clustering more rapidly than the equivalent extensive RD model. A primary cause for such behavior is described in Fig. 2, where the nonextensive *effective inverse temperatures* $\beta^*$ are depicted versus the corresponding *inverse temperatures* $\beta$, with the aid of (21). As is noticed, $\beta^*$ increases rapidly with marginal increases in $\beta$.

The above arguments result in an observation in Fig.1 that is of particular significance. Specifically, even for less relaxed distortion constraints $\langle d(x, \tilde{x}) \rangle_{p(x,\tilde{x})}$, any nonextensive case for $0 < q < 1$ possesses a lower minimum compression information than the corresponding extensive case. The threshold for the minimum achievable compression-information decreases as $q \to 0$. *Note that all nonextensive RD curves inhabit the non-achievable region for the extensive case.* By definition, the *non-achievable region* is the region below a given RD curve, and signifies the domain in the *compression-distortion* plane where compression does not occur.

*Further, nonextensive RD models possessing a lower nonextensivity parameter* $q$ *inhabit the non-achievable regions of nonextensive RD models possessing a higher value of* $q$*.* These features imply the superiority of nonextensive models to perform data compression *vis-á-vis* any comparable model derived from Boltzmann-Gibbs statistics.

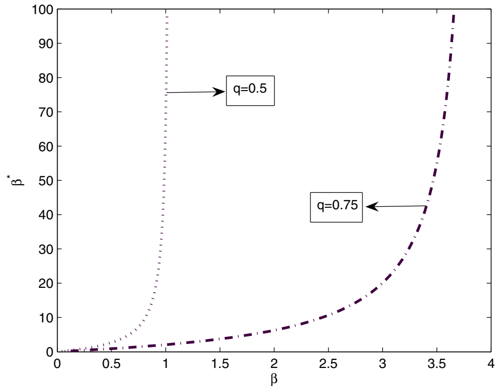

**Figure 2. Curves for $\beta$ v/s $\beta^*$**

## 6 Ongoing Work

A variational principle for a generalized RD theory and IB method has been presented. Ongoing work is focused upon two objectives. First, employing the T-M-P constraints a different set of variational principles is obtained. A comparative analysis between nonextensive variational principles for the RD theory and the IB method is currently underway. Next, the RD model and the Blahut-Arimoto numerical scheme studied in this paper represent an idealized scenario, involving well behaved sources and distortion measures.

An ongoing study treats a more realistic RD scenario by extending the works of Rose [20] and Banerjee *et. al.* [21]. This is accomplished via the formulation and analysis of a generalized Bregman RD (GBRD) model. One of the important features of the GBRD model is the derivation of a *Tsallis-Bregman lower bound* for the RD function. The *Tsallis-Bregman lower bound* will enable in establishing a principled theoretical rationale for the lower minimum compression information demonstrated by the generalized RD models, *vis-á-vis* equivalent extensive RD models. Results of these studies will be presented elsewhere.

### Acknowledgements

This work was supported by *RAND-MSR contract CSM-DI & S-QIT-101107-2005*. Gratitude is expressed to A. Plastino, S. Abe, and, K. Rose for helpful discussions.

## References

[1] C. Tsallis. Possible Generalizations of Boltzmann-Gibbs Statistics. *J. Stat. Phys.*, **542**, pp 479-487, 1988.

[2] M. Gell-Mann and C. Tsallis (Eds.). *Nonextensive Entropy-Interdisciplinary Applications*. Oxford University Press, New York, 2004.

[3] C. Tsallis. Generalized Entropy-Based Criterion for Consistent Testing. *Phys. Rev. E*, **58**, pp 1442-1445, 1998.

[4] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley & Sons, New York, NY, 1991.

[5] T. Berger. *Rate Distortion Theory*. Prentice-Hall, Englewood Cliffs, NJ, 1971.

[6] K. Rose, E. Gurewitz, and, G.C. Fox. A Deterministic Annealing Approach to Clustering. *Phys. Rev. Lett.*, **65**, pp 945948 , 1990.

[7] K. Rose. Deterministic Annealing for Clustering, Compression, Classification, Regression, and Related Optimization Problems. *Proc. IEEE*, **86**,11, pp 22102239 , 1999.

[8] N. Tishby, F. C. Pereira, W. Bialek. The Information Bottleneck Method . *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*, pp 368-377, 1999.

[9] P.T. Landsberg and V. Vedral. Distributions and Channel Capacities in Generalized Statistical Mechanics. *Phys. Lett. A*, **247**, pp. 211-217, 1998.

[10] T. Yamano. Information Theory based on Nonadditive Information Content. *Phys. Rev. E*, **63**, pp. 046105-046111, 2001.

[11] T. Yamano. Generalized Symmetric Mutual Information Applied for the Channel Capacity. *Phys. Rev. E.*, To Appear, 2006. Manuscript available at http://arxiv.org/abs/cond-mat/0102322.

[12] R. E. Blahut. Computation of Channel Capacity and Rate Distortion Functions. *IEEE Trans. on Inform. Theory*, **IT18**, pp. 460473, 1972.

[13] E. Borges. A Possible Deformed Algebra and Calculus Inspired in Nonextensive Thermostatistics. *Physica A*, **340**, pp. 95-111, 2004.

[14] E. M. F. Curado and C. Tsallis. Generalized Statistical Mechanics: Connection with Thermodynamics. *Journal of Physics A*, **24**, pp. L69-72, 1991.

[15] C Tsallis, RS Mendes, and A. R. Plastino. The Role of Constraints Within Generalized Nonextensive Statistics *Physica A*, **261**, pp. L534, 1998.

[16] G. L. Ferri, S. Martinez, and A. Plastino. Equivalence of the Four Versions of Tsallis Statistics. *J. Stat. Phys.*, **04**, pp. 04009-04024 , 2004. Manuscript available at http://arxiv.org/abs/cond-mat/0503441.

[17] S. Abe and A. K. Rajagopal. Nonadditive Conditional Entropy and its Significance for Local Realism. *Physica A*, **289**,(1), pp. 157-164, 2001. Manuscript available at http://arxiv.org/abs/quant-ph/0001085.

[18] N. Slonim. *The Information Bottleneck: Theory and Applications*. Ph.D. dissertation, Hebrew University of Jerusalem NJ, 2003. Manuscript available at http://www.princeton.edu/ nslonim/.

[19] G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley, N. Y., 1996.

[20] K. Rose. A Mapping Approach to Rate Distortion Computation and Analysis. *IEEE Trans. on Inform. Theory*, **40**,6 pp. 19391952, 1994.

[21] A. Banerjee, S. Merugu, I. Dhillon, and, J. Ghosh. Clustering with Bregman Divergences. *Journal of Machine Learning Research (JMLR)*, **6**, pp 1705-1749, 2005.