An Efficient Algorithm for Maximum Tsallis Entropy using Fenchel-duality

Tamir Hazan Amnon Shashua

School of Computer Science and Engineering
The Hebrew University of Jerusalem
Jerusalem, 91904
Israel

June 12, 2007

Abstract

We derive a dual-primal recursive algorithm based on the Fenchel duality framework, extending Dykstra's successive projections and Csiszar's I-projections schemes, to handle Tsallis MaxEnt. The Tsallis entropy $S_q(p)$ is a one-parameter extension of Shannon's entropy $H(p)$ in the sense that $S_{q\to 1}(p) = H(p)$. The solution of the Tsallis MaxEnt falls under a q-deformed Gibbs distribution which is a power-law distribution which contains the Gibbs distribution as a special case. The algorithm allows one to fit an extended distribution model to observations, which goes beyond the exponential distribution, yet is as simple as the original algorithms it extends.

1 Introduction

Maximum entropy (MaxEnt) is a general purpose method for making predictions or inferences from incomplete information. Its origins lie in statistical mechanics [8] starting from the late 50s and remains today an active area of research. Notable application examples in machine learning include natural language processing [1], text processing [10], visual recognition [9] and species distribution modeling [11].

The distribution model recovered by the MaxEnt principle is the exponential (Boltzmann-Gibbs) distribution. Distributions that are far from uniform, especially those whose energy is localized in a small number of areas, do not tend to be approximated well by an exponential. In statistical mechanics it has been observed that there are non-extensive physical phenomena which do not fall naturally under the Boltzmann-Gibbs distribution (long-range particle interaction, fractal phase-space, non-Markovian memory) but do fall under an extended power-law distribution model called the Tsallis entropy [14]. The Tsallis entropy $S_q(p)$ is a one-parameter extension of the Shannon-Boltzmann-Gibbs entropy $H(p)$ in the sense that $S_{q\to 1}(p) = H(p)$.

From a distribution modeling standpoint, having a power-law family of distributions which *contain* the Gibbs distribution as a special case is attractive. It would allow us to fit distributions to information which are not naturally amenable to an exponential model. In any case, we have all to gain and nothing to loose if we have an extended family of distributions to work with.

[1] Hebrew University, School of Computer Science and Engineering, Technical Report TR-110/2007, June 2007.

In this paper we derive a globally optimum algorithm for finding the Tsallis MaxEnt solution. The algorithm is derived using the Fenchel duality framework and turns out to be a natural extension of Dykstra's L_2 successive projections [7] and Csiszar's I-projections for Shannon's MaxEnt [3, 6].

The next three subsections provide necessary background material on MaxEnt, Tsallis entropy and the q-exponential function and the framework of Fenchel duality.

1.1 MaxEnt and the Boltzmann-Gibbs Distribution

Maximum entropy (MaxEnt) is an implementation of Laplace's *principle of insufficient reason* originally proposed in 1957 by E.T. Jaynes [8]. The general idea is when faced with the problem of approximating an unknown probability distribution $\pi(x)$, $x \in X$ where X is a finite domain $|X| = n$, from a small sample $\pi(x_i)$, $i = 1, ..., m$, where additional global information in the form of expectations are given, then among all the distributions $\hat{\pi}$ that satisfy the global constraints choose the one the is most *uniform*, i.e., maximizes Shannon's entropy $H(\hat{\pi}) = -\sum_{x \in X} \hat{\pi}(x) \ln \hat{\pi}(x)$.

The global constraints are referred to as "feature constraints" in machine learning and take the form of linear constraints $\sum_{x \in X} \pi(x) a_{xj} = \sum_i \pi(x_i) a_{ij} = b_j$, $j = 1, ..., k$, which state that the empirical expectations b_j are equal to the true expectations of "features" $\mathbf{a}_j : X \rightarrow R$. Setting this up as a mathematical optimization problem we get:

$$\max_{\mathbf{x}} -\sum_{i=1}^{n} x_i \ln x_i \;\; s.t. \;\; \mathbf{x} \geq 0, \sum_i x_i = 1, \sum_i x_i a_{ij} = b_j \tag{1}$$

for $j = 1, ..., k$. For simplicity of notation the values x_i stand for $\pi(x)$ for all $x \in X$. The optimal solution of MaxEnt as described by eqn. 1 falls under a Boltzmann-Gibbs distribution:

$$x_i^* = \frac{1}{Z} e^{-\sum_j \mu_j a_{ij}},$$

where Z is a normalizing factor (so that $\sum_i x_i^* = 1$) and μ_j are the Lagrange multipliers associated with the feature constraints. There are efficient algorithms to find the global optimal solution, as for example the I-projection dual-primal successive projections [3, 6], or the Iterated Scaling algorithm of [12, 4] which solve for the Lagrange multipliers by exploiting the MaxEnt vs. Max-Likelihood duality.

MaxEnt in a way is an interpolation scheme where unlike conventional interpolation that seeks $\hat{\pi}$ to be as close as possible to π at the sample points under some regularization (smoothness) constraint for all other entries, in MaxEnt the assumption is that the set of feature constraints "captures" the essence of the distribution with the remaining degree of freedom captured by striving to uniformity — or equivalently that the space of possible interpolation surfaces behave as an exponential (Gibbs distribution). The advantage of MaxEnt is that the number of samples could be very small (which would make conventional interpolation unsuccessful) where in return the number of feature constraints could be very large.

In practice, there is a need to *select* a small subset of features to avoid over-fitting and this is normally done by replacing the feature constraint $\sum_i x_i a_{ij} = b_j$ by inequality constraints $|\sum_i x_i a_{ij} - b_j| \leq \epsilon_j$ where ϵ_j can be determined by the data variances. The KKT theorem of non-linear programming would guarantee that $\mu_j^* = 0$ whenever $|\sum_i x_i^* a_{ij} - b_j| < \epsilon_j$ thus winnowing out the j'th feature. There are efficient methods to incorporate the inequalities in the MaxEnt/ML duality approach [5].

One of the shortcomings of the MaxEnt approach is that the solution is confined to an exponential model. An exponential model is not inherently bounded above and can give very large predicted values for entries outside the range presented by the empirical distribution. In what follows we will use the Tsallis entropy model, which is a one-parameter extension of the Shannon entropy, to present a power law distribution which includes the exponential model as a special case.

1.2 Maximum Tsallis Entropy and q-Gibbs

Tsallis entropy [14] defined below,

$$S_q(\mathbf{x}) = \frac{1 - \sum_i x_i^q}{q - 1} \quad (\mathbf{x} \geq 0, \sum_{i=1}^{n} x_i = 1)$$

where q is a real parameter, is a generalization of the Shannon-Boltzmann-Gibbs entropy. In the limit as $q \to 1$, we have that $x_i^{q-1} = e^{(q-1)\ln x_i} \approx 1 + (q-1)\ln x_i$, hence $S_1 = -\sum_i x_i \ln x_i$, which is Shannon's entropy.

Tsallis entropy can also be thought of a q-deformation of Shannon entropy by noting that $S_q(\mathbf{x}) = -\sum_i x_i^q \ln_q x_i$ where $\ln_q(x) = (x^{1-q} - 1)/(1-q)$ is the q-logarithm with the property $\ln_q(x) \to \ln x$ when $q \to 1$.

As for properties, $S_q \geq 0$, for $q > 0$, and equals to zero when all probabilities but one vanishes; like Shannon's entropy, S_q attains its maximum (for $q > 0$) for uniform distribution ($x_i = 1/n$), thus becoming $S_q = ln_q$. Moreover, like Shannon's entropy, it is possible to set-up axioms which uniquely define S_q [13].

The MaxEnt framework has a natural generalization with Tsallis entropy, but there exist several variants which differ in the way the feature constraints are defined [15]. The basic setup is maximization of $S_q(\mathbf{x})$ subject to power constraints $\sum_i x_i^q a_{ij} = b_j$ also known as q-expectations:

$$\max_{\mathbf{x}} S_q(\mathbf{x}) \;\; s.t. \;\; \mathbf{x} \geq 0, \; \mathbf{x}^\top \mathbf{1} = 1, \; \sum_i x_i^q a_{ij} = b_j \tag{2}$$

where $j = 1, ..., k$. Like in Shannon's MaxEnt, the optimal solution $\mathbf{x}^*$ belongs to a distribution family which can be represented as a q-deformed Gibbs distribution:

$$x_i^* = \frac{1}{Z} e_q^{-\sum_j \mu_j a_{ij}} \tag{3}$$

where Z is a normalization factor (so that $\sum_i x_i^* = 1$), μ_j correspond to the Lagrange multipliers associated with the feature (power) constraints, and e_q^x is the *q-exponential* defined by

$$e_q^x = \left\{ \begin{array}{cc} [1 + (1-q)x]^{1/(1-q)} & if \; 1 + (1-q)x > 0 \\ 0 & otherwise \end{array} \right\} \tag{4}$$

which has the properties $e_{q\to 1}^x = exp(x)$ and $e_q^{\ln_q x} = x$ (see Fig. 1a). The second variant replaces the power constraints with normalized power constraints:

$$\sum_i \frac{x_i^q}{\sum_r x_r^q} a_{ij} = b_j.$$

The optimal solution is also a member of a q-Gibbs distribution $x_i^* \cong e_q^{-\sum_j \hat{\mu}_j a_{ij}}$ where $\hat{\mu}_j$ are not the Lagrange multipliers associated with the normalized power constraints, however there exists a simple transformation:

$$\hat{\mu}_j = \frac{\mu_j}{\sum_r x_r^{*q} + (1-q)\mu_j b_j} \tag{5}$$

Finally, the third variant uses linear constraints $\sum_i x_i a_{ij} = b_j$. The third variant can be derived from the second by change of variables $y_i = x_i^q / \sum_r x_r^q$ (known as escort probabilities) and $q \to 1/q$ to obtain:

$$\max_{\mathbf{y}} S_{q'}(\mathbf{y}) \; s.t. \; \mathbf{y} \geq 0, \; \mathbf{y}^\top \mathbf{1} = 1, \; \sum_i y_i a_{ij} = b_j \tag{6}$$

with the solution $y_i^* \cong e_{q'}^{-\sum_j \hat{\mu}_j a_{ij}}$ where $q' = 1/q$. It is the third variant that interests us when generalizing MaxEnt for machine learning applications because the linear constraints represent global feature constraints whereas power constraints do not have a natural interpretation in the context of defining features. In all cases though, the optimal solution belongs to the q-deformed Gibbs distribution which in the limit $q \to 1$ becomes the Boltzmann-Gibbs distribution.

1.3 Fenchel Duality

Our main derivations are based on a form of duality known as *conjugate* or *Fenchel duality* (see [2], ch. 7). The duality is based on a transformation which associates to any function f a convex function g called the *conjugate* of f. Under convexity assumptions, the transformation is symmetric in that f is recovered by taking the conjugate of g.

The Fenchel conjugate of a function $f(\mathbf{x})$ is defined as the convex function $g(\boldsymbol{\lambda}) = \max_{\mathbf{x}}\{\boldsymbol{\lambda}^\top \mathbf{x} - f(\mathbf{x})\}$. If the function f is closed, proper and convex then the Fenchel conjugate of $g(\boldsymbol{\lambda})$ is $f(\mathbf{x})$. Likewise, the conjugate of $-f(\mathbf{x})$ is defined as the concave function $g(\boldsymbol{\lambda}) = \min_{\mathbf{x}}\{\boldsymbol{\lambda}^\top \mathbf{x} + f(\mathbf{x})\}$. The basic framework of Fenchel duality is the following primal-dual relationship:

$$\min_{\mathbf{x}} f_1(\mathbf{x}) - f_2(\mathbf{x}) = \max_{\boldsymbol{\lambda}} g_2(\boldsymbol{\lambda}) - g_1(\boldsymbol{\lambda}) \tag{7}$$

where $f_1, -f_2$ are convex, proper and closed functions and g_1 is the convex conjugate function of f_1, and $g_2(\boldsymbol{\lambda}) = \min_{\mathbf{x}}\{\boldsymbol{\lambda}^\top \mathbf{x} - f_2(\mathbf{x})\}$ is the conjugate concave function of f_2. The relationship between the optimal primal and dual solutions, $\mathbf{x}^*$ and $\boldsymbol{\lambda}^*$ are given by *Lagrangian optimality*:

$$\mathbf{x}^* = \text{argmin}_{\mathbf{x}}\{\mathbf{x}^\top \boldsymbol{\lambda}^* - f_2(\mathbf{x})\} \tag{8}$$

For a classical use of this framework, consider the problem:

$$\mu = \min_{\mathbf{x}\in C} f(\mathbf{x}),$$

where C is some convex set and f is a convex function. Let $\delta_C(\mathbf{x})$ be an indicator function $\delta_C(\mathbf{x}) = 0$ if $\mathbf{x} \in C$ and $\delta_C(\mathbf{x}) = \infty$ if $\mathbf{x} \notin C$. The constraint problem is obviously equivalent to the following unconstrained one:

$$\mu = \min_{\mathbf{x}} \delta_C(\mathbf{x}) - (-f(\mathbf{x})),$$

which is of the form $\min_{\mathbf{x}} f_1(\mathbf{x}) - f_2(\mathbf{x})$. The convex conjugate of δ_C is known as the *support function of* C defined by $\sigma_C(\boldsymbol{\lambda}) = \max_{\mathbf{x}\in C} \boldsymbol{\lambda}^\top \mathbf{x}$, thus from Fenchel duality eqn. 7 we have the following primal-dual relationship:

$$\min_{\mathbf{x}\in C} f(\mathbf{x}) = \max_{\boldsymbol{\lambda}} \left\{ \min_{\mathbf{x}}\{\boldsymbol{\lambda}^\top \mathbf{x} + f(\mathbf{x})\} - \sigma_C(\boldsymbol{\lambda}) \right\} \tag{9}$$

The generalized form of Fenchel duality is

$$\min_{\mathbf{x}} f_1(Q\mathbf{x}) - f_2(\mathbf{x}) = \max_{\boldsymbol{\lambda}} g_2(Q^\top \boldsymbol{\lambda}) - g_1(\boldsymbol{\lambda})$$

where Q is a matrix. By constructing $Q = [I.\ I,\ ..,I]^\top$ as a concatenation of identity matrices one can obtain a natural generalization of eqn. 9 in the case $C = \bigcap_1^k C_i$ is the intersection of closed convex sets:

$$\min_{\mathbf{x}\in\cap_{i=1}^k C_i} f(\mathbf{x}) = \max_{\boldsymbol{\lambda}_1,\ldots,\boldsymbol{\lambda}_k} \left\{ \min_{\mathbf{x}}\{(\sum_i \boldsymbol{\lambda}_i)^\top \mathbf{x} + f(\mathbf{x})\} - \sum_i \sigma_{C_i}(\boldsymbol{\lambda}_i) \right\} \tag{10}$$

2 The Recursive Primal-Dual Algorithm for Tsallis MaxEnt

Given the preceding material we wish to maximize $S_q(\mathbf{x})$ where $\mathbf{x} \in R^n$ is a probability vector ($\mathbf{x} \geq 0$ and $\sum_i x_i = 1$) subject to feature constraints $|\sum_i x_i a_{ij} - \tilde{b}_j| \leq \epsilon_j,\ j = 1,...,m-1$. We assume that ϵ_j can be estimated from the data (see [11] for details). Each inequality constraint $|\mathbf{x}^\top \mathbf{a}_j - \tilde{b}_j| \leq \epsilon_j$ can be substituted by two constraints $\mathbf{x}^\top \mathbf{a}_j \leq \epsilon_j + \tilde{b}_j$ and $\mathbf{x}^\top(-\mathbf{a}_j) \leq \epsilon_j - \tilde{b}_j$ which are of the general type $\mathbf{x}^\top \mathbf{f}_j \leq b_j, j = 1,...,k-2$ (where $k = 2m$). Furthermore, since $\mathbf{x} \geq 0$ we can replace $\sum_i x_i^q$ with $\sum_i |x_i|^q = \|\mathbf{x}\|_q^q$ thereby obtaining the following convex optimization problem:

$$\min_{\mathbf{x}} \frac{\alpha}{q}\|\mathbf{x}\|_q^q \ \ s.t.\ \ \mathbf{x} \geq 0,\ \mathbf{x}^\top \mathbf{1} = 1,\ \mathbf{x}^\top \mathbf{f}_j \leq b_j,\ \ j = 1,...,k-2 \tag{11}$$

where $\alpha = q/(q-1)$. The optimal solution $\mathbf{x}^*$ is a member of the q-Gibbs distribution family

$$x_i^* = \frac{1}{Z_\lambda} e_q(\sum_j \hat{\mu}_j f_{ji})$$

where $\hat{\mu}_j$ is related to μ_j the Lagrange multiplier of the constraint $\mathbf{x}^\top \mathbf{f}_j \leq b_j$ by eqn. 5 and in particular $\hat{\mu}_j$ and μ_j vanish together. From KKT theorem of non-linear programming μ_j vanishes if $\mathbf{x}^{*\top} \mathbf{f}_j < b_j$. As a result, vanishing $\hat{\mu}_j$ correspond to features $\mathbf{f}_j$ which are *winnowed out* in the

modeling process thereby achieving both MaxEnt modeling together with feature selection (as in traditional MaxEnt [5]).

We will not be solving for μ_j or $\hat{\mu}_j$ directly but instead use Fenchel duality to obtain a dual ascend where each update uses the primal-dual relationship as follows. Eqn. 11 corresponds to the primal problem:

$$\min_{\mathbf{x} \in C_1 \cap \cdots \cap C_k} \frac{\alpha}{q} \|\mathbf{x}\|_q^q \tag{12}$$

where $C_1, ..., C_k$ are convex sets corresponding to $\mathbf{x}^\top \mathbf{1} = 1$, $\mathbf{x}^\top \mathbf{f}_j \leq b_j$, $j = 1, ..., k-2$, and the set $\mathbf{x} \geq 0$. In other words, the feasible solutions are in the intersection of the k convex sets. From (generalized) Fenchel duality eqn. 10 we have the following:

$$\max_{\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_k} \left\{ \min_{\mathbf{x}} \left\{ (\sum_j \boldsymbol{\lambda}_j)^\top \mathbf{x} + \frac{\alpha}{q} \|\mathbf{x}\|_q^q \right\} - \sum_k \sigma_{C_j}(\boldsymbol{\lambda}_j) \right\} \tag{13}$$

where $\boldsymbol{\lambda}_j \in R^n$. The dual can be simplified by deriving the result of $\min_{\mathbf{x}} \{ (\sum_j \boldsymbol{\lambda}_j)^\top \mathbf{x} + \frac{\alpha}{q} \|\mathbf{x}\|_q^q \}$ (which is the conjugate function of $-\frac{\alpha}{q} \|\mathbf{x}\|_q^q$).

Proposition 1

$$-\frac{\alpha^{1-p}}{p} \| \sum_j \boldsymbol{\lambda}_j \|_p^p = \min_{\mathbf{x}} \left\{ (\sum_j \boldsymbol{\lambda}_j)^\top \mathbf{x} + \frac{\alpha}{q} \|\mathbf{x}\|_q^q \right\}$$

where $\frac{1}{p} + \frac{1}{q} = 1$.

The proof is the Appendix. Following substitution, the dual becomes:

$$\max_{\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_k} F(\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_k) = \left\{ -\frac{\alpha^{1-p}}{p} \| \sum_j \boldsymbol{\lambda}_j \|_p^p - \sum_k \sigma_{C_j}(\boldsymbol{\lambda}_j) \right\} \tag{14}$$

We consider next a cyclic iterative scheme for maximizing the dual vectors $\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_k$ where at each step we maximize $F(\cdot)$ with respect to $\boldsymbol{\lambda}_i$ while keeping the other dual vectors $\boldsymbol{\lambda}_j$, $j \neq i$, fixed. A classical way of doing this is to generate a sequence $\boldsymbol{\lambda}_1^{(t)}, ..., \boldsymbol{\lambda}_k^{(t)}$ where t is a cycle counter such that $\boldsymbol{\lambda}_i^{(t)}$ maximizes $F(\boldsymbol{\lambda}_1^{(t)}, ..., \boldsymbol{\lambda}_{i-1}^{(t)}, \mathbf{u}, \boldsymbol{\lambda}_{i+1}^{(t-1)}, ..., \boldsymbol{\lambda}_k^{(t-1)})$ over $\mathbf{u} \in R^n$.

$$\boldsymbol{\lambda}_i^{(t)} = \text{argmax}_{\boldsymbol{\lambda}} \left\{ -\frac{\alpha^{1-p}}{p} \|\boldsymbol{\lambda} - \mathbf{d}_i^{(t)}\|_p^p - \sigma_{C_i}(\boldsymbol{\lambda}) \right\} \tag{15}$$

where

$$-\mathbf{d}_i^{(t)} = \sum_{j=1}^{i-1} \boldsymbol{\lambda}_j^{(t)} + \sum_{j=i+1}^{k} \boldsymbol{\lambda}_j^{(t-1)} \tag{16}$$

Eqn. 15 is of the form $\max_{\boldsymbol{\lambda}} \{ g_2(\boldsymbol{\lambda}) - \sigma_C(\boldsymbol{\lambda}) \}$ and from Fenchel Duality we know that the convex conjugate $f_2(\mathbf{x})$ of $g_2(\boldsymbol{\lambda})$ satisfies:

$$\min_{\mathbf{x}} \{ \delta_C(\mathbf{x}) - f_2(\mathbf{x}) \} = \max_{\boldsymbol{\lambda}} \{ g_2(\boldsymbol{\lambda}) - \sigma_C(\boldsymbol{\lambda}) \} \tag{17}$$

Proposition 2 *The conjugate of* $-\frac{\alpha^{1-p}}{p} \|\boldsymbol{\lambda} - \mathbf{d}_i^{(t)}\|_p^p$ *is*

$$f_2(\mathbf{x}) = -\frac{\alpha}{q} \|\mathbf{x}\|_q^q + \mathbf{x}^\top \mathbf{d}_i^{(t)}$$

where $\frac{1}{p} + \frac{1}{q} = 1$.

The proof is the Appendix. Taken together, eqn. 17 becomes:

$$\max_{\boldsymbol{\lambda}} \left\{ -\frac{\alpha^{1-p}}{p} \|\boldsymbol{\lambda} - \mathbf{d}_i^{(t)}\|_p^p - \sigma_{C_i}(\boldsymbol{\lambda}) \right\} = \min_{\mathbf{x} \in C_i} \left\{ \frac{\alpha}{q} \|\mathbf{x}\|_q^q - \mathbf{x}^\top \mathbf{d}_i^{(t)} \right\} \tag{18}$$

The key to what will follow is that the optimization problem on the right-hand side:

$$\mathbf{x}^* = P_i(\mathbf{d}) = \text{argmin}_{\mathbf{x}\in C_i}\left\{\frac{\alpha}{q}\|\mathbf{x}\|_q^q - \mathbf{x}^\top\mathbf{d}\right\} \tag{19}$$

can be easily solved for the convex sets C_i of the form $\mathbf{x}^\top\mathbf{1} = 1$, $\mathbf{x}^\top\mathbf{f}_j \leq b_j$, $j = 1, ..., k-2$, and the set $\mathbf{x} \geq 0$. The details are in the appendix, thus we can assume that for any $\mathbf{d} \in R^n$ the operator $P_i(\mathbf{d})$ is at our disposal. The connection between $\boldsymbol{\lambda}_i^{(t)}$ and $P_i(\mathbf{d}_i^{(t)})$ is available from Lagrangian optimality:

$$\mathbf{x}^* = \text{argmin}_{\mathbf{x}}\{\mathbf{x}^\top\boldsymbol{\lambda}^* - f_2(\mathbf{x})\}$$

Proposition 3 *The solution for*

$$\mathbf{x}^* = \textit{argmin}_{\mathbf{x}}\left\{\mathbf{x}^\top\boldsymbol{\lambda}^* + \frac{\alpha}{q}\|\mathbf{x}\|_q^q - \mathbf{x}^\top\mathbf{d}\right\}$$

gives rise to:

$$\lambda_i^* = d_i - \alpha\cdot\text{sgn}(x_i^*)|x_i^*|^{q-1}, \quad i = 1, ..., n \tag{20}$$

where $\text{sgn}(x) = 1$ *when* $x \geq 0$ *and* -1 *otherwise.*

The proof is the Appendix. Let $\mathbf{h}_q : R^n \to R^n$ defined by

$$\mathbf{h}_q(\mathbf{x}) = -\alpha(\text{sgn}(x_1)|x_1|^{q-1}, ..., \text{sgn}(x_n)|x_n|^{q-1})^\top. \tag{21}$$

From Prop. 3 we obtain:

$$\boldsymbol{\lambda}_i^{(t)} - \mathbf{d}_i^{(t)} = \mathbf{h}_q(P_i(\mathbf{d}_i^{(t)})). \tag{22}$$

Eqn. 22 is a key equation: from it we will derive the recursive dual-primal algorithm as follows. We start with a basic relation:

Proposition 4

$$\boldsymbol{\lambda}_{i-1}^{(t)} - \mathbf{d}_{i-1}^{(t)} = \boldsymbol{\lambda}_i^{(t-1)} - \mathbf{d}_i^{(t)} \tag{23}$$

The proof is the Appendix. Let $\mathbf{x}_i^{(t)} = P_i(\mathbf{d}_i^{(t)})$. From eqns. 22,23 we obtain

$$\mathbf{d}_i^{(t)} = \boldsymbol{\lambda}_i^{(t-1)} - \mathbf{h}_q(\mathbf{x}_{i-1}^{(t)}) \tag{24}$$

Taken together, we have the following algorithm:

Algorithm 1 (Dual-Primal $\mathbf{h}_q$ Deformed Projections) *Set* $\boldsymbol{\lambda}_1^{(0)} = ... = \boldsymbol{\lambda}_k^{(0)} = 0$. *Use the convention* $\mathbf{x}_0^{(t)} = \mathbf{x}_k^{(t-1)}$ *and set* $\mathbf{x}_k^{(0)}$ *such that there holds* $h_q(\mathbf{x}_k^{(0)}) = 0$. *Explicitly, for* $q \neq 1$ *set* $\mathbf{x}_k^{(0)} = 0$ *and for* $q = 1$ *set* $\mathbf{x}_k^{(0)} = 1$.

1. *For* $t = 1, 2, ...$
2. *For* $i = 1, ...k$*:*
   - *(a)* $\mathbf{d}_i^{(t)} = \boldsymbol{\lambda}_i^{(t-1)} - \mathbf{h}_q(\mathbf{x}_{i-1}^{(t)})$
   - *(b)* $\mathbf{x}_i^{(t)} \leftarrow P_i(\mathbf{d}_i^{(t)})$.
   - *(c)* $\boldsymbol{\lambda}_i^{(t)} \leftarrow \mathbf{d}_i^{(t)} + \mathbf{h}_q(\mathbf{x}_i^{(t)})$.

The algorithm converges to the global optimal (guaranteed by convexity and Fenchel duality results) solution $\mathbf{x}^* = \mathbf{x}_i^{(t)}$ as $t \to \infty$ for every i.

2.1 Previous Work: Csiszar's I-projection and Dykstra's Algorithms

The dual-primal algorithm above is distinguished from previously known successive projection schemes by the "deformation" function $\mathbf{h}_q$. There are two noteworthy special cases where $q = 2$ and $q \to 1$. When $q = 2$, the original optimization eqn. 11 becomes $1/(q-1)\|\mathbf{x}\|_q^q = \|\mathbf{x}\|_2^2$ and the general primal eqn. 12 becomes the classical projection (in this case of the origin) onto the intersection of k convex sets. The deformation function becomes $\mathbf{h}_{q=2}(\mathbf{x}) = -2\mathbf{x}$ and in turn Alg. 1 reduces to Dykstra's [7] successive projection algorithm:

$$\begin{aligned} \mathbf{x}_i^{(t)} &\leftarrow P_i(\boldsymbol{\lambda}_i^{(t-1)} + 2\mathbf{x}_{i-1}^{(t)}) \\ \boldsymbol{\lambda}_i^{(t)} &\leftarrow \boldsymbol{\lambda}_i^{(t-1)} + 2\mathbf{x}_{i-1}^{(t)} - 2\mathbf{x}_i^{(t)} \end{aligned}$$

where $P(\mathbf{d}) = \text{argmin}_{\mathbf{x}\in C_i}\|\mathbf{x} - \mathbf{d}/2\|_2$. Note that we substituted $\mathbf{d}_i^{(t)} = \boldsymbol{\lambda}_i^{(t-1)} + 2\mathbf{x}_{i-1}^{(t)}$ in step (c) of Alg. 1. Note also that the factor 2 arises because normally one minimizes $(1/2)\|\mathbf{x}\|_2^2$ rather than $\|\mathbf{x}\|_2^2$.

When $q \to 1$, the objective primal function in eqn. 11 becomes (negative) Shannon's entropy $1/(q-1)\|\mathbf{x}\|_q^q \to \sum_i x_i \ln x_i$ and the general primal eqn. 12 becomes the classical I-projection problem handled by Csiszar and Dykstra [3, 6]. The deformation function becomes $\mathbf{h}_{q\to 1}(\mathbf{x}) = -1 - \ln \mathbf{x}$. Our basic operator $P(\mathbf{d})$ described in eqn. 19 becomes:

$$P(\mathbf{d}) = \text{argmin}_{\mathbf{x}\in C, \mathbf{x}^\top \mathbf{1}=1, \mathbf{x}\geq 0} \sum_i x_i \ln x_i - \sum_i x_i \ln e^{d_i} = \text{argmin}_{\mathbf{x}\in C, \mathbf{x}^\top \mathbf{1}=1, \mathbf{x}\geq 0} D(\mathbf{x} \,||\, e^{\mathbf{d}}), \tag{25}$$

where $D(\mathbf{x} \,||\, \mathbf{y})$ is the relative entropy between two distributions. Note that we added the probability constraints ($\mathbf{x} \geq 0, \mathbf{x}^\top \mathbf{1} = 1$) as part of the convex set, i.e., we will have $k-2$ sets C_i representing $\mathbf{x}^\top \mathbf{f}_i \leq b_i$ intersected with the probability simplex.

Taken together, step (a) of our algorithm becomes: $\mathbf{d}_i^{(t)} = \boldsymbol{\lambda}_i^{(t-1)} + 1 + \ln \mathbf{x}_{i-1}^{(t)}$ or equivalently $e^{\mathbf{d}_i^{(t)}} = \gamma e^{\boldsymbol{\lambda}_i^{(t-1)}} \mathbf{x}_{i-1}^{(t)}$, where $\gamma = \exp^1$. Step (c) becomes $\boldsymbol{\lambda}_i^{(t)} = \mathbf{d}_i^{(t)} - 1 - \ln \mathbf{x}_i^{(t)}$ or equivalently $e^{\boldsymbol{\lambda}_i^{(t)}} = (e^{\mathbf{d}_i^{(t)}})/(\gamma \mathbf{x}_i^{(t)})$. We can eliminate the intermediate variable $\mathbf{d}$ and obtain Csiszar's I-projection steps:

$$\begin{aligned} \mathbf{x}_i^{(t)} &\leftarrow P_i(\boldsymbol{\lambda}_i^{(t-1} + \ln \mathbf{x}_{i-1}^{(t)} + 1) = \text{argmin}_{\mathbf{x}} D(\mathbf{x} \,||\, e^{\boldsymbol{\lambda}_i^{(t-1)}} \mathbf{x}_{i-1}^{(t)}) \\ e^{\boldsymbol{\lambda}_i^{(t)}} &\leftarrow \frac{e^{\boldsymbol{\lambda}_i^{(t-1)}} \mathbf{x}_{i-1}^{(t)}}{\mathbf{x}_i^{(t)}} \end{aligned}$$

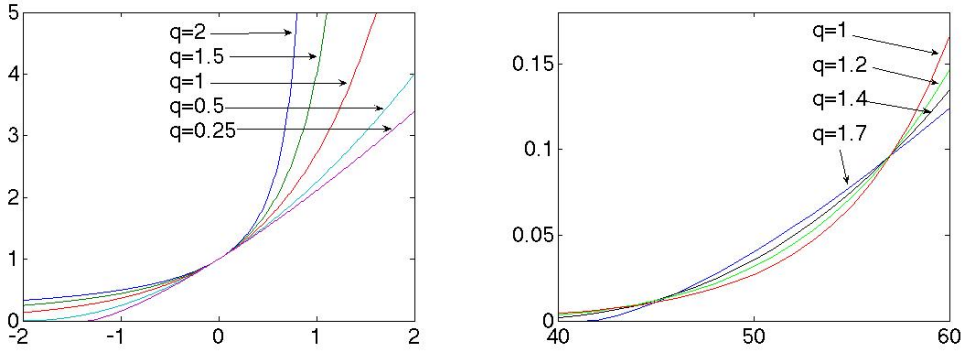
where $\mathbf{x}_{k-2}^{(0)} = 1$. Note that the constant γ in the update of $\mathbf{x}_i^{(t)}$ has dropped due to the constraint $\mathbf{x}^\top \mathbf{1} = 1$. We conclude therefore that our algorithm is a faithful generalization of the currently known special cases of finding the closest, in L_2 or relative-entropy measures, intersection of convex sets.

3 Experiments

We will begin with the classic 6-face dice with probabilities $p_1, ..., p_6$ of falling on the i'th face. This example has been used extensively to demonstrate the advantages (and fallacies [16]) of Laplace principle of insufficient reasoning. The question is that when given the empirical average ϵ calculated from a very large sample of tosses, can we find out the face probabilities? clearly there isn't enough information because we have two constraints ($\sum_i p_i = 1, \sum_i i p_i = \epsilon$) and six unknowns. The principle of insufficient reasoning states that one should prefer the uniform distribution when no information to the contrary exists. Thus, the MaxEnt approach solves the following problem:

$$\max H(\mathbf{p}) \;\; s.t. \;\; \mathbf{p} \geq 0, \; \sum_i p_i = 1, \; \sum_i i p_i = \epsilon,$$

with the solution $p_i^* = (1/Z)e^{-i\mu}$ where μ is a function of ϵ and can be recovered numerically using the I-projection or MaxEnt/ML duality approaches. When $\epsilon = 3.5$ indeed the uniform distribution



(a) (b)

Figure 1: *(a) The q-exponential function for a number of q values. Note that $q = 1$ corresponds to the normal exponential; (b) the recovered q-Gibbs distributions of a 60-face dice with average toss value of* 55 *for $q = 1, 1.2, 1.4, 1.7$ (the average value on a fair dice is expected to be* 30.5*). As q increases in value the shape of the distribution curve assigns smaller values to the high faces, higher values to intermediate faces, and providing a vanishing probability to the lower faces. For example, when $q = 1.7$ the first* 41 *faces are assigned a vanishing probability and for $q = 1.2$ the first 22 faces are assigned zero probabilities.*

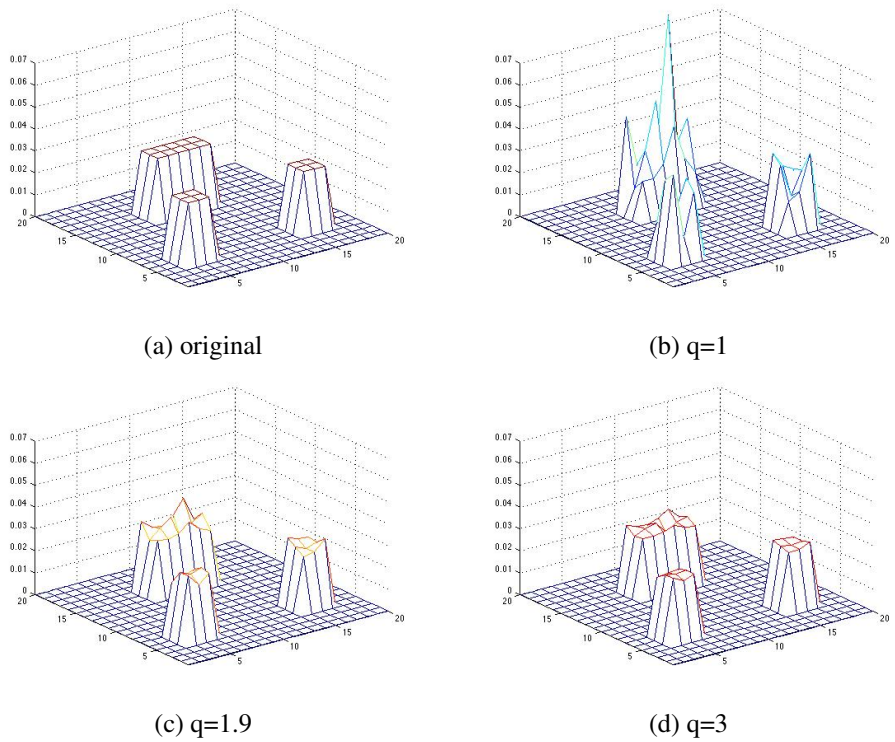
$p_i^* = 1/6$ emerges as the optimal solution and the exponential gets skewed left or right when ϵ deviates substantially from 3.5. The crucial point is that unless $\epsilon = 6$, the probability p_i^* is always *non-vanishing*. In other words, regardless of the number of tosses, if $\epsilon = 5.99$ there is still a non-zero chance that the dice will fall on face 1, i.e., $p_1^* > 0$. This is not necessarily the case when maximizing Tsallis entropy $S_q(\mathbf{p})$ with $q \neq 1$ since by definition e_q^x will vanish if $1 + (1-q)x < 0$. Fig. 1 shows the dice face distribution for a 60-face dice and $\epsilon = 55$ (note that the average of a uniform distribution is 30.5) for varying values of $q = 1, 1.2, 1.4, 1.7$. For $q = 1$ we have $S_1(\mathbf{p}) = H(\mathbf{p})$ and the probability distribution is the exponential function (skewed to the right). As q increases in value the shape of the distribution curve assigns smaller values to the high faces, higher values to intermediate faces, and providing a vanishing probability to the lower faces. For example, when $q = 1.7$ the first 41 faces are assigned a vanishing probability and for $q = 1.2$ the first 22 faces are assigned zero probabilities.

Note that the dice example does not have a "right" or "wrong" answer, but it seems that from an aesthetic standpoint the setting of $q = 1$ (Gibbs distribution) would be preferred since the exponential will never assign vanishing probabilities.

We next address an example in which there is an advantage of having the extra degree of freedom given by the parameter q. Consider a step-wise distribution over the 2D plane, i.e., the probability $p(x)$ vanishes expect in certain blocks. To make this simple we consider $p(x) = const$ for locations x within the blocks and $p(x) = 0$ everywhere else (see Fig. 2a). The generating distribution is therefore far from uniform and in such cases there is an advantage to a power-law distribution model because it could provide a better fit by an appropriate choice of q. We chose randomly feature vectors $\mathbf{f}_1, ..., \mathbf{f}_{20}$ which have their non-vanishing values confined to the block areas, i.e., the feature constraints support the block-wise structure of the original distribution. Fig. 2 shows the recovered distribution for values $q = 1, 1.9, 3$. One can clearly see that the Gibbs distribution ($q = 1$) provides the worst fit whereas increasing values of q allow the model to fit zeroes outside the blocks and provide a better fit within the blocks. For example, when $q = 1$ the values outside the blocks is around 10^{-4} which is a non-negligible number thus preventing the model to focus properly on the fit inside the blocks.

4 Summary

We have derived a dual-primal recursive update algorithm to the solve for the Tsallis MaxEnt problem. The algorithm is a generalization of the I-projection (for Shannon's MaxEnt) and Dykstra's L_2 successive projection scheme. The generalization appears in two places in our algorithm: (i) the "deformation" function $\mathbf{h}_q$ defined in eqn. 21 which reduces to the identity transformation for



(a) original (b) q=1

(c) q=1.9 (d) q=3

Figure 2: *Tsallis MaxEnt on step-wise sparse distribution: (a) original distribution, (b)* $q = 1$ *Shannon's MaxEnt approximation (c)* $q = 1.9$*, (d)* $q = 3$*.*

$q = 2$ (L_2 norm), and the "primitive" building block $P_i(\mathbf{d})$ in eqn. 19 which reduces to a projection $\|\mathbf{x} - \mathbf{d}\|_2^2$ when $q = 2$ and to the relative entropy (eqn. 25) when $q \to 1$. The algorithm converges to the global optimum solution (guaranteed by the Fenchel duality body of results).

Since our algorithm is a generalization of the current MaxEnt and recovers the Gibbs distribution for $q \to 1$ it has all the advantages of a generalization with little or none drawbacks. As for generalization, the degree of freedom introduced by q allows to for a larger family of models than just the exponential. The power-law distribution model (which contains the exponential) gives the modeler more flexibility in fitting the right model to observations. Moreover, the exponential model does not allow for vanishing entries whereas the power-law distributions do allow for it. As a result, distributions which are far from uniform (like the block-wise model used in our synthetic experiment) stand a better chance of approximation by the power-law model than the exponential.

As for drawbacks, the algorithm is as simple (3 line code) as those it generalizes (I-projection and Dykstra) but it is somewhat more computing intensive than the Iterated Scaling algorithm for Shannon's MaxEnt [12]. But since we are guaranteed to arrive to the global optimum the added computations are worth the advantage carried by the larger family of distribution models.

Further work would focus on (i) designing rules for choosing the right value of q to observations, and (ii) conduct empirical studies on real word data-sets to identify where in practice there is a need for power-law generalizations.

References

[1] A. L. Berger, V. J. Della Pietra, and S. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71, 1996.

[2] D. P. Bertsekas, A. Nedic, and A.E. Ozdaglar. *Convex analysis and optimization*. Athena Scientific, 2003.

[3] I. Csiszar. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, 1975.

[4] J. N. Darroch and D. Ratcliif. Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, 43:1470–1480, 1972.

[5] M. Dudik, S. J. Phillips, and R. E. Schapire. Performance guarantees for regularized maximum entropy density estimation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2004.

[6] R. Dykstra. An iterative procedure for obtaining i-projections onto the intersection of convex sets. *The Annals of Probability*, 13:975–984, 1985.

[7] R.L. Dykstra. An algorithm for restricted least squares regression. *J. of the Amer. Stat. Assoc.*, 78:837–842, 1983.

[8] E.T. Jaynes. Information theory and statistical mechanics. *Physica Review*, 106(4):620–630, 1957.

[9] S. Lazebnik, C. Schmid, and J. Ponce. A maximum entropy framework for part-based texture and object recognition. In *Proceedings of the International Conference on Computer Vision*, Beijing, China, October 2005.

[10] K. Nigam, J. Lafferty, and A. McCallum. Using maximum entropy for text classification. In *Proceedings IJCAI*, pages 61–67, 1999.

[11] S. J. Phillips, M. Dudik, and R. E. Schapire. Maximum entropy modeling of species geographic distributions. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2004.

[12] S. Della Pietra, V. J. Della Pietra, and J. D. Lafferty. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4):1–13, 1997.

[13] H. Suyari. On the most concise set of axioms and the uniqueness theorem for tsallis entropy. *Physica A: Math. Gen.*, 35:10731–10738, 2002.

[14] C. Tsallis. Possible generalization of bolzmann-gibbs statistics. *J. of Math. Phy.*, 52:479–487, 1988.

[15] C. Tsallis, R.S. Mendes, and A.R. Plastino. The role of constraints within generalized nonextensive statistics. *Physica A*, 261:534–554, 1998.

[16] J. Uffink. Can the maximum entropy principle be explained as a consistency requirement. *Studies in History and Philosophy of Modern Physics*, 26, 1995.

A Implementing the projection operator

The Fenchel duality framework reduces the original problem in eqn. 11 to successive projections of the type

$$P(\mathbf{d}) = \text{argmin}_{\mathbf{x}^\top \mathbf{f} \leq b} \left\{ \frac{\alpha}{q} \|\mathbf{x}\|_q^q - \mathbf{x}^\top \mathbf{d} \right\}$$

To achieve computational efficiency we recover the projection of $\mathbf{d}$ onto the hyperplane $\mathbf{x}^\top \mathbf{f} \leq b$ while performing two simple projections, the first onto $\mathbf{R}^n$ and the second onto the affine space $\mathbf{x}^\top \mathbf{f} = b$.

Proposition 5

$$P(\mathbf{d}) = \mathit{argmin}_{\mathbf{x} \in \mathbb{R}^n} \left\{ \frac{\alpha}{q} \|\mathbf{x}\|_q^q - \mathbf{x}^\top \mathbf{d} \right\} \quad \mathit{or} \quad P(\mathbf{d}) = \mathit{argmin}_{\mathbf{x}^\top \mathbf{f} = b} \left\{ \frac{\alpha}{q} \|\mathbf{x}\|_q^q - \mathbf{x}^\top \mathbf{d} \right\}$$

Proof: The objective function $f(\mathbf{x})$ is strictly convex as it is sum of the strictly convex function $\|\mathbf{x}\|_q^q$ and the linear function $\mathbf{x}^\top \mathbf{d}$, therefore it attains a unique global minima $\mathbf{x}^*$ in $\mathbb{R}^n$. If $\mathbf{x}^*$ satisfies the constraint $\mathbf{x}^\top \mathbf{f} \leq b$ then $P(\mathbf{d}) = \mathbf{x}^*$. Otherwise $\mathbf{x}^*$ is outside the feasible hyperplane, and we consider the objective function $f(\mathbf{x})$ restricted to line connecting $\mathbf{x}^*$ and $P(\mathbf{d})$. The function $f(\mathbf{x})$ is strictly convex hence it is strictly decreasing on the line from $P(\mathbf{d})$ and $\mathbf{x}^*$, and the minimality of $P(\mathbf{d})$ restricts the projected point to reside on the boundary of the hyperplane, i.e. $P(\mathbf{d})^\top \mathbf{f} = b$. □

The global minimum $\mathbf{x}^* = \text{argmin}_{\mathbf{x} \in \mathbb{R}^n} \left\{ \frac{\alpha}{q} \|\mathbf{x}\|_q^q - \mathbf{x}^\top \mathbf{d} \right\}$ admits a closed form solution

$$x_i^* = \text{sgn}(d_i) \sqrt[q-1]{|d_i|}/\alpha \tag{26}$$

which can be recovered by differentiating the objective function keeping in mind $d|x|/dx = \text{sgn}(x)$ while arguing that $\text{sgn}(x_i^*) = \text{sgn}(d_i)$.

On the other hand computing the projection restricted to the affine space $\mathbf{x}^\top \mathbf{f} = b$ is more involved and requires Lagrange duality theory: Define the Lagrangian function $L(\mathbf{x}, \mu) = \alpha \|\mathbf{x}\|_q^q / q - \mathbf{x}^\top \mathbf{d} - \mu(\mathbf{a}^\top \mathbf{f} - b)$ and consider the dual function $q(\mu) = \min_{\mathbf{x}} L(\mathbf{x}, \mu)$ described in Proposition. 1 by a straight forward variable substitution. The dual problem

$$\max_{\mu \in R} -\frac{\alpha^{1-p}}{q} \|\mathbf{d} + \mu \mathbf{f}\|_p^p + \mu b$$

correspond to the maximization of a concave function ([2], Proposition 6.2.1) and can be solved with any gradient based method. In particular, the Newton-Raphson update rule is very simple in our case: Repeat until convergence $\mu \leftarrow \mu - q'(\mu)/q''(\mu)$ where

$$\begin{aligned} q'(\mu) &= b - \alpha^{1-p} \sum_i f_i |d_i + \mu f_i|^{p-1} \mathrm{sgn}(d_i + \mu f_i) \\ q''(\mu) &= -\alpha^{1-p}(p-1) \sum_i f_i^2 |d_i + \mu f_i|^{p-2} \end{aligned}$$

Given μ^*, recover $P(\mathbf{d}) = \mathrm{argmin}_{\mathbf{x}} L(\mu^*, \mathbf{x})$, and combined with eqn.26 it takes the form

$$x_i^* = \mathrm{sgn}(d_i + \mu^* f_i) \sqrt[q-1]{|d_i + \mu^* f_i|}/\alpha$$

B Technical proofs

Proposition 1

$$-\frac{\alpha^{1-p}}{p} \| \sum_j \boldsymbol{\lambda}_j \|_p^p = \min_{\mathbf{x}} \left\{ (\sum_j \boldsymbol{\lambda}_j)^\top \mathbf{x} + \frac{\alpha}{q} \|\mathbf{x}\|_q^q \right\}$$

where $\frac{1}{p} + \frac{1}{q} = 1$.

Proof: The proof follows directly from the Lemma below:

Lemma 1 *For every* $\alpha > 0$ *holds*

$$-\frac{\alpha^{1-p}}{p} |\lambda|^p = \min_{x \in R} \left\{ x\lambda + \frac{\alpha}{q} |x|^q \right\}$$

where $\frac{1}{p} + \frac{1}{q} = 1$.

Proof: setting to zero the derivative of the right-hand function with respect to x yields

$$x^* = -\mathrm{sgn}(\lambda) |\lambda|^{\frac{1}{q-1}} \alpha^{\frac{1}{1-q}}.$$

Substituting x^* back into the right-hand function we obtain:

$$\begin{aligned} \lambda x^* + \frac{\alpha}{q} |x^*|^q &= -\alpha^{\frac{1}{1-q}} |\lambda|^{1+\frac{1}{q-1}} + \frac{1}{q} \alpha^{1+\frac{q}{q-1}} |\lambda|^{\frac{q}{q-1}} \\ &= -\alpha^{1-p} |\lambda|^p + \frac{1}{q} \alpha^{1-p} |\lambda|^p \\ &= -\frac{1}{p} \alpha^{1-p} |\lambda|^p \end{aligned}$$

□

Proposition 2 *The conjugate of* $-\frac{\alpha^{1-p}}{p} \|\boldsymbol{\lambda} - \mathbf{d}_i^{(t)}\|_p^p$ *is*

$$f_2(\mathbf{x}) = -\frac{\alpha}{q} \|\mathbf{x}\|_q^q + \mathbf{x}^\top \mathbf{d}_i^{(t)}$$

where $\frac{1}{p} + \frac{1}{q} = 1$.

Proof:

$$\min_{\boldsymbol{\lambda}} \left\{ \boldsymbol{\lambda}^\top \mathbf{x} + \frac{\alpha^{1-p}}{p} \|\boldsymbol{\lambda} - \mathbf{d}_i^{(t)}\|_p^p \right\} \overset{\hat{\boldsymbol{\lambda}} = \boldsymbol{\lambda} - \mathbf{d}_i^{(t)}}{=} \min_{\hat{\boldsymbol{\lambda}}} \left\{ \hat{\boldsymbol{\lambda}}^\top \mathbf{x} + \frac{\alpha^{1-p}}{p} \|\hat{\boldsymbol{\lambda}}\|_p^p \right\} + \mathbf{x}^\top \mathbf{d}_i^{(t)}$$

From Prop. 1 we have:

$$-\frac{1}{q} \beta^{1-q} \|\mathbf{x}\|_q^q = \min_{\boldsymbol{\lambda}} \left\{ \boldsymbol{\lambda}^\top \mathbf{x} + \frac{\beta}{p} \|\boldsymbol{\lambda}\|_p^p \right\}.$$

Substitute α^{1-p} for β while noting that $(1-q)(1-p) = 1$ we obtain $\beta^{1-q} = \alpha$. □

Proposition 3 *The solution for*

$$\mathbf{x}^* = \textit{argmin}_{\mathbf{x}} \left\{ \mathbf{x}^\top \boldsymbol{\lambda}^* + \frac{\alpha}{q} \|\mathbf{x}\|_q^q - \mathbf{x}^\top \mathbf{d} \right\}$$

gives rise to:

$$\lambda_i^* = d_i - \alpha \cdot \mathrm{sgn}(x_i^*) |x_i^*|^{q-1}, \quad i = 1, ..., n$$

where $\mathrm{sgn}(x) = 1$ *when* $x \geq 0$ *and* -1 *otherwise.*

Proof: From Prop. 1 we have:

$$x_i^* = -\mathrm{sgn}(\lambda_i^* - d_i) |\lambda_i^* - d_i|^{\frac{1}{q-1}} \alpha^{\frac{1}{1-q}}.$$

Note that $x_i^* < 0$ when $\lambda_i^* - d_i > 0$ and vice-versa. We have therefore:

$$x_i^* = \left\{ \begin{array}{cc} -(\lambda_i^* - d_i)^{\frac{1}{q-1}} \alpha^{\frac{1}{1-q}} & \lambda_i^* - d_i \geq 0 \\ |\lambda_i^* - d_i|^{\frac{1}{q-1}} \alpha^{\frac{1}{1-q}} & \lambda_i^* - d_i < 0 \end{array} \right\}$$

from which we obtain:

$$\lambda_i^* = \left\{ \begin{array}{cc} \alpha(-x_i^*)^{q-1} + d_i & x_i^* < 0 \\ -\alpha(x_i^*)^{q-1} + d_i & x_i^* \geq 0 \end{array} \right\}$$

□

Proposition 4

$$\boldsymbol{\lambda}_{i-1}^{(t)} - \mathbf{d}_{i-1}^{(t)} = \boldsymbol{\lambda}_i^{(t-1)} - \mathbf{d}_i^{(t)}$$

Proof: Recall the definition of $d_i^{(t)}$ in eqn. 16, then

$$\boldsymbol{\lambda}_{i-1}^{(t)} - \mathbf{d}_{i-1}^{(t)} = (\boldsymbol{\lambda}_{i-1}^{(t)} + \sum_{j=1}^{i-2} \boldsymbol{\lambda}_j^{(t)}) + \sum_{j=i}^{k} \boldsymbol{\lambda}_j^{(t-1)} = \sum_{j=1}^{i-1} \boldsymbol{\lambda}_j^{(t)} + \boldsymbol{\lambda}_i^{(t-1)} + \sum_{j=i+1}^{k} \boldsymbol{\lambda}_j^{(t-1)} = \boldsymbol{\lambda}_i^{(t-1)} - \mathbf{d}_i^{(t)}$$

□