Divergences. Scale invariant Divergences. Applications to linear inverse problems. N.M.F. Blind deconvolution

Henri Lantéri

Divergences. Scale invariant Divergences. Applications to linear inverse problems. - N.M.F. - - Blind deconvolution -

Version 1.0

Construction of the divergences. Problems related to their use.

Henri Lantéri

March 2, 2020

Contents

Résumé. **3**

Abstract. **5**

Preface - Outline of the document. **7**

chapter 1- Introduction **11**
1.1 Some Preliminary Considerations. 11
1.1.1 Brief reminders on inverse problems and the use of divergence functions. 11
1.1.2 The gap function $F(p, q)$ and its arguments "p" and "q". 12
1.1.3 Discrepancy (gap) functions - Divergences. 14
1.1.4 Choice of a divergence - Algorithmic aspects. 16
1.1.5 A few general considerations about the divergences. . 18
1.2 Application to the regularization. 19

chapter 2 - Convex functions and divergences. **21**
2.1 Convex functions - Some properties. 21
2.1.1 Generalities. 21
2.1.2 Some properties of convex functions. 22
2.2 Csiszär's divergences - f (or I) divergences. 25
2.2.1 A few examples to illustrate these difficulties. 27
2.2.2 Consequences of these examples. 29
2.2.3 Convexity of Csiszär's divergences. 29
2.3 Bregman's divergences 30
2.3.1 Example. 32
2.3.2 Some possible variations. 33
2.4 Jensen's divergences. 33

2.5 Relationship between these divergences. 35
2.5.1 Between Csiszär's and Bregman's divergences. 35
2.5.2 Between Jensen's and Bregman's divergences. 35
2.5.3 Between Jensen's and Csiszär's divergences. 36

chapter 3 -
Scale change invariance **37**
3.1 Introduction. 37
3.2 Invariance factor K - General properties. 38
3.2.1 Calculation of the nominal invariance factor. 39
3.3 A few comments regarding the invariance factor K. 40
3.3.1 Fundamental property. 40
3.3.2 Extension of the fundamental property to "non-nominal" invariance factors. 44
3.3.3 Some precisions on the differential equation (3.35). . . 45
3.3.4 General form of the invariance factor $K(p,q)$. 48
3.3.5 Remarks. 50
3.4 Some properties of the Gradient of an invariant divergence. . 50
3.5 A very special case of the invariance factor. 52

chapter 4 -
Divergences and Entropies **53**
4.1 Shannon Entropy related divergences. 53
4.1.1 Direct form. 53
4.1.2 Dual form. 56
4.1.3 Symmetrical form. 58
4.2 Havrda-Charvat Entropy related divergences. 59
4.3 Sharma-Mittal Entropy related divergences. 60
4.4 Renyi Entropy related divergences. 62
4.5 Arimoto entropy related divergences. 63
4.5.1 Direct form. 63
4.5.2 Related dual divergences. 66
4.5.3 Symmetrical divergence. 66
4.5.4 Weighted versions of these divergences. 66
4.5.5 First type of extension. 67
4.5.6 Extension in the sense of the Generalized Logarithm. . 68
4.6 Jensen divergences based on Entropies. 69
4.6.1 Shannon's entropy. 69
4.6.2 Havrda-Charvat entropy. 70
4.6.3 Renyi's entropy. 72

chapter 5 - Alpha, Beta and Gamma divergences **75**
5.1 Introduction. 75
5.2 Csiszär's divergences - Alpha divergences. 77
5.2.1 On $f_1(x)$. 77
5.2.2 On $f_2(x)$. 78
5.2.3 On $f_c(x)$. 79
5.2.4 Some classical particular cases. 81
5.3 Bregman's divergences - Beta divergences. 81
5.3.1 Some classical particular cases. 82
5.4 Jensen's divergences. 82
5.5 Invariance by change of scale. 82
5.5.1 "A" divergences. 82
5.5.2 "B" divergences. 86
5.5.3 "J" divergences. 88
5.5.4 Important note on invariance. 89
5.5.5 A particular case of the invariance factor. 90
5.6 Generalization in the sense of Cichocki and Ghosh. 91
5.6.1 Generalized divergence. 91
5.6.2 Invariance by change of scale on "q". 92
5.6.3 Logarithmic form. 93
5.6.4 Transition to Renyi's divergence. 94
5.6.5 Dual Generalized Divergence. 95
5.6.6 Remark. 97

chapter 6 -
Other classical divergences. **99**
6.1 CHI2 (χ^2) divergences. 99
6.1.1 Neyman's CHI2 divergence. 99
6.1.2 Pearson's Chi2 divergence. 101
6.1.3 Extensions of Chi2 divergences. 103
6.1.4 Chi2 relative divergences. 104
6.2 Hellinger's divergence. 104
6.3 Triangular discrimination. 105
6.4 Harmonic mean divergence of Toussaint. 105
6.5 Henze-Penrose divergence. 107
6.6 Polya information divergence. 110
6.6.1 Some general observations. 110
6.6.2 Invariance by change of scale. 114
6.6.3 Remark. 115
6.7 Bose-Einstein and Fermi-Dirac divergences. 115

6.7.1 Bose-Einstein divergence. 115
6.7.2 Fermi-Dirac divergence. 118
6.8 Divergence of P.K.Bathia and S.Singh. 121
6.8.1 Invariance by scale change on "q". 123

chapter 7 -
Divergences between means **125**
7.1 Square root (Quadratic) - Arithmetic mean divergence. . . . 125
7.1.1 Unweighted version. 125
7.1.2 Weighted version. 127
7.1.3 Logarithmic form. 128
7.2 Square root (Quadratic)- Géométric mean divergence. . . . 129
7.2.1 Unweighted version. 129
7.2.2 Weighted version. 129
7.2.3 Logarithmic form. 130
7.3 Square root (Quadratic) - Harmonic mean divergence. . . . 131
7.3.1 Unweighted version. 131
7.3.2 Weighted version. 132
7.3.3 Logarithmic version. 133
7.4 Arithmetic - Geometric mean divergence 134
7.4.1 Basis unweighted version. 134
7.4.2 Weighted version. 135
7.4.3 Logarithmic form. 136
7.5 Arithmetic - Harmonic mean divergence 137
7.5.1 Unweighted version. 137
7.5.2 Weighted version. 138
7.5.3 Logarithmic form. 139

chapter 8 -
Divergences between mixed fields. **141**
8.1 F divergence (Jensen-Shannon relative divergence). 142
8.1.1 Direct form. 142
8.1.2 Dual form. 143
8.1.3 Symmetrical form. 144
8.1.4 Invariance by change of scale on "q". 145
8.2 G divergence (Arithmetic-Geometric relative divergence). . . 146
8.2.1 Direct form. 146
8.2.2 Dual form. 147
8.2.3 T symmetrical divergence. 149
8.2.4 Invariance by change of scale on "q". 150

8.3 Generalization of the F and G divergences. 150
8.3.1 Direct form. 150
8.3.2 Invariant form by change of scale on "q". 151
8.3.3 Dual form. 152
8.3.4 Symmetrical form. 153
8.4 J relative divergence of Dragomir. 153
8.4.1 Direct form. 153
8.4.2 Dual form. 154
8.4.3 Invariance by change of scale on "q". 155
8.4.4 Generalization of Dragomir's J relative divergence. . . 156
8.5 Generalizations proposed by Taneja. 157

chapter 9 -
Smoothness constraint regularization. **167**
9.1 Invariant divergences derived from the "α" and "β" divergences. 168
9.1.1 Constant default solution = C. 168
9.1.2 Use of the Laplacian operator. 169
9.1.3 Important Note. 172
9.2 Effect of the invariance factor $K^* = \frac{\sum_j p_j}{\sum_j q_j}$. 172
9.2.1 Applications of some divergences to regularization. . . 173

chapitre 10 -
Algorithmic point of view. **177**
10.1 Some preliminary remarks. 177
10.2 Split Gradient Method (S.G.M.). 178
10.2.1 Non-negativity constraint. 178
10.2.2 Standard Algorithms - Non multiplicative 183
10.2.3 Multiplicative Algorithms. 183
10.2.4 Effect of the function expressing the constraints. . . . 184
10.2.5 Specific case of simplified divergences. 186
10.3 Introduction of the sum constraint. 191
10.3.1 Simplified divergences. 191
10.3.2 Non-simplified divergences. 192
10.3.3 Divergences invariant by change of scale. 194

chapter 11 -
Applications to the N.M.F. **199**
11.1 Introduction. 199
11.1.1 Linear unmixing. 199
11.1.2 Multispectral case. 201

11.1.3 Case of the NMF. 202
11.2 Generalities. 202
11.3 Algorithmic. 204
11.3.1 Principle of the method. 204
11.3.2 Iterations with respect to H. 206
11.3.3 Iterations on X. 210
11.3.4 Non-invariant divergences - Overview of the algorithms. 213
11.4 Applications to some typical divergences. 214
11.4.1 "Alpha" divergences. 215
11.4.2 "Beta" divergences. 217
11.5 Fundamental property of scale invariant divergences. 218
11.5.1 General case - All components of Q are concerned simultaneously. 219
11.5.2 The invariance is derived for each column of Q. 220
11.6 Iterative algorithm for invariant divergences. 222
11.6.1 Iterations on H. 222
11.6.2 Iterations on X. 223
11.7 Applications to some particular invariant divergences. 225
11.7.1 Invariant "Alpha" divergences. 225
11.7.2 Invariant "Beta" divergences 229
11.8 Régularisation. Non-invariant divergences 233
11.8.1 Regularization of the columns of the matrix H. 234
11.8.2 Regularization of the columns of the matrix X. 235
11.9 Some examples of regularized algorithms. 236
11.9.1 Regularized "Alpha" divergence. 236
11.9.2 Regularized "Beta" divergence. 239
11.10 Regularization. Scale invariant divergences. 240
11.10.1 Regularized invariant "Alpha" divergence. 242
11.10.2 Regularized invariant "Beta" divergence. 243
11.11 Overview of the techniques implemented. 244
11.11.1 Iterations on H. 244
11.11.2 Iterations on X. 244

chapter 12 -
Application to Blind Deconvolution. 245
12.1 Introduction. 245
12.1.1 Some details related to the calculus. 248
12.2 Algorithmic method. 249
12.3 Application to scale invariant divergences. 249

12.3.1 Property of scale invariant divergences with respect to "q". 249
12.4 Analysis of 2 particular cases. 254
12.4.1 Invariant "Beta" divergence. 254
12.4.2 Invariant mean square deviation. 255
12.4.3 Invariant Alpha divergence. 256
12.4.4 Kullback-Leibler invariant divergence. 257
12.5 Regularization - Scale invariant divergences. 258
12.5.1 Euclidean norm of the solution. 258
12.5.2 Use of the Laplacian norm of the solution. 259
12.6 Regularized algorithm. 261
12.7 Blind deconvolution - Non-invariant divergences. 261
12.7.1 Regularization. 264

Appendix 1 - Generalized Logarithm. 267

Appendix 2 - Means. 269

Appendix 3 - Generalization of GHOSH. 273

Appendix 4 - Gradient of invariant divergences. 275

Appendix 5 - Expressions of the invariance factor. 281

Appendix 6 - Effects of the invariance factor. 285

Appendix 7 - Inequalities between invariant divergences. 289

Appendix 8 - Order relations between divergences. 293

Appendix 9 - Regularization problems for the NMF. 297

List of Figures

2.1 Jensen divergence 23
2.2 Bregman divergence 24

4.1 Function $f_c(x) = x \log x + 1 - x$ 54
4.2 Function $f(x) = x \log x$ 55
4.3 Function $\breve{f}_c(x) = \log \frac{1}{x} + x - 1$ 56
4.4 Function $\breve{f}(x) = \log \frac{1}{x}$ 57
4.5 Function $\hat{f}_c(x) = x \log x - \log x$ 58
4.6 Function $f_{c\delta}(x) = \frac{1}{\delta - 1}\left[\left(\frac{1+x^\delta}{2}\right)^{\frac{1}{\delta}} - \left(\frac{1+x}{2}\right)\right], \delta = 2$ 64
4.7 Function $f_{1,\delta}(x) = \frac{1}{\delta - 1}\left[\left(\frac{1+x^\delta}{2}\right)^{\frac{1}{\delta}} - x\right], \delta = 2$ 65
4.8 Function $f_{2,\delta}(x) = \frac{1}{\delta - 1}\left[\left(\frac{1+x^\delta}{2}\right)^{\frac{1}{\delta}} - 1\right], \delta = 2$ 65

5.1 Function $f_1(x) = \frac{1}{\lambda(\lambda-1)}\left(x^\lambda - 1\right), \lambda = 0.5$ 76
5.2 Function $f_2(x) = \frac{1}{\lambda(\lambda-1)}\left(x^\lambda - x\right), \lambda = 0.5$ 76
5.3 Function $f_c(x) = \frac{1}{\lambda(\lambda-1)}\left[x^\lambda - \lambda x - (1-\lambda)\right], \lambda = 0.5$ 77
5.4 Function $\breve{f}_1(x) = \frac{1}{\lambda(\lambda-1)}\left(x^{1-\lambda} - x\right), \lambda = 0.5$ 78
5.5 Function $\breve{f}_2(x) = \frac{1}{\lambda(\lambda-1)}\left(x^{1-\lambda} - 1\right), \lambda = 0.5$ 79
5.6 Function $\breve{f}_c(x) = \frac{1}{\lambda(\lambda-1)}\left[x^{1-\lambda} - \lambda - (1-\lambda)x\right], \lambda = 0.5$ 80

6.1 Function $f_c(x) = (x-1)^2$ 99
6.2 Function $f_1(x) = x^2 - x$ 100
6.3 Function $\breve{f}_c(x) = \frac{(x-1)^2}{x}$ 102
6.4 Function $f(x) = x - \frac{2x}{1+x}$ 106
6.5 Function $\breve{f}(x) = 1 - \frac{2x}{1+x}$ 106

6.6 Function $f_c(x) = \frac{1}{2}\frac{(x-1)^2}{x+1}$ . . . 107
6.7 Function $f_1(x) = \frac{\beta^2x^2+(1-\beta)^2}{\beta x+(1-\beta)} - \left[\beta^2+(1-\beta)^2\right]$, $\beta = 0.5$ . . . 108
6.8 Function $f_2(x) = \frac{\beta^2x^2+(1-\beta)^2}{\beta x+(1-\beta)} - \left[\beta^2+(1-\beta)^2\right]x$, $\beta = 0.5$ . . 109
6.9 Function $f_c(x) = \frac{\beta^2x^2+(1-\beta)^2}{\beta x+(1-\beta)} - \left[\beta^2+(1-\beta)^2\right] + (x-1)\left[2\beta^3-4\beta^2+\beta\right]$, $\beta = 0.5$ . . . 109
6.10 Function $f_1(x) = x\log\frac{x}{1+\gamma x} + \frac{1}{\gamma}\log\frac{1}{1+\gamma x} + \frac{(1+\gamma)x}{\gamma}\log(1+\gamma)$, $\gamma = 0.8$ . . . 111
6.11 Function $f_2(x) = x\log\frac{x}{1+\gamma x} + \frac{1}{\gamma}\log\frac{1}{1+\gamma x} + \frac{1+\gamma}{\gamma}\log(1+\gamma)$, $\gamma = 0.8$ . . . 112
6.12 Function $f_c(x) = x\log\frac{x}{1+\gamma x} + \frac{1}{\gamma}\log\frac{1}{1+\gamma x} + \frac{1+\gamma x}{\gamma}\log(1+\gamma)$, $\gamma = 0.8$ . . . 112
6.13 Function $\breve{f}_c(x) = \log\frac{1}{x+\gamma} + \frac{x}{\gamma}\log\frac{x}{x+\gamma} + \frac{x+\gamma}{\gamma}\log(1+\gamma)$, $\gamma = 0.8$ 113
6.14 Function $f_{c,\alpha}(x) = x\log x + (\alpha+x)\log\frac{\alpha+1}{\alpha+x}$, $\alpha = 0.8$ . . . 116
6.15 Function $f_{c,\beta}(x) = x\log x + (\beta-x)\log\frac{\beta-x}{\beta-1}$, $\beta = 3$ . . . 119
6.16 Function $f(x) = \frac{x\sinh(\alpha\log x)}{\sinh\alpha}$, $\alpha = 0.5$ . . . 122
6.17 Function $f_c(x) = \frac{x\sinh(\alpha\log x)-\alpha x+\alpha}{\sinh\alpha}$, $\alpha = 0.5$ . . . 122

7.1 Function $f_{c\alpha}(x) = \sqrt{\alpha x^2+1-\alpha} - (\alpha x+1-\alpha)$; $\alpha = 0.5$ . . 127
7.2 Function $f_c(x) = \sqrt{\alpha x^2+(1-\alpha)} - x^\alpha$; $\alpha = 0.5$ . . . 130
7.3 Function $f_{c\alpha}(x) = \sqrt{\alpha x^2+(1-\alpha)} - \frac{x}{(1-\alpha)x+\alpha}$; $\alpha = 0.5$ . . 132
7.4 Function $f_{c\alpha}(x) = \alpha x+(1-\alpha) - x^\alpha$, $\alpha = 0.5$ . . . 136
7.5 Function $f_{c\alpha}(x) = \alpha x+(1-\alpha) - \frac{x}{(1-\alpha)x+\alpha}$, $\alpha = 0.5$ . . . 138

8.1 Function $f_c(x) = x\log\frac{x}{\alpha x+(1-\alpha)} + (1-\alpha)(1-x)$, $\alpha = 0.5$ . . 142
8.2 Function $f(x) = x\log\frac{x}{\alpha x+(1-\alpha)}$, $\alpha = 0.5$ . . . 143
8.3 Function $\breve{f}_c(x) = \log\frac{1}{\alpha+(1-\alpha)x} + (1-\alpha)(x-1)$, $\alpha = 0.5$ . . . 144
8.4 Function $\breve{f}(x) = \log\frac{1}{\alpha+(1-\alpha)x}$, $\alpha = 0.5$ . . . 145
8.5 Function $f_c(x) = [\alpha x+(1-\alpha)]\log\frac{\alpha x+(1-\alpha)}{x} + (1-\alpha)(x-1)$, $\alpha = 0.5$ . . . 147
8.6 Function $f(x) = [\alpha x+(1-\alpha)]\log\frac{\alpha x+(1-\alpha)}{x}$, $\alpha = 0.5$ . . . 148
8.7 Function $\breve{f}_c(x) = [\alpha+(1-\alpha)x]\log[\alpha+(1-\alpha)x] + (1-\alpha)(1-x)$, $\alpha = 0.5$ . . . 148
8.8 Function $\breve{f}(x) = [\alpha+(1-\alpha)x]\log[\alpha+(1-\alpha)x]$, $\alpha = 0.5$ . . 149
8.9 Function $f_c(x) = (1-\alpha)(x-1)\log[\alpha+(1-\alpha)x]$, $\alpha = 0.5$. 154
8.10 Function $\breve{f}_c(x) = (1-\alpha)(1-x)\log\left[\frac{\alpha x+(1-\alpha)}{x}\right]$, $\alpha = 0.5$ . . . 155

10.1 Karush, Kuhn, Tucker conditions for a constraint of the form $x \geq m$ 180

List of Tables

4.1 Divergences resulting from various values of "γ" and "δ". . . 67

Résumé.

Ce livre traite des fonctions permettant d'exprimer la différence entre deux champs de données ou "Divergence", pour des applications à des problèmes linéaires inverses. La plupart des divergences trouvées dans la littérature sont utilisées en théorie de l'information pour quantifier la différence entre deux densités de probabilité, c'est-à-dire des quantités positives dont les sommes sont égales à 1. Dans ce contexte, elles prennent des formes simplifiées qui les rendent inadaptées aux problèmes considérés ici, dans lesquels les champs de données ont des valeurs positives quelconques.
De manière systématique, les divergences classiques sont réexaminés et on en donne des formes qui peuvent être utilisées pour les problèmes inverses que nous considérons. Cela nous amène à préciser les méthodes de construction des écarts et à proposer un certain nombre de généralisations. La résolution des problèmes inverses implique systématiquement la minimisation d'une divergence entre les mesures physiques et un modèle dépendant de paramètres inconnus. Dans le cadre de la reconstruction d'images, le modèle est généralement linéaire et c'est un problème de minimisation sous la contrainte de la non-négativité et (éventuellement) de la somme constante des paramètres inconnus.
Afin de prendre en compte la contrainte de somme de façon simple, nous introduisons la classe des divergences invariantes par changement d'échelle sur les paramètres du modèle (divergences affines invariantes) et nous montrons des propriétés intéressantes de ces divergences.
Une extension de ces divergences permet d'obtenir l'invariance par changement d'échelle par rapport aux deux arguments intervenant dans les divergences ; ceci autorise l'utilisation de ces divergences dans la régularisation des problèmes inverses par contrainte de lissage. Des méthodes algorithmiques de minimisation des divergences sont développées, sous contraintes de non-négativité et de somme des composantes de la solution. Les méthodes présentées sont basées sur les conditions de Karush-Kuhn-Tucker qui doivent être satisfaites à l'optimum. La régularisation au sens de Tikhonov est prise

en compte dans ces méthodes. Le chapitre 11 associé à l'annexe 9 traite des applications de la MMF, tandis que le chapitre 12 est consacré à la déconvolution à l'aveugle. Dans ces deux chapitres, l'accent est mis sur l'intérêt des divergences invariantes.

Abstract.

This book deals with functions allowing to express the dissimilarity (discrepancy) between two data fields or "divergence functions" with the aim of applications to linear inverse problems.
Most of the divergences found in the litterature are used in the field of information theory to quantify the difference between two probability density functions, that is between positive data whose sum is equal to one. In such context, they take a simplified form that is not adapted to the problems considered here, in which the data fields are non-negative but with a sum not necessarily equal to one.
In a systematic way, we reconsider the classical divergences and we give their forms adapted to inverse problems.
To this end, we will recall the methods allowing to build such divergences, and propose some generalizations.
The resolution of inverse problems implies systematically the minimisation of a divergence between physical measurements and a model depending of the unknown parameters.
In the context image reconstruction, the model is generally linear and the constraints that must be taken into account are the non-negativity as well as (if necessary) the sum constraint of the unknown parameters.
To take into account in a simple way the sum constraint, we introduce the class of scale invariant or affine invariant divergences. Such divergences remains unchanged when the model parameters are multiplied by a constant positive factor. We show the general properties of the invariance factors, and we give some interesting characteristics of such divergences.
An extension of such divergences allows to obtain the property of invariance with respect to both the arguments of the divergences; this characteristic can be used to introduce the smoothness regularization of inverse problems, that is a regularisation in the sense of Tikhonov.
We then develop in a last step, minimisation methods of the divergences subject to non-negativity and sum constraints on the solution components.

These methods are founded on the Karush-Kuhn-Tucker conditions that must be fulfilled at the optimum. The Tikhonov regularization is considered in these methods.
Chapter 11 associated with Appendix 9 deal with the application to the NMF, while Chapter 12 is dedicated to the Blind Deconvolution problem.
In these two chapters, the interest of the scale invariant divergences is highlighted.

Preface - Outline of the document.

Solving inverse problems generally involves minimizing, with respect to unknown parameters, a gap function or discrepancy or divergence between measurements and a model describing the physical phenomenon under consideration. The divergences found in the literature are for the most part dedicated to problems related to information theory and are unsuitable for solving inverse problems. In order to have divergences adapted to these problems, we therefore re-examine the classical divergences, specify their construction methods, give some generalisations and present the invariant forms of these divergences. Some algorithmic methods of minimization under constraints are proposed.

Chapter 1 deals with some general considerations on inverse problems and on deviation functions or divergences. In particular, the problems concerning the influence of the inevitable noise on the measurements are briefly discussed. The types of divergences to be considered are specified.

Since these discrepancies are essentially based on the properties of the differentiable convex functions, some basic properties of the convex functions are recalled in Chapter 2 and the *"standard convex functions"* are defined, as opposed to the *"simple convex functions"*.

We then indicate the rules of construction of the Csiszar, Bregman and Jensen divergences based on such functions, then we analyze the convexity of the resulting "*separable*" divergences and we give some relations between the different types of divergences.

In Chapter 3, we introduce the "scale-invariant divergences", discuss in detail the methods of obtaining the invariance factors and indicate some remarkable properties of the resulting invariant divergences.

Chapter 4 is devoted to divergences based on classical entropy functions and some extensions of those divergences.

Chapter 5 deals with the "Alpha and Beta divergences" and we analyzes how they are constructed.
The forms of these divergences invariant by scale change are then given; this allows us to obtain the well known "Gamma divergence" and some divergences similar to this one.
In Chapter 6, a certain number of classical divergences are grouped together and extensions of these divergences are proposed in case the data fields considered are not probability densities.
Chapter 7 deals with divergences based on inequalities between weighted and unweighted generalized means, and extensions are given to the logarithmic forms of these divergences.
Chapter 8 deals with divergences between one of the two data fields and weighted mixtures of the two fields. Generalizations of these divergences are also discussed.
In Chapter 9, we study the application of scale invariant divergences to the problem of regularization of inverse problems by smoothness constraint and we show the interest of invariant divergences by change of scale on the 2 arguments for certain forms of regularization of this type.
Finally, in Chapter 10, we discuss algorithmic methods. After recalling the S.G.M. method, we develop extensions of this method in order to take into account simultaneously constraints of non negativity and sum of the unknown parameters. We develop in particular the case of scale invariant divergences and we show the interest of their specific properties to take into account the sum constraint.
Chapter 11 is devoted to Non-Negative Matrix Factoring (NMF), and more specifically to the introduction of sum constraints on unknown matrix columns.
Concerning this particular constraint, in the case where the divergence used does not have an invariance property, we proceed by changing variables.
On the other hand, if the considered divergence is invariant by change of scale, we show that the specific properties of these divergences lead to new particularly interesting algorithms.
Annex 9 deals with the problem of regularization in MMF and provides some clarifications and corrections to the classical techniques.
Chapter 12 deals with Blind Deconvolution.
It is shown in this chapter that the use of scale invariant divergences is of great interest insofar as the commutativity of the convolution product makes it possible to fully exploit the properties of such divergences when taking into account the sum constraints on the solution images.
As a comparison, the use of non-invariant divergences is also considered.
In these two chapters, the interest of the scale invariant divergences is clearly

highlighted insofar as they make it possible to take into account the sum constraints, in particular with regard to the Blind Deconvolution.

chapter 1- Introduction

1.1 Some Preliminary Considerations.

The goal of this book is to analyze the functions allowing to express the gap between 2 data fields and some methods to minimize these functions in order to apply them to inverse (linear) problems[16].
Two aspects will thus be examined successively: on the one hand the inverse problems aspect on which we will only make a few brief reminders, and on the other hand the "*gap functions*" aspect, i.e. the "*divergences functions*" on which we will extend much more extensively and which constitutes a large part of this work.

1.1.1 Brief reminders on inverse problems and the use of divergence functions.

In the general case, we have experimental measurements of a physical quantity dependent on a temporal, spatial or other variable, i.e. $y(t)$ ("t" not necessarily being time) and a model $m(t,x)$ to describe this quantity; this model depends on a certain number of parameters $x = (x_1, x_2, ...x_n)$ whose values are unknown.
We'll move to the special case where the model is linear with respect to the parameters.

It is further assumed that the measurements and the model are sampled with a step Δt so that $t_i = i\ \Delta t$; thus a set of measurements $y_i = y(t_i)$ is available and the corresponding values of the model are noted $m_i(x) = m(t_i, x)$.

To fix the ideas, we consider the problem of image restoration (reconstruction), and it is more precisely the problem of image deconvolution that will be considered throughout this book.

In this case, the sampling is implicitly linked to the pixels of the observed (convolved) image, the measurements are the intensities in the pixels of this image, the (linear) model corresponds to the convolution of the unknown original image with an impulse response that we will assume to be known; the unknown parameters are the intensities of the pixels of the original image.
In the context of inverse problems, the objective is to determine the optimal values of the unknown parameters of the model, i.e. those which will make it possible to bring the physical model as close as possible to the experimental measurements; one is thus led to compare the measurements with the model.
In order to do this, we need a function to quantify the difference between the two quantities (the measurements and the model).

In terms of inverse problems, the "discrepancy (gap) function" as we just defined expresses "*the data attachment*" or "*the data consistency*".

Obviously, in this exercise, the measurements are fixed quantities and the values of the model parameters are varied until the gap between the measurements and the model is minimal.
It is an inverse problem with all the difficulties that can be inherent in this type of problem [16] [48].
Assuming the model based on physical considerations is known, two steps occur in sequence:

- the choice of a gap function
- the choice of a method allowing to minimize such function with respect to parameters of the physical model (the unknowns).

1.1.2 The gap function $F(p, q)$ and its arguments "p" and "q".

An important point should be stressed immediately: except in some special cases, the variables "p" and "q" considered in this book are implicitly non-negative quantities.

* The data "p".

The first of the arguments in the deviation function concerns measurements. In order to adopt classical notations in this area, they are referred to as "p_i",

(these are the "y_i" from the previous section).
With rare exceptions, these measurements will be corrupted by errors that are attributed to noise whose nature is assumed to be known, which is not always true.
A first question immediately comes to mind: due to the presence of noise, do the measurements have the properties of the physical quantity being studied ?
Specifically, if the physical quantity under consideration is non-negative, "are the measurements non-negative ?".
If not, then the gap function must be able to take into account the negative values of the measurements. If the discrepancy function can only take into account non-negative quantities, then the measurements must be modified either by setting the negative values to zero or by shifting all the measurements to become non-negative, but in doing so, the statistical properties of the noise are changed.
For example, in astronomical imaging, corrections related to "dead" sensor pixels, or corrections related to differences in the gain of the elements of a CCD detector array change the nature of the noise present in the raw measurements.
In general, the range of definition of the deviation function should include the range of values of the measurements.
Clearly, the problem of non-negative measurement is a crucial point.

*** The physical model "q".**

The second argument which is used in the discrepancy function is the model, let us designate it by "q_i" (they are the $m_i(x)$); these values depend on the unknown parameters, which we note "x_l".
The question is, what does the physics say about the possible values of the "q_i" and of the "x_l"?
We can assume that the model is such that if the values of the "x_l" are chosen within a physically acceptable range, then the values of the "q_i" will be acceptable, that is, in particular, they will be non-negative.
In any event, that means not proposing any values for the unknowns "x_l", it will be essential to introduce some constraints on these values.
It is quite obvious, moreover, that a choice of the parameters "x_l" that satisfies the constraints, must lead to a value "q_i" of the model belonging to the domain of definition of the discrepancy function.
For example, in image reconstruction, the unknowns "x_l" must be non-

negative, and the transformation operation (convolution) must lead to non-negative images.

1.1.3 Discrepancy (gap) functions - Divergences.

Most of the "discrepancy functions" or "divergences" that appear in the literature are either Csiszär's divergences [27] [26], Ali-Silvey[1], either Bregman divergences [20], or Jensen divergences [53] and, more generally , when they do not strictly belong to one of these categories (this is the case when considering the generalizations of these divergences), they are based on the use of strictly convex differentiable basic functions, and rely on the properties of such functions.
This classification into different categories is purely formal, and it can be observed that a divergence may belong simultaneously to more than one of these categories.
Furthermore, non-differentiable divergences such as variational distance will not be considered in this book.
The major difficulty with the divergences found in the literature lies in the fact that many works have been developed within the framework of the information theory [8, 10, 78, 91, 87]. In this case, the data fields "p" and "q" are probability densities, which implies that the quantities considered are both non-negative and of sum equal to 1.
Thus, many of the divergences used in this context are constructed specifically for such applications, or have simplifications related to this particular case.
On the other hand, for the applications considered in this book, two important points must be taken into account: on the one hand, the data fields involved are almost never probability densities, so that simplifications concerning the sum of the arguments are not necessary, on the other hand, it is not simply a question of quantifying a difference between the two data fields, but of minimizing a difference between "p" and "q" with respect to the unknown parameters in one of the two fields considered: the model "q".
In addition, situations should be considered where the measurements "p" may be punctually negative due to noise, mainly in the case of Gaussian additive noise; if this is the case, it is important that the divergences used and the basic convex functions on which they are based, be defined for the negative values of the variable.
If this is not the case, the alternative is to pre-process the data to make it non-negative.

These remarks imply to analyze precisely the divergences found in the literature and to make sure that they are well adapted to the considered problem.

*** Csiszär's divergences.**

The above remarks apply in particular to Csiszär divergences, the use of which requires some precautions. Indeed, their mode of construction implies the use of a strictly convex basis function $f(x)$, but moreover, for applications related to information theory (and more precisely if we are dealing with probability densities), we impose on the basis function to have the property $f(1) = 0$ (this will correspond to particular convex functions that we will designate by "*simple convex functions*").
On the contrary, in our problems, the fields "p" and "q" have almost never spontaneously equal sums, (this will be one of the constraints of the problem), no more than sums equal to 1.
Then, in order to construct Csiszär divergences usable for our applications, the basic functions, in addition to the preceding properties, will have to be such that $f'(1) = 0$ (this will correspond to convex functions which we will designate by "*standard convex functions*"); we will come back to this point in the following sections.

*** Bregman's and Jensen's divergences.**

Unlike Csiszär's divergences, Bregman and Jensen's divergences, which are what one might call "*convexity measures*", imply only the use of a strictly convex base function $f(x)$. Except in the particular case where simplifications have been introduced after construction of the divergence, their use does not pose a problem for our applications, as long as the particular divergence considered is convex with respect to the true unknowns of the problem. Indeed, as we will see in the following sections, the fact that a Bregman or Jensen divergence is constructed on a strictly convex base function does not necessarily imply its convexity.

General properties of divergences.

Subject to differentiability, the divergences considered shall have the following properties:

- They'll have to be positive.

- They'll have to be zero when $p_i = q_i \ \forall i$.

Those two conditions are enough if you're comparing two different probability densities.
However, for applications to inverse problems that involve minimizing deviations from the model parameters "q", additional properties are required:

- Their gradient with respect to "q" must be zero for $p_i = q_i \ \forall i$.

- They'll have to be convex to the true unknowns "x". (usually going through the convexity in relation to "q").

Generally speaking:

- They are not necessarily symmetrical; if the need for this property arises for a particular problem, one can always build a "*ad hoc*" divergence.

- They do not necessarily meet the triangular inequality, so they are not necessarily distances (but this is not prohibited).

The consequence of these remarks is that a few checks are required before using a divergence, and this explains in part the analyses that are carried out in each case.

1.1.4 Choice of a divergence - Algorithmic aspects.

From a physical point of view, when considering the data attachment term, the choice of the gap function may be dictated by the statistical properties of the measurement noise, as is the case for maximum likelihood methods [48].
When the measurements are simulated, this point does not present any difficulty because the noise statistics are perfectly known. On the other hand, in the case of real measurements and in particular when a pre-treatment has been carried out, it is necessary to estimate the probability density of the residual noise before use.
In the Bayesian framework [48], given the lack of information on the real

nature of the noise, it is generally proposed to model the probability density of the noise using a Gaussian law.
This is justified by the fact that it is the "*least compromising*" choice given the lack of information on the exact nature of the measurement noise; one is thus led to minimize a divergence which is the root mean square deviation.
However, in some cases such as low photon count imaging, the use of a binomial distribution or Poisson's law seems more appropriate, leading to the minimization of a Kullback-Leibler divergence. Similarly, in [35], it is proposed to use the "*Beta*" divergence; a special case of this divergence is the Itakura-Saito divergence which seems to be better suited to the problem of musical signal processing.
Generally speaking, depending on the problem at hand, one could consider using any one of the various divergences discussed in this book to express the data attachment term.
As far as the minimization problem is concerned, the choice of method depends on the properties of the divergence used, and again the question is, which divergence to select?
From a strict optimization point of view, the simplest answer is to choose a strictly convex and differentiable discrepancy function, so that gradient-type descent methods can be used; however, for the chosen function, there must be a minimum corresponding to $p_i = q_i \ \forall i$, obtained by nulling the gradient of the discrepancy function (in the simplest case, i.e. without constraints). This is not always the case, as we will see in the following.
In addition, during the minimization process, constraints on the solution properties must be taken into account.
In this framework, properties reflecting strict constraints, such as the non-negativity of the components of the solution, or a constraint on the sum of these components, will be easily taken into account by conventional Lagrangian techniques of constrained minimization.
However, in the case of ill-posed problems [16] [48], solutions obtained by estimating the maximum likelihood under constraints, using iterative methods (usually), reveal instabilities as the number of iterations increases; an acceptable solution can only be obtained by empirically limiting the number of iterations. This implicitly regularizes the problem.
The classical alternative to solve this instability problem is to perform an explicit regularization.
The characteristic of these regularization methods consists in searching for a solution to make a tradeoff between the fidelity to the measured data and a fidelity to an information "*a priori*" [94].
To this end, one seeks to minimize a composite criterion made by introduc-

ing, next to the data attachment term , a penalty term allowing to reinforce certain suitable properties of the solution and which summarize our knowledge "*a priori*"; the relative importance of the two terms of the composite criterion thus made up being adjusted by a "*regularization factor*".
This "*energetic*" point of view is the one we'll adopt; the data attachment term and the penalty term will be expressed by two divergences of the type studied in this book; the penalty term expressing a "*gap*" between the current solution and a "*default*" solution reflecting the desirable property(ies) of the solution.
Such a regularisation can be interpreted in a Bayesian framework cited by [48] in which one must model a "*a priori*" law of probability of the solution, which should make it possible to take into account the desirable and known properties of the solution.
Applying Bayes' theorem yields the "*a posteriori*" distribution.
The estimation of the maximum of the "*a posteriori*" law is equivalent to finding the minimum of the composite criterion.

1.1.5 A few general considerations about the divergences.

Basically, a divergence is used to express the difference between 2 data fields: Field 1 ($C1$) and Field 2 ($C2$), and the divergence is written $D(C1\|C2)$. With the notations indicated previously, the basic fields (if one can say so) are "p" and "q", but many authors have introduced a 3rd field which is the weighted sum of the basic fields, i.e. $\alpha p + (1-\alpha) q \quad 0 \leq \alpha \leq 1$; thus, 3 different fields are involved, and one can easily imagine the variants of the divergences related to the assignment of these 3 fields on the 2 fields appearing in the divergence.
Add to that the fact that the divergences are generally not symmetrical, which adds to the number of possible divergences
In any case, all of these divergences will always express a gap between the two basic fields "p" and "q."
Finally, to further add to the variety of possible divergences, we will rely on the fact that if a $D(p\|q)$ divergence is expressed as a difference of 2 positive terms $D(p\|q) = A(p,q) - B(p,q)$, a generalization can be introduced by applying to each of the terms $A(p,q)$ and $B(p,q)$ an increasing function (e.g. the generalized logarithm or the logarithm) without changing the (positive) sign of the divergence.
From these few remarks, one can see the wide variety of divergences that can appear.

This being said, the important problem when minimizing a divergence remains that of the convexity of the divergence considered in relation to the true unknowns of the problem.

1.2 Application to the regularization.

The resolution of the inverse problem as we have presented it, implies the minimization (under constraints) with respect to "x", of a divergence between the measurements and the corresponding physical model $D_1\left(y \| m\left(x\right)\right)$. This divergence reflects the "attachment to the data."
However, because of the ill-posed character of the inverse problems, one is led to carry out a regularization of the problem [16, 48, 94] by introducing, next to the data attachment term, a "*penalty*" or "*regularization*" term which makes it possible to give particular properties to the solution. In the most classical case, where (for example) a certain "*smoothness*" is imposed on the solution, this term is expressed as a discrepancy (divergence) $D_2\left(x \| x_d\right)$ between the current solution and a "default solution x_d" having the required properties (in this case a smoothed version of the solution); on the other hand, it is reasonable to think that this default solution must also fulfill the constraints imposed on the solution.
We are thus led to minimize with respect to "x", under constraints, an expression of the form:

$$J\left(x\right) = D_1\left(y \| m\left(x\right)\right) + \gamma D_2\left(x \| x_d\right) \tag{1.1}$$

In this expression, "γ" is the positive regularization factor.
The constraints considered in this book are the non-negativity constraint and the sum constraint of the components of the solution:

$$x_i \geq 0 \quad \forall i \qquad ; \qquad \sum_i x_i = C \tag{1.2}$$

The terms D_1 and D_2 are divergences that will be analyzed in this work. After having analyzed a certain number of classical divergences, and in order to simply take into account the sum constraint, we will introduce the scale invariant divergences and we will show the advantages of the latter to satisfy this constraint. In particular, we will show that scale invariant divergences with respect to both arguments are particularly adapted to the problem of regularization by smoothness constraint in the sense of Tikhonov [93], and, more specifically, when the default solution "x_d" depends explicitly on the variable "x".

chapter 2 - Convex functions and divergences.

In this chapter, we recall some properties of the convex functions [17, 19, 46, 84] that will be useful in constructing the divergences.
We then look at the constructive modes of different types of classical divergences. We study their convexity and give some relationships that link them.

2.1 Convex functions - Some properties.

2.1.1 Generalities.

From a strictly convex function $f(x)$, we define the "mirror function" (Basseville [8, 10]) or "dual function" or "*conjugated function" (Osterreicher [75]) by:

$$\breve{f}(x) = xf\left(\frac{1}{x}\right) \tag{2.1}$$

This function will help build the Csiszär dual divergences.
The properties of the dual function will be:
* if $f(x)$ is convex, then $\breve{f}(x)$ is convex .
* if $f(1) = 0$, then we have $\breve{f}(1) = 0$.
A convex function with this property will be referred to as a ***"simple convex function."***
* if $f(1) = 0$, we have: $\breve{f}'(1) = -f'(1)$, which leaves the possibility of having $\breve{f}'(1) = -f'(1) = 0$ in the case of "self-mirror" functions.

From functions $f(x)$ and $\breve{f}(x)$, we define the function:

$$\hat{f}(x) = \frac{f(x) + \breve{f}(x)}{2} \tag{2.2}$$

This function will allow us to construct symmetrical Csiszär divergences. For this function, we have the following properties:
* si $f(1) = 0$, on a $\hat{f}(1) = 0$
* si $f(1) = 0$, on a $\hat{f}'(1) = 0$

This last property will be fundamental to our purpose.
This leads us, starting from a function $f(x)$ with no other property than the strict convexity, to define the function $f_c(x)$ which we can designate by *"**standard convex function**"*:

$$f_c(x) = f(x) - f(1) - (x-1) f'(1) \tag{2.3}$$

This function will be strictly convex, with $f_c(1) = 0$ and $f_c'(1) = 0$, so it will be positive or zero.
If we divide the right-hand side of (2.3) by $f''(1)$, then we get $f_c''(1) = 1$; the latter property is not necessary for our purposes.

Contrary to the *"simple convex functions"*, the *"standard convex functions"* will allow us to construct Csiszär divergences that can be used with data fields that are not probability densities, the so-called by Zhang [98] "measure invariant divergences" or "measures extended to allow denormalized densities".

Remarks

- a function of the type $\hat{f}(x)$ is a *"standard convex function"*.
- the reciprocal is not true.
- for the "auto-mirror" functions, we have $f(x) = \breve{f}(x) = \hat{f}(x) = f_c(x)$.

2.1.2 Some properties of convex functions.

For a convex $f(x)$ function, provided that we remain in its definition domain, that its derivative is defined and continuous on the same domain, we have the following properties:

$$(q-p) f'(p) \leq f(q) - f(p) \tag{2.4}$$

$$f(q) - f(p) \leq (q-p) f'(q) \tag{2.5}$$

We also have Jensen's inequality [53]: for $\alpha_i > 0 \ \forall i$ and $\sum_i \alpha_i = 1$.

$$\sum_i \alpha_i f(x_i) \geq f\left(\sum_i \alpha_i x_i\right) \tag{2.6}$$

This is writen with 2 points (see figure (2.1)):

$$\alpha f(p) + (1-\alpha) f(q) \geq f[\alpha p + (1-\alpha) q] \tag{2.7}$$

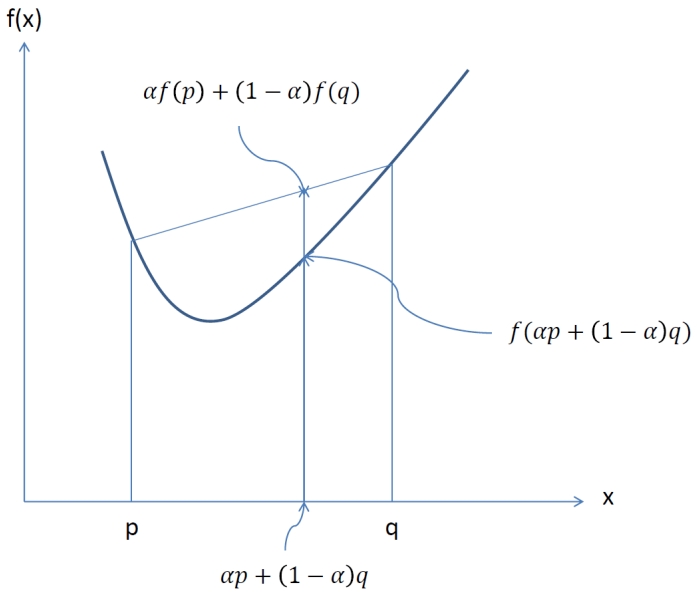

Figure 2.1: Jensen divergence

In the basic case, α =1/2, it comes:

$$\frac{f(p) + f(q)}{2} \geq f\left(\frac{p+q}{2}\right) \tag{2.8}$$

On the other hand, from (2.5), we can write:

$$f(p) - f(q) - (p-q) f'(q) \geq 0 \tag{2.9}$$

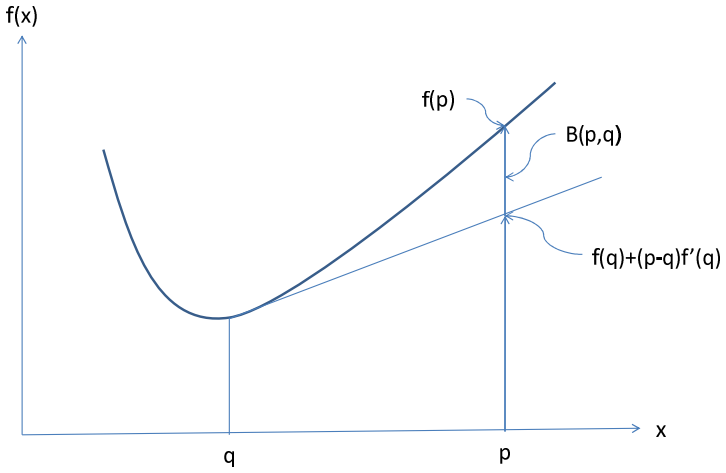

Figure 2.2: Bregman divergence

This inequality will be at the origin of Bregman's divergences, [20], see figure(2.2).
Similarly, from (2.4), we have:

$$f(q) - f(p) - (q-p) f'(p) \geq 0 \tag{2.10}$$

Which will allow us to build the dual Bregman divergence of the previous one. Finally, by summing up (2.10) and (2.9), we have:

$$(q-p)\left[f'(q) - f'(p)\right] \geq 0 \tag{2.11}$$

This inequality will intervene in the Burbea-Rao divergences [21].
Note that Burbea and Rao define a more general expression that is written:

$$(q-p)\left[\frac{\phi(q)}{q} - \frac{\phi(p)}{p}\right] \geq 0 \tag{2.12}$$

where $\frac{\phi(x)}{x}$ is an increasing function, which coincides with (2.11) if $f(x)$ is convex.

More generally, it can be considered that any inequality can give rise to a divergence.

The question of the definition domain is essential, because, as we shall see, this is what often restricts the field of use of the divergences to non-negative variables (for example).
In the general case that will be ours, the variables "p" and "q" are discretized (with the same step size); they are considered as vectors with components "p_i" and "q_i".
A divergence between "p" and "q" will be written in the general case as the separable form:

$$D(p\|q) = \sum_i d(p_i\|q_i) \tag{2.13}$$

And, from time to time, in the non-separable form, which is more delicate to analyse:

$$D(p\|q) = h\left[\sum_i d(p_i\|q_i)\right] \tag{2.14}$$

If we limit ourselves (for that time) to the separable form (2.13), the convexity of this divergence passes through the convexity of one of the terms of the sum: $d(p_i\|q_i)$.
We rely on the following definition:

Definition: The term $d(p_i\|q_i)$ is jointly convex with respect to "p_i" and "q_i" if its Hessian matrix H is positive defined ; it is convex with respect to "p_i" if h_{11} is positive and convex with respect to "q_i" if h_{22} is positive.

2.2 Csiszär's divergences - f (or I) divergences.

These divergences were introduced by Csiszär [27] and simultaneously by Ali and Silvey [1].
Lets consider a strictly convex function $f(x)$, a Csiszär divergence between two data fields "p" and "q" constructed on the function $f(x)$ is written:

$$C_f(p\|q) = \sum_i c_f(p_i\|q_i) = \sum_i q_i f\left(\frac{p_i}{q_i}\right) \tag{2.15}$$

For applications in information theory, the fields "p" and "q" are probability densities, they are positive quantities with equal sums (to 1 moreover); in this context, we simply impose on the basic function, the property:

$f(1) = 0$ [78] [75].
As previously stated, a convex function possessing only this property will be referred to as: "*simple convex function*".
Divergences built on such a function are not usable outside of this context without some precautions.
It should also be noted that if in a Csiszâr divergence, one explicitly introduces the fact that the variables are specifically summed to 1, the resulting simplified divergence is no longer a Csiszâr divergence in the strict sense, since there is no convex function to construct it according to the relation (2.15).
A symmetrical Csiszär divergence can always be obtained in Jeffreys' sense [52] applying the constructive process (2.15) using the base function $\hat{f}(x)$ (2.2).
In our problems, as noted above, the "*standard convex function*" will allow us to construct Csiszär's divergences that can be used with data fields that are not probability densities, what Zhang [98] calls "measure invariant divergences" or "measures extended to allow denormalized densities".
To show the necessity of using standard convex functions, let's consider Csiszär's divergences built on a simple convex function $f(x)$ on the one hand, and on the standard convex function $f_c(x)$ associated with it on the other hand. In the first case, the gradient with respect to "q_j" is written:

$$\frac{\partial C_f(p\|q)}{\partial q_j} = f\left(\frac{p_j}{q_j}\right) - \frac{p_j}{q_j} f'\left(\frac{p_j}{q_j}\right) \tag{2.16}$$

If we want this gradient to be zero for $p_j = q_j \ \forall j$, it requires that $f(1) = f'(1)$.
Given the property $f(1) = 0$, that implies that one must have: $f'(1) = 0$.
This observation obviously leads to the standard convex function deduced from $f(x)$:

$$f_c(x) = f(x) - f(1) - (x-1) f'(1) \tag{2.17}$$

The gradient with respect to "q_j" of the Csiszär's divergence built on $f_c(x)$ will be written:

$$\frac{\partial C_{f_c}(p\|q)}{\partial q_j} = f\left(\frac{p_j}{q_j}\right) - \frac{p_j}{q_j} f'\left(\frac{p_j}{q_j}\right) - f(1) + f'(1) \tag{2.18}$$

It will be spontaneously zero for $p_j = q_j \ \forall j$.
In the next section, we give two examples to illustrate this point.
Let's note that in the context of inverse problems, in general, if we want to use simplified divergences, we must take into account the simplifications

that have been introduced, for example by changing variables, that is, by introducing normalized variables.

2.2.1 A few examples to illustrate these difficulties.

Exemple 1.

We consider the standard convex function:

$$f_c(x) = (x-1)^2 \tag{2.19}$$

The corresponding Csiszär divergence is known as Neyman's Chi2:

$$\chi^2_N(p\|q) = \sum_i \frac{(p_i - q_i)^2}{q_i} \tag{2.20}$$

Its gradient with respect to "q" will be written:

$$\frac{\partial \chi^2_N(p\|q)}{\partial q_j} = 1 - \frac{p_j^2}{q_j^2} \tag{2.21}$$

The components of the gradient can cancel for $p_j = q_j$.
Let's now move to the simplified form in which, we introduced: $\sum_i p_i = \sum_i q_i$; the divergence becomes:

$$\chi^2_N(p\|q) = \sum_i \frac{p_i^2}{q_i} - p_i \tag{2.22}$$

This divergence is constructed in the sense of Csiszär on the simple convex function $f(x)$, associated with $f_c(x)$ which is written:

$$f(x) = x^2 - x \tag{2.23}$$

The corresponding gradient will be::

$$\frac{\partial \chi^2_N(p\|q)}{\partial q_j} = -\frac{p_j^2}{q_j^2} \tag{2.24}$$

Now, the gradient doesn't cancel out for $p_j = q_j$, it doesn't even cancel out at all unless $q_j \to \infty$, hence the problem appear.
We can associate with the same standard convex function another simple convex function that's written:

$$f(x) = x^2 - 1 \tag{2.25}$$

The corresponding Csiszär divergence is given by:

$$\chi^2_N(p\|q) = \sum_i \frac{p_i^2}{q_i} - q_i \tag{2.26}$$

The corresponding gradient will be:

$$\frac{\partial \chi^2_N(p\|q)}{\partial q_j} = -\frac{p_j^2}{q_j^2} - 1 \tag{2.27}$$

As in the previous example, it never cancels.
Let's go even further and set the sum of the variables to 1: $\sum_i p_i = \sum_i q_i = 1$; then, the divergence becomes:

$$\chi^2_N(p\|q) = \left[\sum_i \frac{p_i^2}{q_i}\right] - 1 \tag{2.28}$$

This function taken as such is no longer a Csiszär divergence, there is no basic convex function to obtain it according to the relation (2.15), its gradient with respect to "q" is given by (2.24).

Example 2.

Lets consider the standard convex function:

$$f_c(x) = x\log x + x - 1 \tag{2.29}$$

The Csiszär divergence built on this function is the Kullback-Leibler divergence [57].

$$KL(p\|q) = \sum_i p_i \log\frac{p_i}{q_i} + q_i - p_i \tag{2.30}$$

Its gradient with respect to "q" will be:

$$\frac{\partial KL(p\|q)}{\partial q_j} = 1 - \frac{p_j}{q_j} \tag{2.31}$$

It will be zero for $p_j = q_j$.
On the other hand, if we introduce in the expression (2.30), the simplification $\sum_i p_i = \sum_i q_i$, we obtain "Kullback information" [10]:

$$IKL(p\|q) = \sum_i p_i \log\frac{p_i}{q_i} \tag{2.32}$$

This expression is a Csiszär divergence built on the simple convex function:

$$f(x) = x \log x \tag{2.33}$$

Its gradient with respect to "q" will be:

$$\frac{\partial IKL(p\|q)}{\partial q_j} = -\frac{p_j}{q_j} \tag{2.34}$$

He'll never be equal to zero.
Note that here, the extreme simplification $\sum_i p_i = \sum_i q_i = 1$ will not add anything.

2.2.2 Consequences of these examples.

In our problem, it is a matter of making one of the two fields (the one that represents the model) evolve through the variations of the model parameters (the true unknowns), until the model is as close as possible to the measurements, in the sense of the divergence considered.

The divergences used being convex with respect to the unknown parameters, the classical optimization methods always imply to look for the set of unknown parameters that corresponds to "zero gradient", (even in constrained problems, it provides part of the solution).

It is quite obvious from the previous remarks and examples that if the divergence to be minimized, although convex, does not have a finite minimum, or if the minimum obtained by this method does not have a suitable physical meaning, it is inappropriate for our problem.

Therefore, in our problems, if the divergence used is a Csiszär divergence, it is imperative that we consider only those that are constructed on the basis of a standard convex function.
Furthermore, the divergences, regardless of how they are constructed, must not have undergone any simplification related to a particular application.

2.2.3 Convexity of Csiszär's divergences.

Taking into account (2.13) and (2.15), we must calculate the Hessian of a term of the form:

$$c_f(p\|q) = q f\left(\frac{p}{q}\right) \tag{2.35}$$

We classically consider that "p" and "q" are positive, and we have:

$$\frac{\partial^2 c_f(p\|q)}{\partial p^2} = \frac{1}{q} f"\left(\frac{p}{q}\right) > 0 \tag{2.36}$$

$$\frac{\partial^2 c_f(p\|q)}{\partial q^2} = \frac{p^2}{q^3} f"\left(\frac{p}{q}\right) > 0 \tag{2.37}$$

$$\frac{\partial^2 c_f(p\|q)}{\partial q \partial p} = \frac{\partial^2 c_f(p\|q)}{\partial p \partial q} = \frac{p}{q^2} f"\left(\frac{p}{q}\right) > 0 \tag{2.38}$$

Separate convexity is clear from (2.36) and (2.37).
The joint convexity is analyzed by considering the Hessian determinant that is written:

$$\frac{p^2}{q^4}\left[f"\left(\frac{p}{q}\right)\right]^2 - \left[\frac{p}{q^2} f"\left(\frac{p}{q}\right)\right]^2 = 0 \tag{2.39}$$

One of the eigenvalues is zero, the other is equal to the "Trace" of the Hessian matrix, therefore positive; the expression (2.35) is therefore jointly convex.
The Csiszär divergence is therefore jointly convex as a sum of jointly convex terms.

2.3 Bregman's divergences .

These divergences are typically convexity measurements.
They are based on a property of convex functions; they therefore imply the use of a basis convex function.
From this function, nothing is required other than strict convexity.
The property used to construct these divergences is expressed as:

- a convex curve is always located above any tangent to that curve; the Bregman divergence [20],[22] is the difference between the curve and the tangent (taken in that order).
- we can also say that for a strictly convex $f(x)$ function, the Bregman divergence is the difference between the function and its first order Taylor's development.

$$B_f(p\|q) = \sum_i b_f(p_i\|q_i) = \sum_i \left[f(p_i) - f(q_i) - (p_i - q_i) f'(q_i)\right] \tag{2.40}$$

Of course, since it is a basic property of convexity, a simple convex function and the corresponding standard convex function, which even have

a second derivative, lead to the same Bregman's divergence, in other words, two functions that differ from each other by a linear function lead to the same Bregman's divergence.
Therefore, as one would expect, whether the divergence is constructed on the simple convex function or on the associated standard convex function, the gradients with respect to "q" lead to the same expression being written:

$$\frac{\partial B_f(p\|q)}{\partial q_j} = \frac{\partial B_{f_c}(p\|q)}{\partial q_j} = (q_j - p_j) f''(q_j) = (q_j - p_j) f_c''(q_j) \tag{2.41}$$

That expression is obviously zero for $p_j = q_j \ \forall j$.
The convexity of Bregman's divergences is studied using the Hessian for one of the terms of the sum.

Proposal: If $f(x)$ is convex, $B_f(p\|q)$ is always convex with respect to the first "p" argument, but may be non convex with respect to the second argument "q".

Demonstration: From one of the sum terms appearing in (2.40), we calculate the elements of the Hessian.

$$\frac{\partial^2 b_f(p\|q)}{\partial p^2} = f''(p) \geq 0 \tag{2.42}$$

$$\frac{\partial^2 b_f(p\|q)}{\partial q^2} = f''(q) + (q-p) f'''(q) \tag{2.43}$$

$$\frac{\partial^2 b_f(p\|q)}{\partial p \partial q} = \frac{\partial^2 b_f(p\|q)}{\partial q \partial p} = -f''(q) \tag{2.44}$$

Thus, (2.42) implies convexity with respect to "p".
The sign of (2.43) depends of course on $f(x)$, so the convexity with respect to "q" depends on $f(x)$.
For joint convexity, we express the determinant of the Hessian matrix:

$$Det_H(p,q) = -\left[f''(q)\right]^2 + f''(p) f''(q) - (p-q) f''(p) f'''(q) \tag{2.45}$$

Dividing by $f''(p)\left[f''(q)\right]^2 > 0$, we have:

$$\frac{Det_H(p,q)}{f''(p)\left[f''(q)\right]^2} = -\frac{1}{f''(p)} + \frac{1}{f''(q)} - (p-q)\left[-\frac{1}{f''(q)}\right]' \tag{2.46}$$

This quantity is a Bregman's divergence built on the function $\left[-\frac{1}{f''(x)}\right]$, it will be positive if $\left[-\frac{1}{f''(x)}\right]$ is convex, i.e. if $\left[\frac{1}{f''(x)}\right]$ is concave.
If so, the Hessian determinant will be positive, the Hessian will be positive defined and the corresponding Bregman divergence will be "jointly convex".

Property: For a strictly convex $f(x)$ function, Bregman's divergence is jointly convex with respect to "p" and "q" if and only if $\left[1/f''(x)\right]$ is concave.
We'll get a similar property for Jensen's divergences.
From the expression (2.4), we obtain the adjoint (dual) form of the Bregman divergence:

$$B_f(q\|p) = \sum_i b_f(q_i\|p_i) = \sum_i \left[f(q_i) - f(p_i) - (q_i - p_i) f'(p_i)\right] \quad (2.47)$$

Following the same reasoning as for $B_f(p\|q)$ we can easily show that $B_f(q\|p)$ is jointly convex to "p" and "q" if and only if $\left[1/f''(x)\right]$ is concave.
Similarly, from the expression (2.11), we can deduce a form of divergence proposed by Burbea and Rao [21] which is expressed as follows:

$$A_f(p,q) = \sum_i a_f(p_i,q_i) = \sum_i \left[(p_i - q_i)\left(f'(p_i) - f'(q_i)\right)\right] \quad (2.48)$$

This symmetrical divergence is related to Bregman's $B_f(p\|q)$ and $B_f(q\|p)$ divergences by the relationship:

$$A_f(p,q) = B_f(p\|q) + B_f(q\|p) \quad (2.49)$$

By relying on the convexity of Bregman's divergences, $A_f(p,q)$ is jointly convex with respect to "p" and "q" if and only if $\left[1/f''(x)\right]$ is concave.
It should be noted, however, that $A_f(p,q)$ is not a Bregman divergence in that there is no convex function to construct it directly.

2.3.1 Example.

We consider the functions $f_1(x) = x^2 - 1$, $f_2(x) = x^2 - x$ and $f_c(x) = (x-1)^2$; these convex functions defined for any x, differ from each other by a linear function; only $f_c(x)$ is a standard convex function, but all lead to the same Bregman's divergence which is the mean square deviation (which is otherwise symmetric and respects the triangular inequality; it is therefore a distance).

2.3.2 Some possible variations.

Considering the Bregman divergence, based on a strictly convex function $f(x)$, between a convex combination of the variables "$\alpha p + (1-\alpha) q$" and "p" or "q", and taking into account the fact that the order of the arguments can be reversed, we can consider the divergence:

$$B_f^{\alpha}(p\|q) = B_f[\alpha p + (1-\alpha) q\|q] \tag{2.50}$$

which is written in more detail:

$$B_f^{\alpha}(p\|q) = \sum_i f[\alpha p_i + (1-\alpha) q_i] - f(q_i) - \alpha(p_i - q_i)\nabla f(q_i) \tag{2.51}$$

But we can also look at the divergences:
* $B_f[\alpha p + (1-\alpha) q\|p]$
* $B_f[p\|\alpha p + (1-\alpha) q]$
* $B_f[q\|\alpha p + (1-\alpha) q]$

That being said, one may ask if it's of much interest...

2.4 Jensen's divergences.

These divergences are applications of Jensen's inequality [53] (2.6), (2.7) or (2.8), with simple or standard convex functions.
Since it is, as for Bregman's divergences, a convexity measure, two convex functions having equal second derivatives (i.e. differing from each other by a linear function) will lead to the same Jensen's divergence.
For a strictly convex $f(x)$ base function and for $0 \leq \alpha \leq 1$, we define:

$$j_f^{\alpha}(p_i\|q_i) = \alpha f(p_i) + (1-\alpha) f(q_i) - f[\alpha p_i + (1-\alpha) q_i] \tag{2.52}$$

The corresponding Jensen's divergence is then:

$$J_f^{\alpha}(p\|q) = \sum_i j_f^{\alpha}(p_i\|q_i) \tag{2.53}$$

Therefore, as expected, whether the divergence is built on the simple convex function or on the associated standard convex function, the gradients with respect to "q" lead to the same expression which is written:

$$\frac{\partial J_f^{\alpha}(p\|q)}{\partial q_j} = \frac{\partial J_{f_c}^{\alpha}(p\|q)}{\partial q_j} = (1-\alpha)\left[f'(q_j) - f'(\alpha p_j + (1-\alpha) q_j)\right] \tag{2.54}$$

That expression is obviously zero for $p_j = q_j \ \forall j$.
So the Jensen's 1/2 divergence will be expressed as follows:

$$J_f^{\frac{1}{2}}(p\|q) = \sum_i \left\{ \frac{1}{2} f(p_i) + \frac{1}{2} f(q_i) - f\left[\frac{p_i + q_i}{2}\right] \right\} \tag{2.55}$$

This last divergence is symmetrical.
Its convexity is subject to the following rule:

Rule: For a strictly convex $f(x)$ function, Jensen's "α" divergence is jointly convex with respect to "p" and "q" if and only if $[1/f''(x)]$ is concave. (see. Burbea-Rao [21]).

Démonstration: From the expression (2.53), we calculate the elements of the Hessian matrix of one of the terms of the sum:

$$\frac{\partial^2 j_f^\alpha(p,q)}{\partial p^2} = \alpha f''(p) - \alpha^2 f''[\alpha p + (1-\alpha) q] \tag{2.56}$$

$$\frac{\partial^2 j_f^\alpha(p,q)}{\partial q^2} = (1-\alpha) f''(q) - (1-\alpha)^2 f''[\alpha p + (1-\alpha) q] \tag{2.57}$$

$$\frac{\partial^2 j_f^\alpha(p,q)}{\partial p \partial q} = \frac{\partial^2 j_f^\alpha(p,q)}{\partial q \partial p} = -\alpha(1-\alpha) f''[\alpha p + (1-\alpha) q] \tag{2.58}$$

The Hessian will be positive defined if its determinant is positive, i.e. if:

$$\begin{aligned} \left\{ f''(p) - \alpha f''[\alpha p + (1-\alpha) q] \right\} \left\{ f''(q) - (1-\alpha) f''[\alpha p + (1-\alpha) q] \right\} \\ -\alpha(1-\alpha) \left\{ f''[\alpha p + (1-\alpha) q] \right\}^2 > 0 \end{aligned} \tag{2.59}$$

Which gives immediately:

$$f''(p) f''(q) - \left[(1-\alpha) f''(p) + \alpha f''(q) \right] f''[\alpha p + (1-\alpha) q] > 0 \tag{2.60}$$

And finally:

$$\frac{\alpha}{f''(p)} + \frac{1-\alpha}{f''(q)} - \frac{1}{f''[\alpha p + (1-\alpha) q]} < 0 \tag{2.61}$$

Which expresses the concavity of $[1/f''(x)]$.
This result is identical to what we found with the Bregman's divergences.

2.5 Relationship between these divergences.

2.5.1 Between Csiszär's and Bregman's divergences.

In Censor and Zenios [22], we find the relationship:

$$C_f\left(p\|q\right)=\sum_i q_i\, f\left(\frac{p_i}{q_i}\right)=\sum_i q_i\, b_f\left(\frac{p_i}{q_i}\|1\right) \tag{2.62}$$

Or, similarly:

$$f\left(\frac{p_i}{q_i}\right)=b_f\left(\frac{p_i}{q_i}\|1\right) \tag{2.63}$$

We can show this easily, by expressing $b_f\left(\frac{p_i}{q_i}\|1\right)$ and considering that $f\left(1\right)=f^{'}\left(1\right)=0$, which is the case for the *"standard convex functions"*, on the other hand, the relations (2.62) and (2.63) are not valid for *"simple convex functions"*.

2.5.2 Between Jensen's and Bregman's divergences.

* - In the direction: $Bregman \rightarrow Jensen$.
A first relation is given by Basseville [8], [10]:

$$J_f^{\alpha}\left(p\|q\right)=\alpha B_f\left[p\|\alpha p+\left(1-\alpha\right)q\right]+\left(1-\alpha\right)B_f\left[q\|\alpha p+\left(1-\alpha\right)q\right] \tag{2.64}$$

But we can also establish another relationship that can be expressed as:

$$J_f^{\alpha}\left(p\|q\right)=\alpha B_f\left(p\|q\right)-B_f\left[\alpha p+\left(1-\alpha\right)q\|q\right] \tag{2.65}$$

* - In the direction: $Jensen \rightarrow Bregman$.
Basseville [8] gives the following relationship:

$$B_f\left(p\|q\right)=\lim_{\alpha\rightarrow 0}\; J_f^{\alpha}\left(p\|q\right) \tag{2.66}$$

But we can also establish the relationship:

$$B_f\left(p\|q\right)=\sum_{n=1}^{\infty}\frac{1}{\alpha^n}J_f^{\alpha}\left[\alpha^{n-1}p+\left(1-\alpha^{n-1}\right)q\|q\right] \tag{2.67}$$

To obtain this relationship, we start from expression (2.65) we write:

$$B_f\left(p\|q\right)=\frac{1}{\alpha}J_f^{\alpha}\left(p\|q\right)+\frac{1}{\alpha}B_f\left[\alpha p+\left(1-\alpha\right)q\|q\right] \tag{2.68}$$

Then replace $B_f\left[\alpha p+(1-\alpha)\,q\|q\right]\equiv B_f\left[z\|q\right]$ with its expression deduced from (2.68), that is:

$$B_f\left(z\|q\right)=\frac{1}{\alpha}J_f^{\alpha}\left(z\|q\right)+\frac{1}{\alpha}B_f\left[\alpha z+(1-\alpha)\,q\|q\right] \tag{2.69}$$

hence, by replacing "z" with its expression:

$$B_f\left(\alpha p+(1-\alpha)q\|q\right)=\frac{1}{\alpha}J_f^{\alpha}\left(\alpha p+(1-\alpha)q\|q\right)+\frac{1}{\alpha}B_f\left[\alpha^2 p+\left(1-\alpha^2\right)q\|q\right] \tag{2.70}$$

This expression is reported in (2.68) which gives:

$$B_f\left(p\|q\right)=\frac{1}{\alpha}J_f^{\alpha}\left(p\|q\right)+\frac{1}{\alpha^2}J_f^{\alpha}\left(\alpha p+(1-\alpha)q\|q\right)+\frac{1}{\alpha^2}B_f\left[\alpha^2 p+\left(1-\alpha^2\right)q\|q\right] \tag{2.71}$$

and so on and so forth, recursively.

2.5.3 Between Jensen's and Csiszär's divergences.

Based on the previous results, we can establish the relationship that yields Csiszär's divergences when we know Jensen's divergences.
From (2.67), we have:

$$b_f\left(\frac{p_i}{q_i}\|1\right)=\sum_{n=1}^{\infty}\frac{1}{\alpha^n}J_f^{\alpha}\left[\alpha^{n-1}\frac{p_i}{q_i}+\left(1-\alpha^{n-1}\right).1\|1\right] \tag{2.72}$$

Or also:

$$b_f\left(\frac{p_i}{q_i}\|1\right)=\sum_{n=1}^{\infty}\frac{1}{\alpha^n}J_f^{\alpha}\left[\frac{\alpha^{n-1}p_i+\left(1-\alpha^{n-1}\right)q_i}{q_i}\|1\right] \tag{2.73}$$

Then, with (2.62):

$$C_f\left(p\|q\right)=\sum_i q_i\sum_{n=1}^{\infty}\frac{1}{\alpha^n}J_f^{\alpha}\left[\frac{\alpha^{n-1}p_i+\left(1-\alpha^{n-1}\right)q_i}{q_i}\|1\right] \tag{2.74}$$

chapter 3 -
Scale change invariance

In this chapter, we are interested in the divergences invariant by change of scale and we indicate their construction mode; we also specify for this type of divergence, some useful properties to build minimization algorithms under sum and non-negativity constraints. We will consider in this chapter that the variables "p" and "q" involved in the expression of divergences are non-negative.

3.1 Introduction.

In the context of linear inverse problems, one is generally led to minimize with respect to the true unknown "x", a divergence between measures $y \equiv p$ and a linear model $Hx \equiv q$; this minimization is frequently associated with a non-negativity constraint $x_i \geq 0 \ \forall i$ and a sum constraint of the type $\sum_i x_i = Cte$.
In order to simply take into account this last constraint, we develop a class of divergences which are invariant by changing the scale with respect to "q", i.e. such that:

$$DI\left(p\|q\right) = DI\left(p\|aq\right) \tag{3.1}$$

where "DI" denotes an invariant divergence and "a" is a positive scalar.
The underlying idea being that during the iterative process of minimization, after each iteration, we will be able to renormalize "x" and respect the sum constraint, without changing the value of the divergence to be minimized; indeed, according to (3.1) we will obviously have:

$$DI\left(y\|Hx\right) = DI\left(y\|a\left(Hx\right)\right) = D\left(y\|H\left(ax\right)\right) \tag{3.2}$$

This being the case, we will show that the divergences invariant by change of scale possess a property that will be particularly interesting when the non-

negativity constraint is associated with the sum constraint in a minimization problem such as the one mentioned above.
Some examples of applications of invariant divergences have been proposed in [63], [65], [64] .

3.2 Invariance factor K - General properties.

The general idea is that starting with a divergence of any kind, we're going to transform it into a divergence of $DI(p\|q)$ (related to $D(p\|q)$), which is invariant with respect to the variable "q".
More precisely, in order to make a divergence scale invariant with respect to "q", we look for the expression of a **positive scalar factor** $K(p,q)$ such that the divergence $D(p\|Kq) = DI(p\|q)$ remains unchanged when the components of "q" are multiplied by a positive scalar.

The solution of this problem is not unique, indeed all expressions of $K(p,q)$ having the following properties are possible solutions of this problem:

1 - $K(p,q)$ must be scalar and positive.,

2 - In order to obtain a divergence which is invariant by scale change with respect to "q", the vector $[K(p,q).\ q]$ must be invariant when multiplying "q" by a constant.

Finally, in general terms:

*** If $p \to q$, we must have $K(p,q) \to 1$.**
consequently:

*** If $p \to q$, $D(p\|Kq) \to D(p\|q) \to 0$.**

We will note that the constraints imposed on the $K(p,q)$ factor do not necessarily relate it to any given divergence.

These observations lead to the following very important remark: an invariance factor having the properties listed above will make any divergence invariant with respect to "q".

3.2.1 Calculation of the nominal invariance factor.

In order to obtain an expression of $K(p,q)$, the method defined in [34] consists of calculating the invariance factor:

$$K_0(p,q) = \arg\min_{K>0} D(p\|Kq) \tag{3.3}$$

In this mode of calculation, the invariance factor $K_0(p,q)$, which we will designate by "Nominal invariance factor" is specifically associated with the divergence $D(p\|q)$ and therefore implicitly associated with a basic convex function and a constructive mode of divergence.
So we have to solve with respect to K (positive), the equation:

$$\frac{\partial D(p\|Kq)}{\partial K} = 0 \tag{3.4}$$

If we put the resulting K expression, $K_0(p,q)$, into the divergence $D(p\|q)$, we get the invariant divergence $DI(p\|q) = D(p\|K_0q)$.
In this calculation, K is considered a scalar quantity.

However, it should be noted that (3.4) does not necessarily have an explicit solution.

*** Property.**

Taking into account the definition (3.3), the resulting invariant divergence $D(p\|K_0q)$ is less than or equal to any other invariant divergence derived from $D(p\|q)$ using an invariance factor $K_1(p,q)$ different from $K_0(p,q)$.

$$D(p\|K_0q) \leq D(p\|K_1q) \tag{3.5}$$

With equality if $p_i = q_i \ \forall i$.
Furthermore, when $q \to p$, we have:
$K_0(p,q) \ \to K_1(p,q) \to 1$
therefore
$D(p\|K_0q) \ \to D(p\|K_1q) \ \to \ D(p\|q) \to 0.$

Examples allowing to show this more precisely are given in Annex 7.

3.3 A few comments regarding the invariance factor K.

For divergences that remain invariant by change of scale, whatever their form, we note:

$$D\left(p\|Kq\right)=\sum_{i}d\left(p_{i}\|Kq_{i}\right) \tag{3.6}$$

3.3.1 Fundamental property.

For divergences invariant by scale change on "q", we will first establish a fundamental property which is writtens:

$$\sum_{j}q_{j}\frac{\partial D\left(p\|Kq\right)}{\partial q_{j}}=0 \tag{3.7}$$

In a first step, we show that this relation is verified when the invariance factor is a nominal factor, i.e. when it is calculated explicitly by resolution of (3.4); this invariance factor $K_0\left(p,q\right)$ is thus directly associated to the divergence $D\left(p\|q\right)$ considered.

Démonstration.

Knowing that K is a function of "p" and "q", the gradient of $D\left(p\|Kq\right)$ with respect to "q" is written:

$$\frac{\partial D\left(p\|Kq\right)}{\partial q_{j}}=\sum_{i}\frac{\partial d\left(p_{i}\|Kq_{i}\right)}{\partial\left(Kq_{i}\right)}\frac{\partial\left(Kq_{i}\right)}{\partial q_{j}} \tag{3.8}$$

But, we have:

$$\frac{\partial\left(Kq_{i}\right)}{\partial q_{j}}=\frac{\partial\left(Kq_{i}\right)}{\partial K}\frac{\partial K}{\partial q_{j}}+\frac{\partial\left(Kq_{i}\right)}{\partial q_{i}}\frac{\partial q_{i}}{\partial q_{j}}=q_{i}\frac{\partial K}{\partial q_{j}}+K\delta_{ij} \tag{3.9}$$

which leads to:

$$\frac{\partial D\left(p\|Kq\right)}{\partial q_{j}}=\sum_{i}\frac{\partial d\left(p_{i}\|Kq_{i}\right)}{\partial\left(Kq_{i}\right)}\frac{\partial K}{\partial q_{j}}q_{i}+K\frac{\partial d\left(p_{j}\|Kq_{j}\right)}{\partial\left(Kq_{j}\right)} \tag{3.10}$$

From which it can be deduced:

$$\sum_{j}q_{j}\frac{\partial D\left(p\|Kq\right)}{\partial q_{j}}=\left[\sum_{i}\frac{\partial d\left(p_{i}\|Kq_{i}\right)}{\partial\left(Kq_{i}\right)}q_{i}\right]\left[\sum_{j}q_{j}\frac{\partial K}{\partial q_{j}}+K\right] \tag{3.11}$$

With:

$$q_i = \frac{\partial (Kq_i)}{\partial K} \tag{3.12}$$

one can also write (3.11) in the form:

$$\sum_j q_j \frac{\partial D(p\|Kq)}{\partial q_j} = \left[\sum_i \frac{\partial d(p_i\|Kq_i)}{\partial K}\right] \left[\sum_j q_j \frac{\partial K}{\partial q_j} + K\right] \tag{3.13}$$

The fundamental relationship (3.7) is thus made explicit in the form:

$$\left[\sum_i \frac{\partial d(p_i\|Kq_i)}{\partial K}\right] \left[\sum_j q_j \frac{\partial K}{\partial q_j} + K\right] = 0 \tag{3.14}$$

Or also:

$$\left[\sum_i \frac{\partial d(p_i\|Kq_i)}{\partial (Kq_i)} q_i\right] \left[K + \sum_j q_j \frac{\partial K}{\partial q_j}\right] = 0 \tag{3.15}$$

So it's verified if one of the terms in the product (3.14) is zero.
We examine each of these two terms in turn.

First term of the product.

The nominal invariance factor K_0 is calculated by solving with respect to K the equation:

$$\frac{\partial D(p\|Kq)}{\partial K} = \sum_i \frac{\partial d(p_i\|Kq_i)}{\partial K} = 0 \tag{3.16}$$

that is:

$$\sum_i q_i \frac{\partial d(p_i\|Kq_i)}{\partial (Kq_i)} = 0 \tag{3.17}$$

For such an invariance factor $K_0(p,q)$, specifically associated with the divergence $D(p\|q)$ considered, the first term of the product (3.14) is null; we can therefore conclude that:

The relationship (3.7) is verified if $K(p,q,) = K_0(p,q,)$.

This property is of fundamental importance as we will see in the chapter dealing with the algorithmic developments of minimization under non-negativity and sum constraints.

We first examine the expressions of the term (3.17) corresponding to the different forms of classical divergences.

Relationship with the different construction modes of divergences (Csiszär, Bregman, Jensen).

The factor $K(p,q)$, when calculated by resolving (3.4) (3.17), is specifically related to a given $D(p\|q)$ divergence, so it is associated with both a basic convex function and the constructive mode of the divergence (Csiszär, Bregman or Jensen); generalizations outside this framework will not be considered here.
We will therefore establish the relations that link the K factor and the convex functions that allows us to construct the divergences. To do so, we will give the particular expressions of (3.17) for each of the three constructive modes of divergences.

1 - Csiszär's divergences.

The function $f(x)$ considered here is a standard convex function; the Csiszär divergence $C_f(p\|q)$ is constructed according to the relationship:

$$C_f(p\|q) = \sum_i q_i f\left(\frac{p_i}{q_i}\right) \tag{3.18}$$

The factor $K(p,q)$ making this divergence invariant with respect to "q" is associated with the function "f" and Csiszär's constructive mode. The invariant divergence corresponding to (3.18) will be written:

$$C_f(p\|Kq) = \sum_i c_f(p_i\|Kq_i) = \sum_i Kq_i\ f\left(\frac{p_i}{Kq_i}\right) \tag{3.19}$$

After a few simple calculations, we have:

$$\frac{\partial c_f(p_i\|Kq_i)}{\partial(Kq_i)} = f\left(\frac{p_i}{Kq_i}\right) - \frac{p_i}{Kq_i} f'\left(\frac{p_i}{Kq_i}\right) \tag{3.20}$$

And the equation (3.17) is written:

$$\sum_i q_i \left[f\left(\frac{p_i}{Kq_i}\right) - \frac{p_i}{Kq_i} f'\left(\frac{p_i}{Kq_i}\right)\right] = 0 \tag{3.21}$$

2 - Bregman's divergences.

The invariant divergence associated with this constructive mode is by definition written:

$$B_f(p\|Kq) = \sum_i b_f(p_i\|Kq_i) \tag{3.22}$$

that is:

$$B_f(p\|Kq) = \sum_i \left[f(p_i) - f(Kq_i) - (p_i - Kq_i) f'(Kq_i)\right] \tag{3.23}$$

we then obtain:

$$\frac{\partial b_f(p_i\|Kq_i)}{\partial(Kq_i)} = (Kq_i - p_i) f''(Kq_i) \tag{3.24}$$

And the relationship (3.17) is written:

$$\sum_i q_i (Kq_i - p_i) f''(Kq_i) = 0 \tag{3.25}$$

3 - Jensen's divergences.

By definition, we have:

$$J_f(p\|Kq) = \sum_i j_f(p_i\|Kq_i) \tag{3.26}$$

that is:

$$J_f(p\|Kq) = \sum_i \{\alpha f(p_i) + (1-\alpha) f(Kq_i) - f[\alpha p_i + (1-\alpha) Kq_i]\} \tag{3.27}$$

Therefore:

$$\frac{\partial j_f(p_i\|Kq_i)}{\partial(Kq_i)} = (1-\alpha)\left[f'(Kq_i) - f'(\alpha p_i + (1-\alpha) Kq_i)\right] \tag{3.28}$$

And the relationship (3.17) is written:

$$\sum_i q_i \left[f'(Kq_i) - f'(\alpha p_i + (1-\alpha) Kq_i)\right] = 0 \tag{3.29}$$

4 - Overview

If we make an analysis of the relationships corresponding to the 3 types of divergences considered, we can observe that the relation (3.17) will be satisfied for Csiszär divergences, if:

$$\sum_i q_i \left[f\left(\frac{p_i}{Kq_i}\right) - \frac{p_i}{Kq_i} f'\left(\frac{p_i}{Kq_i}\right) \right] = 0 \tag{3.30}$$

Similarly, this equation will be satisfied for Bregman's divergences if:

$$\sum_i q_i \left(Kq_i - p_i\right) f''\left(Kq_i\right) = 0 \tag{3.31}$$

Finally, this relationship will be satisfied for Jensen's divergences, if:

$$\sum_i q_i \left[f'\left(Kq_i\right) - f'\left(\alpha p_i + (1-\alpha) Kq_i\right) \right] = 0 \tag{3.32}$$

It is quite obvious that the equations (3.30), (3.31) and (3.32) are nothing else but the translation of the equation (3.17) corresponding to the 3 types of divergences.

The resolution of (3.30), (3.31) or (3.32) depending on the type of divergence and more generally of (3.17), when possible, allows us to obtain an expression of the nominal invariance factor $K_0\left(p,q\right)$ and thus to satisfy (3.7).

The question is then: is the relation (3.7) still true if we use an invariance factor different from K_0?

3.3.2 Extension of the fundamental property to "non-nominal" invariance factors.

The fundamental property (3.7) has been expressed in the form:

$$\left[\sum_i \frac{\partial d\left(p_i \| Kq_i\right)}{\partial\left(Kq_i\right)} q_i \right] \left[K + \sum_j q_j \frac{\partial K}{\partial q_j} \right] = 0 \tag{3.33}$$

Or also:

$$\left[\sum_i \frac{\partial d\left(p_i \| Kq_i\right)}{\partial K} \right] \left[K + \sum_j q_j \frac{\partial K}{\partial q_j} \right] = 0 \tag{3.34}$$

The effect of the first product term was discussed in the previous section.

The effect of the second term of the product is examined here.

The relations (3.33) (3.34) are always satisfied if the factor $K(p,q)$ is a solution to the differential equation:

$$K + \sum_j q_j \frac{\partial K}{\partial q_j} = 0 \tag{3.35}$$

This differential equation doesn't depend on the divergence under consideration.

Its resolution allows to extend the fundamental property (3.7) to invariance factors different from K_0, i.e. non-nominal.

3.3.3 Some precisions on the differential equation (3.35).

Clearly, only the "q" dependency is exhibited in this equation. That means that the set of solutions of this equation contains the $K(p,q)$ expressions which satisfy (3.7)(3.14)(3.15).
In order for a solution $K(p,q)$ of the differential equation (3.35) to be an acceptable invariance factor, it must also possess other properties that have already been mentioned:

* $K(p,q)$ **must be a positive scalar**,

on the other hand, and, this is a crucial point already indicated:

*** In order to obtain an invariant divergence by scale-change on "q", the vector "$[K(p,q)\,.\,q]$" must be invariant when multiplying "q" by a constant.**

Finally, in a general way:

*** If $p \to q$, then we must have $K(p,q) \to 1$.**

However, the general solution of (3.35) mentioned in [80] (p.94), is written $\forall j$:

$$K(p,q) = \frac{1}{q_j}\Phi\left(p, \frac{q_1}{q_j}, \frac{q_2}{q_j}, ..., \frac{q_{j-1}}{q_j}, 1, \frac{q_{j+1}}{q_j}, ..., \frac{q_n}{q_j}\right) \tag{3.36}$$

Where Φ is any function.
This means that any function $K(p,q)$ of the form (3.36) will satisfy (3.14)(3.15), i.e. (3.7).
We will observe that with expressions of $K(p,q)$ of this form, the vector $[K(p,q).\,q]$ is invariant when multiplying "q" by a constant. However, among the solutions of the form (3.36), only the positive scalar expressions of $K(p,q)$ can play the role of invariance factors.

Finally, one will be able to check that all the expressions of $K_0(p,q)$ calculated (when possible), by explicit resolution of (3.17) that is to say, according to the case, of (3.30), (3.31) where (3.32) will be of the form (3.36).

These observations induce the following remarkable property:

An expression of $K(p,q)$, positive scalar, as long as it is a solution of the differential equation (3.35), (i.e. as long as it is of the form (3.36)), will make invariant any divergence (because the vector $[K(p,q).\,q]$ is invariant with respect to "q"), and the relation (3.7) will be satisfied.

We can therefore say, in order to globalize these observations concerning the property (3.7), that two cases can arise:

- either the invariance factor is computed by explicit resolution of (3.17), the invariance factor is then referred to as $K_0(p,q)$, this will be the "nominal" invariance factor for the considered divergence.

- or the factor $K(p,q)$ is solution of (3.35) and has the specific properties of the invariance factors, without being solution of (3.17), i.e. without any relation with the starting divergence, then, one always obtains an invariant divergence as shown on the examples of the following paragraph.

- the nominal invariance factors belong to the set of solutions of the differential equation (3.35).

Exemple 1: Kullback-Leibler divergence.

We can consider that this divergence is constructed in the sense of Csiszär

on the standard convex function:

$$f_c(x) = x \log x + 1 - x \tag{3.37}$$

It is written:

$$KL(p\|q) = \sum_i \left[p_i \log \frac{p_i}{q_i} + q_i - p_i \right] \tag{3.38}$$

The calculation of the nominal invariance factor leads to the explicit solution:

$$K_0(p,q) = \frac{\sum_j p_j}{\sum_j q_j} \tag{3.39}$$

This factor will be the solution of the equation (3.30), but this divergence can also be obtained in the Bregman sense by relying on the same convex function (it is the common point between the 2 types of divergences), it will thus be made invariant by the same factor K_0 which is the solution of the equation (3.31). The expression of $K_0(p,q)$ (3.39) can be put in the form (3.36); it is thus the solution of the differential equation (3.35) and thus makes it possible to make any divergence invariant.
This can be verified, for example, on the Euclidean distance:

$$EQM(p\|q) = \sum_i (p_i - q_i)^2 \tag{3.40}$$

With the invariance factor (3.39) ***"which is not the nominal invariance factor for this divergence"***, we obtain:

$$EQM(p\|K_0 q) = \sum_i \left(p_i - \frac{\sum_j p_j}{\sum_j q_j} q_i \right)^2 \tag{3.41}$$

This divergence is invariant under scale change on "q"; it can also be written as follows:

$$EQM(p\|K_0 q) = \left(\sum_j p_j \right)^2 \sum_i \left(\frac{p_i}{\sum_j p_j} - \frac{q_i}{\sum_j q_j} \right)^2 \tag{3.42}$$

or also:

$$EQM(p\|K_0 q) = \left(\sum_j p_j \right)^2 \sum_i (\bar{p}_i - \bar{q}_i)^2 \tag{3.43}$$

When using this very particular form of the invariance factor K given by (3.39), the resulting invariant divergence is analogous (except for one factor that depends only on "p_j") to the initial divergence, where "p_i" has been replaced with $\bar{p}_i = \frac{p_i}{\sum_j p_j}$ and "q_i" has been replaced with $\bar{q}_i = \frac{q_i}{\sum_j q_j}$.
We can check that property on any discrepancies that come up later.

Exemple 2: Mean square deviation (E.Q.M.).

We consider here, the Mean square deviation (Euclidean distance):

$$EQM\left(p\|q\right) = \sum_i \left(p_i - q_i\right)^2 \tag{3.44}$$

This can be considered as being a Bregman divergence based on the standard convex function:

$$f_c\left(x\right) = \left(x-1\right)^2 \tag{3.45}$$

It is made invariant to a scale change on "q" by calculating the nominal factor that is written:

$$K_0\left(p,q\right) = \frac{\sum_i p_i q_i}{\sum_i q_i^2} \tag{3.46}$$

But the E.Q.M. is also a Jensen (1/2) divergence based on the same convex function.
With this expression of $K\left(p,q\right)$, the equation (3.31) will be satisfied because the Euclidean distance is a Bregman divergence, but simultaneously the equation (3.32) will be satisfied because it is also a Jensen divergence, and of course, this expression of $K\left(p,q\right)$ is a solution of the differential equation (3.35).
If we now use the expression of $K_0\left(p,q\right)$ given in (3.46) as the invariance factor in the Kullback-Leibler divergence, we can see that, although this expression is not the nominal invariance factor for this divergence, we obtain an invariant form by scale change on "q".

3.3.4 General form of the invariance factor $K\left(p,q\right)$.

Given these observations, and all the constraints on the invariance factor $K\left(p,q\right)$, a general expression acceptable for K can be written in the form:

$$K\left(p,q\right) = \left(\frac{\sum_i p_i^\alpha q_i^\beta}{\sum_i p_i^\delta q_i^\gamma}\right)^\mu \quad ; \quad \alpha+\beta = \delta+\gamma \quad ; \quad \mu\left(\gamma-\beta\right) = 1 \tag{3.47}$$

Indeed, this expression which can be put in the form (3.36) represents a family of solutions to the differential equation (3.35).
Note that the constraints expressed in (3.47) reflect the following properties:

$$\alpha+\beta=\delta+\gamma \quad \Leftrightarrow \quad K(p,q) \ \textit{is a scalar} \tag{3.48}$$

$$\mu(\gamma-\beta)=1 \quad \Leftrightarrow \quad [K(p,q)\,.\,q] \quad \textit{is invariant with respect to } q \tag{3.49}$$

Taking into account the constraints mentioned in (3.47), this general expression of K can also be written with a smaller number of parameters, in the form:

$$K(p,q)=\left(\frac{\sum_i p_i^{\alpha} q_i^{\beta}}{\sum_i p_i^{\alpha-\frac{1}{\mu}} q_i^{\beta+\frac{1}{\mu}}}\right)^{\mu} \tag{3.50}$$

But, another expression of $K(p,q)$ is more explicit; indeed, taking into account the relations between the parameters, it can be written, after some calculations, in the form:

$$K(p,q)=\left[\sum_i \left(\frac{p_i}{q_i}\right)^{\alpha-\delta} \frac{p_i^{\delta} q_i^{\gamma}}{\sum_j p_j^{\delta} q_j^{\gamma}}\right]^{\frac{1}{\alpha-\delta}} \tag{3.51}$$

These expressions are of the form:

$$K(p,q)=\left[\sum_i \left(\frac{p_i}{q_i}\right)^{t} \frac{p_i^{\delta} q_i^{\gamma}}{\sum_j p_j^{\delta} q_j^{\gamma}}\right]^{\frac{1}{t}} \tag{3.52}$$

These are, for quantities of the form (p_i/q_i), a weighted generalized mean of the order "t" with the exponent $t=\alpha-\delta$ (see.Appendix 2), and weighting coefficients w_i such that $\sum_i w_i=1$ which are given by:

$$w_i=\frac{p_i^{\delta} q_i^{\gamma}}{\sum_j p_j^{\delta} q_j^{\gamma}} \tag{3.53}$$

In the special case of the dual Kullback -Leibler divergence, the nominal invariance factor will be expressed as follows:

$$K_0=\exp\left[\sum_i w_i \log\left(\frac{p_i}{q_i}\right)\right] \tag{3.54}$$

It is a weighted generalized mean of p_i/q_i, based on the function $\psi(x) = \log x$, with weighting factors $\omega_i = \frac{p_i}{\sum_j p_j}$.

$$K_0 = \psi^{-1}\left[\sum_i w_i \psi\left(\frac{p_i}{q_i}\right)\right] \tag{3.55}$$

It can be verified after some calculations that expressions of the invariance factor given by (3.52) and (3.54), associated with weighting coefficients of the form (3.53), satisfy the differential equation (3.35).

3.3.5 Remarks.

Based on this general form, we can go even further in the discussion; by returning to the expression $K(p,q) = \frac{\sum_j p_j q_j}{\sum_j q_j^2}$ calculated explicitly for K.L.'s divergence, we can imagine an expression for the invariance factor written $K(p,q) = \left[\frac{\sum_j p_j^2}{\sum_j q_j^2}\right]^{\frac{1}{2}}$ which is of the form (3.47), but which does not correspond "*a priori*" to any divergence.
This expression of $K(p,q)$ is a solution of the differential equation, so it makes invariant any divergence.
Similarly, from the expression of the invariance factor corresponding to the mean square deviation: $K(p,q) = \frac{\sum_j p_j q_j}{\sum_j q_j^2}$, we can imagine an expression of the invariance factor $K(p,q) = \frac{\sum_j p_j q_j^2}{\sum_j q_j^3}$ which is of the form (3.47), but which does not correspond "*a priori*" to any divergence; this expression of $K(p,q)$ is the solution of the differential equation (3.35); this factor makes invariant any divergence.

3.4 Some properties of the Gradient of an invariant divergence.

We recall the relation (3.10) giving the expression of the gradient with respect to the variable "q" of an invariant divergence $D(p\|Kq)$ for an invariance factor $K(p,q)$:

$$\frac{\partial D(p\|Kq)}{\partial q_j} = \sum_i \frac{\partial d(p_i\|Kq_i)}{\partial(Kq_i)}\frac{\partial K}{\partial q_j}q_i + K\frac{\partial d(p_j\|Kq_j)}{\partial(Kq_j)} \tag{3.56}$$

This expression can also be written:

$$\frac{\partial D\left(p\|Kq\right)}{\partial q_j}=\frac{\partial K}{\partial q_j}\sum_i q_i\frac{\partial d\left(p_i\|Kq_i\right)}{\partial\left(Kq_i\right)}+K\frac{\partial d\left(p_j\|Kq_j\right)}{\partial\left(Kq_j\right)} \tag{3.57}$$

Or also:

$$\frac{\partial D\left(p\|Kq\right)}{\partial q_j}=\frac{\partial K}{\partial q_j}\sum_i \frac{\partial d\left(p_i\|Kq_i\right)}{\partial K}+K\frac{\partial d\left(p_j\|Kq_j\right)}{\partial\left(Kq_j\right)} \tag{3.58}$$

For a given divergence, suppose that the expression of the invariance factor can be calculated explicitly, namely $K_0\left(p,q\right)$ this expression; it is of course of the general form (3.50).
Let us now consider another expression of the invariance factor $K_1\left(p,q\right)$ respecting the general form (3.50), but which is not in correspondence with the divergence considered.

The question is: what happens to the gradient expressions (3.56), (3.57) or (3.58) depending on whether we use $K_0\left(p,q\right)$ or $K_1\left(p,q\right)$?
In the first case, $K_0\left(p,q\right)$ is the solution to the differential equation (3.35), then the first term of the second member of the equation (3.58) is zero and it simply remains:

$$\frac{\partial D\left(p\|Kq\right)}{\partial q_j}=K_0\left[\frac{\partial d\left(p_j\|Kq_j\right)}{\partial\left(Kq_j\right)}\right]_{K=K_0} \tag{3.59}$$

If $K=K_1$, the first term of the second member of the equation (3.58) is no longer zero, and we have:

$$\frac{\partial D\left(p\|Kq\right)}{\partial q_j}=\frac{\partial K_1}{\partial q_j}\left[\sum_i \frac{\partial d\left(p_i\|Kq_i\right)}{\partial K}\right]_{K=K_1}+K_1\left[\frac{\partial d\left(p_j\|Kq_j\right)}{\partial\left(Kq_j\right)}\right]_{K=K_1} \tag{3.60}$$

However, in both cases, one still has the fundamental property:

$$\sum_j q_j\frac{\partial D\left(p\|Kq\right)}{\partial q_j}=\sum_j q_j\frac{\partial DI\left(p\|q\right)}{\partial q_j}=0 \tag{3.61}$$

Some examples to show this in more detail are given in Appendix 4.

3.5 A very special case of the invariance factor.

We saw that a particular expression of the invariance factor could be:

$$K(p,q) = \frac{\sum_j p_j}{\sum_j q_j} \tag{3.62}$$

This expression is the explicit result of calculating the invariance factor for a Kullback-Leibler divergence.
Using this factor, any divergence can be made invariant.

Forms of the divergences related to this invariance factor.

By introducing this invariance factor with respect to "q", in any divergence, one makes it appear in the resulting invariant divergence, normalized variables noted $\bar{p}_i = \frac{p_i}{\sum_j p_j}$ and $\bar{q}_i = \frac{q_i}{\sum_j q_j}$, so that we systematically obtain invariant divergences corresponding to $\bar{p}$ and $\bar{q}$ variables of sum equal to 1.

Except for a multiplicative factor (which depends only on $\sum_j p_j$), the resulting invariant divergences have the same expression as the initial divergences, with the normalized variables simply replacing the initial variables. This operation being carried out, some simplifications can appear.
One thus obtains invariant divergences similar to the simplified divergences applicable to densities of probabilities, provided that one introduces explicitly in the simplified divergences, normalized variables.

Moreover, if we disregard the multiplicative factor, these divergences are invariant not only with respect to the "q" variable, but also with respect to the "p" variable, as can be seen from the examples given in Appendix 6.

chapter 4 - Divergences and Entropies

This chapter presents the divergences that can be related to the various forms of entropy found in the literature. In most cases they will be constructed in the sense of Csiszär on the basis of a convex function; in this constructive mode, we will be led to distinguish the case of *"standard convex functions"*. The relation with the entropies will of course be rather related to the *"simple convex functions"*.
Finally, the extensions of the divergences by using the function "Generalized Logarithm" (see Appendix 1), will lead us to deviate from the constructive mode of Csiszär, but will make it possible to make the connection with the entropies of Sharma-Mittal [89] and Renyi [82] [81].
In a final section, we'll discuss Jensen's differences based on Entropies.
General references for this chapter will be given in: [6],[8],[10],[78] et [91].

4.1 Shannon Entropy related divergences.

In this section, the divergences linked to Shannon's entropy [88] are analyzed.

4.1.1 Direct form.

The standard convex function used in this case, see.figure (4.1), is written:

$$f_c(x) = x \log x + 1 - x \tag{4.1}$$

The corresponding Csiszär's divergence is the Kullback-Leibler divergence [57]:

$$KL(p\|q) = \sum_i p_i \log \frac{p_i}{q_i} + q_i - p_i \tag{4.2}$$

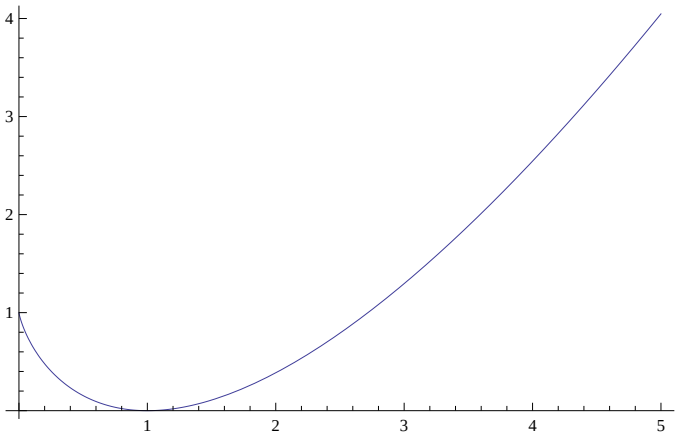

Figure 4.1: Function $f_c(x) = x \log x + 1 - x$

The gradient with respect to "q" is written:

$$\frac{\partial KL(p\|q)}{\partial q_j} = -\frac{p_j}{q_j} + 1 \tag{4.3}$$

It will be zero for $p_j = q_j \ \forall j$.
If, at this point, we consider that we have $\sum_i p_i = \sum_i q_i$, the divergence (4.2) simplifies and is written:

$$IKL(p\|q) = \sum_i p_i \log \frac{p_i}{q_i} \tag{4.4}$$

That's Kullback's information [8].
The gradient with respect to "q" is written:

$$\frac{\partial IKL(p\|q)}{\partial q_j} = -\frac{p_j}{q_j} \tag{4.5}$$

We can see that this gradient will never be equal to zero, therefore $IKL(p\|q)$ is not usable in our problem without special precautions; indeed, to use such a divergence, it will be necessary to explicitly introduce the fact that $\sum_i p_i = \sum_i q_i$.
This difficulty will appear every time one wants to use simplified divergences, that is to say, those built on "simple" convex functions.
An additional simplification such as $\sum_i p_i = \sum_i q_i = 1$ will not bring any additional simplification, but if we don't make this assumption, the divergence (4.4) will probably not be positive.

The form (4.4) is mainly used in works dealing with probability densities; this form is deduced in the Csiszär sense from the simple convex function (see figure 4.2):

$$f(x) = x \log x \tag{4.6}$$

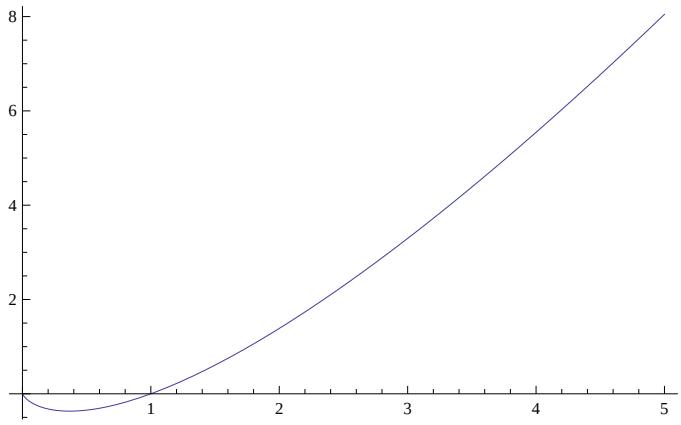

Figure 4.2: Function $f(x) = x \log x$

Invariance by change of scale.

The invariance factor $K_0(p, q)$ corresponding to the Kullback-Leibler divergence (4.2) can be calculated according to the method indicated in Chapter 3; it is expressed in explicit form:

$$K_0 = \frac{\sum_j p_j}{\sum_j q_j} \tag{4.7}$$

It is a special case already mentioned in Chapter 3 that leads to the invariant divergence:

$$KLI(p\|q) = \sum_j p_j \sum_i \bar{p}_i \log \frac{\bar{p}_i}{\bar{q}_i} \tag{4.8}$$

with $\bar{p}_j = \frac{p_j}{\sum_l p_l}$ et $\bar{q}_j = \frac{q_j}{\sum_l q_l}$.
We can observe that we get a divergence equivalent to (4.4), but here we explicitly have $\sum_i \bar{p}_i = \sum_i \bar{q}_i (= 1)$.
On the other hand, disregarding the multiplicative factor $\sum_j p_j$, we can observe that the divergence obtained is invariant not only with respect to

"q" but also with respect "p".
Its gradient with respect to "q" is written:

$$\frac{\partial KLI\left(p\|q\right)}{\partial q_l} = \frac{1}{\sum_j q_j}\left(1 - \frac{\bar{p}_l}{\bar{q}_l}\right) \tag{4.9}$$

4.1.2 Dual form.

To obtain the dual form, we are using the mirror function of (4.1) (see figure 4.3), which is written:

$$\breve{f}_c\left(x\right) = \log\frac{1}{x} + x - 1 \tag{4.10}$$

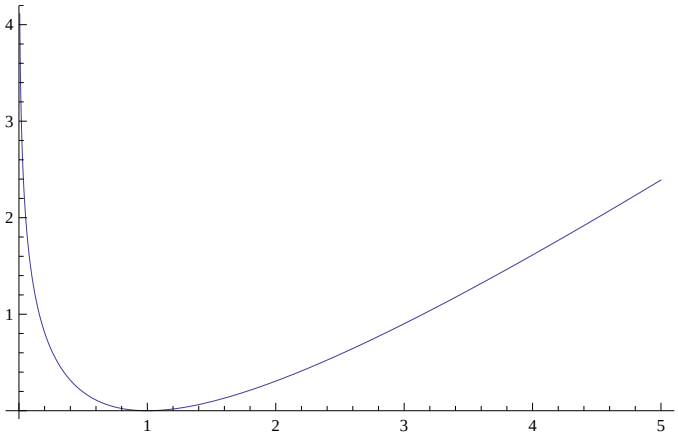

Figure 4.3: Function $\breve{f}_c\left(x\right) = \log\frac{1}{x} + x - 1$

So, we obtain:

$$KL\left(q\|p\right) = \sum_i q_i \log\frac{q_i}{p_i} + p_i - q_i \tag{4.11}$$

The gradient with respect to "q" is written:

$$\frac{\partial KL\left(q\|p\right)}{\partial q_j} = \log\frac{q_j}{p_j} \tag{4.12}$$

It will be zero for $p_j = q_j \ \forall j$.
Again, assuming that $\sum_i p_i = \sum_i q_i$, we have the simplified form:

$$IKL\left(q\|p\right) = \sum_i q_i \log\frac{q_i}{p_i} \tag{4.13}$$

The gradient with respect to "q" is written:

$$\frac{\partial IKL\left(q\|p\right)}{\partial q_j} = \log \frac{q_j}{p_j} + 1 \tag{4.14}$$

This gradient will not be zero for $p_j = q_j$, therefore $IKL\left(q\|p\right)$ is not usable in our problem without special precautions.
An additional specification such as $\sum_i p_i = \sum_i q_i = 1$ will not provide any additional simplification, but if we don't make that assumption, the divergence (4.13) is unlikely to be positive.
The divergence (4.13) is constructed in the Csiszär sense on the simple convex function (see figure (4.4):

$$\breve{f}\left(x\right) = \log \frac{1}{x} \tag{4.15}$$

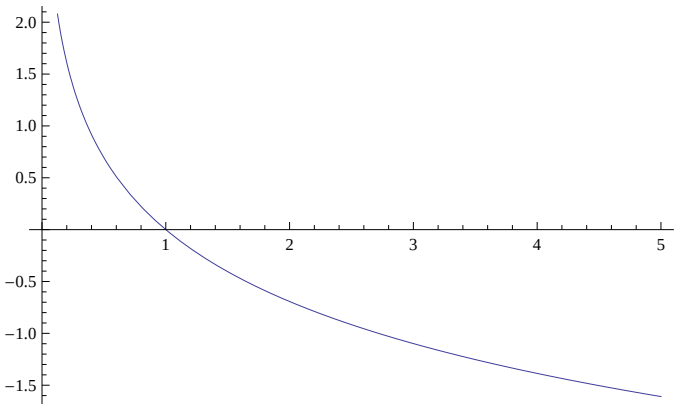

Figure 4.4: Function $\breve{f}\left(x\right) = \log \frac{1}{x}$

Invariant form of dual divergence (4.11).

The derivation of the nominal invariance factor for this divergence leads to the expression:

$$K_0 = \exp \sum_i \frac{q_i}{\sum_j q_j} \log \frac{p_i}{q_i} \tag{4.16}$$

It is a weighted generalized average of terms of the form (p_i/q_i) that is written:

$$K_0 = \psi^{-1}\left[\sum_i w_i \psi\left(\frac{p_i}{q_i}\right)\right] \tag{4.17}$$

With $\psi(x) = \ln(x)$ and weighting factors $w_i = \frac{q_i}{\sum_j q_j}$.
By introducing this invariance factor in the divergence (4.11), we obtain after some simple calculations, the invariant form that can be written:

$$KLI(q\|p) = \sum_i \left[p_i - q_i \exp \sum_j \frac{q_j}{\sum_l q_l} \log \frac{p_j}{q_j} \right] = \sum_i [p_i - K_0 q_i] \quad (4.18)$$

It is easily verified that this divergence is not modified if "q" is multiplied by a positive constant.

4.1.3 Symmetrical form.

It is obtained in the sense of Jeffreys [52] by using Csiszär's constructive mode, based on the standard convex function (see figure 4.5):

$$\hat{f}_c(x) = \frac{1}{2}\left[f_c(x) + \breve{f}_c(x)\right] = \frac{1}{2}\left[f(x) + \breve{f}(x)\right] \quad (4.19)$$

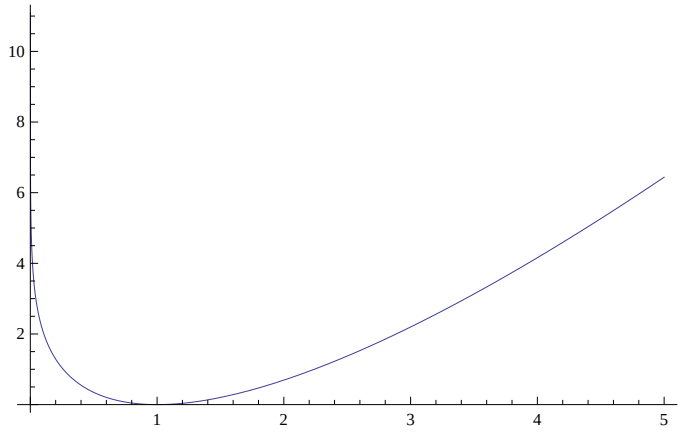

Figure 4.5: Function $\hat{f}_c(x) = x \log x - \log x$

So, we have:

$$KL(p,q) = \sum_i (p_i - q_i) \log \frac{p_i}{q_i} \quad (4.20)$$

Whose gradient with respect to "q" is written:

$$\frac{\partial KL(p,q)}{\partial q_j} = \frac{q_j - p_j}{q_j} - \log \frac{p_j}{q_j} \quad (4.21)$$

This expression is of course equal to zero for $p_j = q_j \ \forall j$.

4.2 Havrda-Charvat Entropy related divergences.

This section deals with Csiszär's divergences related to the entropy of Havrda-Charvat [42] and more specifically based on the standard convex function:

$$f_c(x) = \frac{1}{\alpha(\alpha-1)}\left[x^\alpha - \alpha x + (1-\alpha)\right] \tag{4.22}$$

To these divergences we will associate those built on the corresponding simple convex functions:

$$f_1(x) = \frac{1}{\alpha(\alpha-1)}\left[x^\alpha - x\right] \tag{4.23}$$

and

$$f_2(x) = \frac{1}{\alpha(\alpha-1)}\left[x^\alpha - 1\right] \tag{4.24}$$

We will see in the **chapter 5** that the Csiszär divergences built on these convex functions, in particular on the function (4.22), belong to the class of "Alpha divergences" of Amari [2].
The divergence associated with $f_c(x)$ will be written:

$$HC_\alpha(p\|q) = \frac{1}{\alpha(\alpha-1)}\left\{\sum_i p_i^\alpha q_i^{1-\alpha} - \sum_i \left[\alpha p_i + (1-\alpha) q_i\right]\right\} \tag{4.25}$$

It's clearly a difference between a generalized geometric mean and a generalized arithmetic mean.
The gradient with respect to "q" is given by:

$$\frac{\partial HC_\alpha(p\|q)}{\partial q_j} = \frac{1}{\alpha}\left(1 - p_j^\alpha q_j^{-\alpha}\right) \tag{4.26}$$

It will be zero for $p_j = q_j \ \forall j$.
The divergences related to $f_1(x)$ and $f_2(x)$ will be written respectively:

$$A_1(p\|q) = \frac{1}{\alpha(\alpha-1)}\left[\sum_i p_i^\alpha q_i^{1-\alpha} - p_i\right] \tag{4.27}$$

and

$$A_2(p\|q) = \frac{1}{\alpha(\alpha-1)}\left[\sum_i p_i^\alpha q_i^{1-\alpha} - q_i\right] \tag{4.28}$$

We can notice that if in the expression (4.25), one makes $\sum_i p_i = \sum_i q_i$, one finds (4.27) or (4.28).
The corresponding gradients with respect to "q" will be written respectively:

$$\frac{\partial A_1(p\|q)}{\partial q_j} = \frac{1}{\alpha}\left(p_j^{\alpha} q_j^{-\alpha}\right) \tag{4.29}$$

that will never be zero, and:

$$\frac{\partial A_2(p\|q)}{\partial q_j} = \frac{1}{\alpha(\alpha-1)}\left[(1-\alpha)\, p_j^{\alpha} q_j^{-\alpha} - 1\right] \tag{4.30}$$

which may be zero, but for $p_j \neq q_j$.
Furthermore, if we make the additional assumption that $\sum_i p_i = \sum_i q_i = 1$, we get 3 identical divergences that are written:

$$HCS_{\alpha}(p\|q) = \frac{1}{\alpha(\alpha-1)}\left[\sum_i p_i^{\alpha} q_i^{1-\alpha} - 1\right] \tag{4.31}$$

whose the gradient relative to "q" given by (4.29). This last divergence (4.31) is the Havrda-Charvat divergence [42] quoted by Arndt [6] and Basseville [10].
The dual divergences are constructed by using as convex functions, the mirror functions of the previous ones; they will be indicated in **chapter 5**.

4.3 Sharma-Mittal Entropy related divergences.

The various divergences highlighted in this section are no longer Csiszär's divergences, in fact, they cannot be obtained with this constructive method, but they can be considered as a form of generalisation of the divergences obtained in the previous section. They are related to the entropy of Sharma-Mittal [89].
This generalization can be summarized by the following simple rule.

Rule: When a divergence is expressed as the difference between two positive terms, it can be generalized by applying to each of the terms an increasing function (which will not change the sign of the divergence obtained); the increasing function used is often the "Generalized Logarithm" function (see Appendix 1), which makes it possible to change, by action on a single parameter, from the linear function (which leaves the initial divergence unchanged) to the logarithmic function itself.

Of course, at the end of this operation, we obtain a new divergence whose convexity properties are not guaranteed, even if the initial divergence was convex.
If we apply this rule on the divergence $HC_\alpha(p\|q)$ (4.25), using the generalized logarithm with the exponent $1-d=\frac{s-1}{\alpha-1}$, we obtain:

$$SM_{\alpha,s}(p\|q)=\frac{1}{\alpha(s-1)}\left\{\left[\sum_i p_i^\alpha q_i^{1-\alpha}\right]^{\frac{s-1}{\alpha-1}}-\left[\sum_i \alpha p_i+(1-\alpha)q_i\right]^{\frac{s-1}{\alpha-1}}\right\} \tag{4.32}$$

Similarly, if we apply this rule on the very simplified divergence (4.31), we get:

$$SMS_{\alpha,s}(p\|q)=\frac{1}{\alpha(s-1)}\left\{\left[\sum_i p_i^\alpha q_i^{1-\alpha}\right]^{\frac{s-1}{\alpha-1}}-1\right\} \tag{4.33}$$

It's this last divergence that is commonly referred to as the Sharma-Mittal divergence [89] [6].
Calculating the gradient of $SM_{\alpha,s}(p\|q)$ (4.32) with respect to "q" gives:

$$\frac{\partial SM_{\alpha,s}(p\|q)}{\partial q_j}=\frac{1}{\alpha}\left\{\left[\sum_i \alpha p_i+(1-\alpha)q_i\right]^{\frac{s-\alpha}{\alpha-1}}-\left[\sum_i p_i^\alpha q_i^{1-\alpha}\right]^{\frac{s-\alpha}{\alpha-1}}p_j^\alpha q_j^{-\alpha}\right\} \tag{4.34}$$

He'll cancel for $p_i=q_i\ \forall i$.
On the other hand, the calculation of the gradient of $SMS_\alpha(p\|q)$ (4.33) with respect to "q" yields:

$$\frac{\partial SMS_{\alpha,s}(p\|q)}{\partial q_j}=-\frac{1}{\alpha}\left[\sum_i p_i^\alpha q_i^{1-\alpha}\right]^{\frac{s-\alpha}{\alpha-1}}p_j^\alpha q_j^{-\alpha} \tag{4.35}$$

It will never cancel.

Scale invariance with respect to "q".

The divergence given by the relation (4.32) is made invariant using the invariance factor $K^*(p,q)=\frac{\sum_j p_j}{\sum_j q_j}$ which is not the nominal invariance factor; it is written after simplifications:

$$SM_{\alpha,s}I(p\|q)=\frac{1}{\alpha(s-1)}\left\{\left[\sum_i \bar{p}_i^\alpha \bar{q}_i^{1-\alpha}\right]^{\frac{s-1}{\alpha-1}}-\left[\sum_i \alpha\bar{p}_i+(1-\alpha)\bar{q}_i\right]^{\frac{s-1}{\alpha-1}}\right\} \tag{4.36}$$

Its gradient with respect to "q" is written:

$$\frac{\partial SM_{\alpha,s}I\left(p\|q\right)}{\partial q_l} = \frac{1}{\alpha\sum_j q_j}\left[\sum_i \bar{p}_i^{\alpha}\bar{q}_i^{1-\alpha}\right]^{\frac{s-\alpha}{\alpha-1}}\left(\sum_i \bar{p}_i^{\alpha}\bar{q}_i^{1-\alpha} - \bar{p}_l^{\alpha}\bar{q}_l^{\alpha}\right) \quad (4.37)$$

4.4 Renyi Entropy related divergences.

These divergences related to Renyi's entropy [81] [82] [6] [14], correspond to the limit $d \to 1$ in the "generalized Logarithm", that is $s \to 1$ in the divergences of the previous section.
If we perform this operation on the expressions (4.32) and (4.33), we get respectively:

$$R_{\alpha}\left(p\|q\right) = \frac{1}{\alpha\left(\alpha-1\right)}\left\{\log\sum_i p_i^{\alpha}q_i^{1-\alpha} - \log\sum_i \alpha p_i + \left(1-\alpha\right)q_i\right\} \quad (4.38)$$

The form (4.38) can be seen as an extension of Renyi's divergence to data fields whose sum is not equal to 1.
In the case of probability densities, i.e. with $\sum_i p_i = \sum_i q_i = 1$, it comes:

$$RS_{\alpha}\left(p\|q\right) = \frac{1}{\alpha\left(\alpha-1\right)}\left\{\log\sum_i p_i^{\alpha}q_i^{1-\alpha}\right\} \quad (4.39)$$

The expression (4.39) is Renyi's divergence in its classical form related to Renyi's entropy and probability densities [6].
The gradients with respect to "q" can be deduced from the gradient expressions (4.34) and (4.35) by making $s = 1$; they are written respectively:

$$\frac{\partial R_{\alpha}\left(p\|q\right)}{\partial q_j} = \frac{1}{\alpha}\left\{\left[\sum_i \alpha p_i + \left(1-\alpha\right)q_i\right]^{-1} - \left[\sum_i p_i^{\alpha}q_i^{1-\alpha}\right]^{-1}p_j^{\alpha}q_j^{-\alpha}\right\} \quad (4.40)$$

and

$$\frac{\partial RS_{\alpha}\left(p\|q\right)}{\partial q_j} = -\frac{1}{\alpha}\left[\sum_i p_i^{\alpha}q_i^{1-\alpha}\right]^{-1}p_j^{\alpha}q_j^{-\alpha} \quad (4.41)$$

We can observe that $\frac{\partial R_{\alpha}(p\|q)}{\partial q_j}$ will be equal to zero if $p_i = q_i \ \forall i$, whereas $\frac{\partial RS_{\alpha}(p\|q)}{\partial q_j}$ could never be zero.

Scale invariance with respect to "q".

For this divergence given by the relation (4.38), the invariance factor cannot be calculated explicitly, so we use as invariance factor the expression $K^*(p,q) = \frac{\sum_j p_j}{\sum_j q_j}$.
In these conditions, the invariant divergence is written after simplifications:

$$RI_\alpha(p\|q) = \frac{\log \sum_j p_j}{\alpha(\alpha-1)} \left[\log \sum_i \bar{p}_i^{\alpha} \bar{q}_i^{1-\alpha} \right] \tag{4.42}$$

In this expression, the multiplicative factor $\log \sum_j p_j$ can be omitted, and the gradient with respect to "q" is given by:

$$\frac{\partial RI_\alpha(p\|q)}{\partial q_l} = \frac{1}{\alpha \sum_j q_j} \left[1 - \frac{\bar{p}_l^{\alpha} \bar{q}_l^{-\alpha}}{\sum_j \bar{p}_j^{\alpha} \bar{q}_j^{1-\alpha}} \right] \tag{4.43}$$

Of course, it will be equal to zero if $p_i = q_i \quad \forall i$.

4.5 Arimoto entropy related divergences.

4.5.1 Direct form.

These divergences developed by Osterreicher [74] [76] rely on the use of generalized averages as in the Arimoto Entropy [4]. In their initial form, they are constructed in the Csiszär sense on the basis of the standard convex function shown in figure (4.6) for $\delta = 2$:

$$f_{c\delta}(x) = \frac{1}{\delta - 1} \left[\left(\frac{1+x^\delta}{2} \right)^{\frac{1}{\delta}} - \left(\frac{1+x}{2} \right) \right] \quad \delta \neq 1 \tag{4.44}$$

This leads to the divergence:

$$AR(p\|q) = \frac{1}{\delta - 1} \left[\sum_i \left(\frac{p_i^\delta + q_i^\delta}{2} \right)^{\frac{1}{\delta}} - \sum_i \left(\frac{p_i + q_i}{2} \right) \right] \tag{4.45}$$

This divergence is symmetrical, so it can be denoted as $AR(p,q)$.
The second term is clearly the unweighted arithmetic mean of the two data fields, whereas the first term corresponds, according to the value of "δ", to the different unweighted means between the two fields; indeed:
* if $\delta = 2$, the first term is the square root mean.

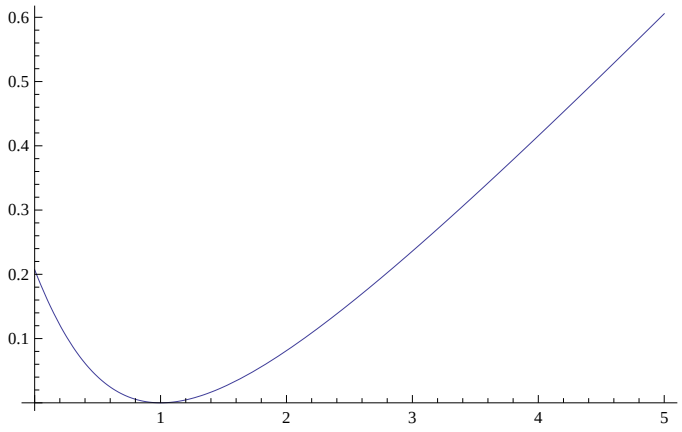

Figure 4.6: Function $f_{c\delta}(x) = \frac{1}{\delta - 1}\left[\left(\frac{1+x^{\delta}}{2}\right)^{\frac{1}{\delta}} - \left(\frac{1+x}{2}\right)\right], \delta = 2$

* if δ =1, the first term is the harmonic mean.
* if $\delta \to 0$, the first term is the geometric mean.
This approach allows us to find the divergences between averages, which will be developed in **chapter 7**.
The gradient with respect to "q" will be written:

$$\frac{\partial AR(p\|q)}{\partial q_j} = \frac{1}{2(\delta - 1)}\left[\left(\frac{p_j^{\delta} + q_j^{\delta}}{2}\right)^{\frac{1-\delta}{\delta}} q_j^{\delta - 1} - 1\right] \tag{4.46}$$

This gradient is equal to zero if $p_i = q_i \ \forall i$.
Since the basic convex function is a standard convex function, we can try to show the associated simple convex functions; after a few thinking about them, we can show 2 of these functions (as always), which are represented on figures (4.7) and (4.8) for $\delta = 2$::

$$f_{1,\delta}(x) = \frac{1}{\delta - 1}\left[\left(\frac{1 + x^{\delta}}{2}\right)^{\frac{1}{\delta}} - x\right] \tag{4.47}$$

and

$$f_{2,\delta}(x) = \frac{1}{\delta - 1}\left[\left(\frac{1 + x^{\delta}}{2}\right)^{\frac{1}{\delta}} - 1\right] \tag{4.48}$$

They will respectively lead to the divergences:

$$ARS1(p\|q) = \frac{1}{\delta - 1}\left[\sum_i \left(\frac{p_i^{\delta} + q_i^{\delta}}{2}\right)^{\frac{1}{\delta}} - \sum_i p_i\right] \tag{4.49}$$

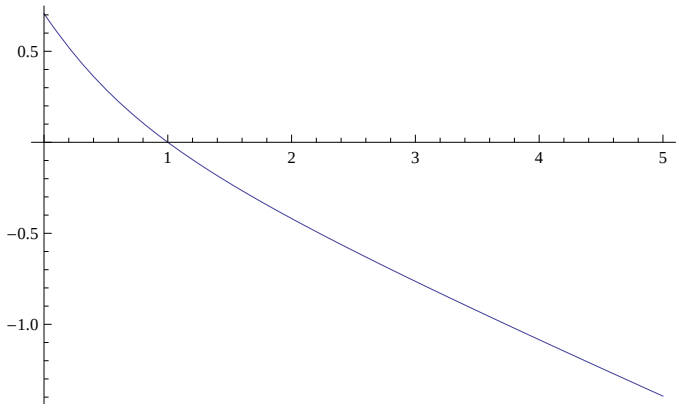

Figure 4.7: Function $f_{1,\delta}(x) = \frac{1}{\delta-1}\left[\left(\frac{1+x^{\delta}}{2}\right)^{\frac{1}{\delta}} - x\right], \delta = 2$

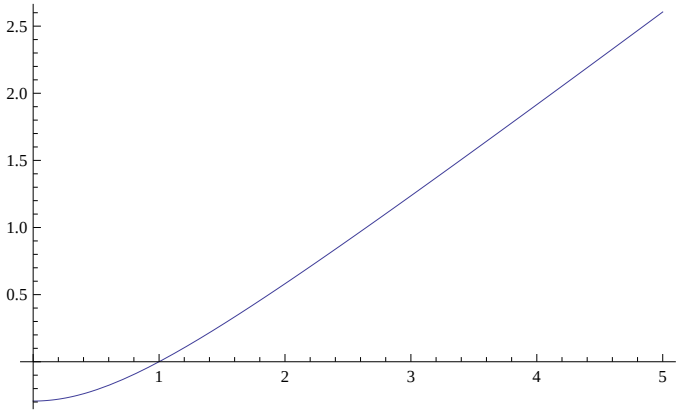

Figure 4.8: Function $f_{2,\delta}(x) = \frac{1}{\delta-1}\left[\left(\frac{1+x^{\delta}}{2}\right)^{\frac{1}{\delta}} - 1\right], \delta = 2$

and

$$ARS2(p\|q) = \frac{1}{\delta-1}\left[\sum_i \left(\frac{p_i^{\delta}+q_i^{\delta}}{2}\right)^{\frac{1}{\delta}} - \sum_i q_i\right] \tag{4.50}$$

If in these divergences as well as in $AR(p\|q)$ one introduces the simplification $\sum_i p_i = \sum_i q_i = 1$, one obtains the divergence of Arimoto [6]:

$$ARS(p\|q) = \frac{1}{\delta-1}\left[\sum_i \left(\frac{p_i^{\delta}+q_i^{\delta}}{2}\right)^{\frac{1}{\delta}} - 1\right] \tag{4.51}$$

Of course, such a divergence can only be used to compare data fields whose sum is explicitly equal to 1.
The gradients with respect to "q" of the $ARS1\left(p\|q\right)$ and $ARS\left(p\|q\right)$ divergences will never cancel, while the gradient of $ARS2\left(p\|q\right)$ may become zero, but not for $p_i = q_i \ \forall i$.

4.5.2 Related dual divergences.

They are built on the dual convex functions of those used in the previous section.
We can note that $\breve{f}_{1,\delta}\left(x\right) = f_{2,\delta}\left(x\right)$ and that $\breve{f}_{2,\delta}\left(x\right) = f_{1,\delta}\left(x\right)$.

4.5.3 Symmetrical divergence.

Taking into account the remark in the previous section, it is constructed, in the sense of Jeffreys, on the basis of $\hat{f}_{\delta}\left(x\right) = \frac{\breve{f}_{1,\delta}(x)+f_{1,\delta}(x)}{2} = \frac{\breve{f}_{2,\delta}(x)+f_{2,\delta}(x)}{2} = f_{c,\delta}\left(x\right)$, and of course leads to the divergence $AR\left(p,q\right)$.

4.5.4 Weighted versions of these divergences.

In a first simple variant of these divergences, weighted versions of these divergences can be introduced by replacing the first term of the divergence $AR\left(p,q\right)$ (4.45) by a weighted generalized mean and second term by a weighted arithmetic mean.
This is equivalent to constructing a Csiszâr's divergence by using the basic standard convex function:

$$f_{c,\delta,\alpha}\left(x\right) = \frac{1}{\left(1-\alpha\right)\left(\delta-1\right)}\left\{\left[\alpha x^{\delta} + \left(1-\alpha\right)\right]^{\frac{1}{\delta}} - \left[\alpha x + \left(1-\alpha\right)\right]\right\} \tag{4.52}$$

With $0 \leq \alpha \leq 1$, this leads to the divergence:

$$AR_{\delta,\alpha}\left(p\|q\right) = \frac{1}{\left(1-\alpha\right)\left(\delta-1\right)}\left\{\sum_i \left[\alpha p_i^{\delta} + \left(1-\alpha\right) q_i^{\delta}\right]^{\frac{1}{\delta}} - \sum_i \left[\alpha p_i + \left(1-\alpha\right) q_i\right]\right\} \tag{4.53}$$

Of course, by varying the "δ" values as indicated above, we can review in the first term, the various weighted means.
Note that the symmetry property of the unweighted version has disappeared; it only exists for $\alpha = 1/2$.
The gradient with respect to "q" will be written:

$$\frac{\partial AR_{\delta,\alpha}\left(p\|q\right)}{\partial q_j} = \frac{1}{\delta-1}\left\{\left[\alpha p_j^{\delta} + \left(1-\alpha\right) q_j^{\delta}\right]^{\frac{1-\delta}{\delta}} q_j^{\delta-1} - 1\right\} \tag{4.54}$$

It will be zero if $p_j = q_j \ \forall j$.
One can observe that a simplified version of this divergence corresponding to $\sum_i p_i = \sum_i q_i = 1$ exists, but the gradient of this simplified form will not be equal to zero if $p_j = q_j \ \forall j$.

4.5.5 First type of extension.

In order to retrieve any divergences based on differences between the weighted means developed in **chapter 7** (the unweighted case will be obtained immediately with $\alpha = \frac{1}{2}$), we can consider, as proposed in [66], the standard convex function:

$$f_{c,\delta,\gamma,\alpha}(x) = \frac{1}{(1-\alpha)(\delta-1)} \left\{ \left[\alpha x^{\delta} + (1-\alpha)\right]^{\frac{1}{\delta}} - \left[\alpha x^{\gamma} + (1-\alpha)\right]^{\frac{1}{\gamma}} \right\} \tag{4.55}$$

With such a function, the resulting Csiszär divergence is written:

$$\begin{aligned} AR_{\delta,\gamma,\alpha}(p\|q) = & \frac{1}{(1-\alpha)(\delta-1)} \left\{ \sum_i \left[\alpha p_i^{\delta} + (1-\alpha) q_i^{\delta}\right]^{\frac{1}{\delta}} \right. \\ & \left. - \sum_i \left[\alpha p_i^{\gamma} + (1-\alpha) q_i^{\gamma}\right]^{\frac{1}{\gamma}} \right\} \end{aligned} \tag{4.56}$$

The divergences obtained for different values of the parameters "γ" and "δ" are summarized in the following table.

δ \ γ	-1	0	1	2
-1	0	GM-HM	AM-HM	SM-HM
0	HM-GM	0	AM-GM	SM-GM
1			Jensen-Shannon	
2	SM-HM	SM-GM	SM-AM	0

Table 4.1: Divergences resulting from various values of "γ" and "δ".

The divergences not listed in this table are either negative or non-convex; moreover the Jensen-Shannon divergence implies the computation of limits $\delta \to 1$, $\gamma \to 1$.
There is no simplified version of the divergence (4.56); its gradient with

respect to "q" is written:

$$\frac{\partial AR_{\delta,\gamma,\alpha}(p\|q)}{\partial q_j} = \frac{1}{\delta-1}\left\{\left[\alpha p_j^{\delta}+(1-\alpha)\,q_j^{\delta}\right]^{\frac{1-\delta}{\delta}} q_j^{\delta-1} - \left[\alpha p_j^{\gamma}+(1-\alpha)\,q_j^{\gamma}\right]^{\frac{1-\gamma}{\gamma}} q_j^{\gamma-1}\right\} \tag{4.57}$$

Such a gradient will be zero pour $p_i = q_i \ \forall i$

4.5.6 Extension in the sense of the Generalized Logarithm.

Since the divergence (4.56) shows a difference of 2 positive terms, an increasing function can be applied to each of the two terms, for example, the generalized logarithm, with the exponent "$1-d$" noted $1-d=\frac{s-1}{\delta-1}$; this gives the divergence:

$$ARSM_{\delta,\gamma,\alpha,s}(p\|q) = \frac{1}{(1-\alpha)(s-1)}\left\{\left[\sum_i\left[\alpha p_i^{\delta}+(1-\alpha)\,q_i^{\delta}\right]^{\frac{1}{\delta}}\right]^{\frac{s-1}{\delta-1}} - \left[\sum_i\left[\alpha p_i^{\gamma}+(1-\alpha)\,q_i^{\gamma}\right]^{\frac{1}{\gamma}}\right]^{\frac{s-1}{\delta-1}}\right\} \tag{4.58}$$

From this expression, the operation $s\to\delta$ $(d\to 0)$ leads to the divergence of the previous section, while the transition $s\to 1$ $(d\to 1)$ leads to the following logarithmic form:

$$LARSM_{\delta,\gamma,\alpha}(p\|q) = \frac{1}{(1-\alpha)(\delta-1)}\left\{\log\sum_i\left[\alpha p_i^{\delta}+(1-\alpha)\,q_i^{\delta}\right]^{\frac{1}{\delta}} - \log\sum_i\left[\alpha p_i^{\gamma}+(1-\alpha)\,q_i^{\gamma}\right]^{\frac{1}{\gamma}}\right\} \tag{4.59}$$

The specific cases corresponding to the various values of "δ" and "γ" will be found in **chapter 7**.
The gradients with respect to "q" of these two divergences are written re-

spectively:

$$\frac{\partial ARSM_{\delta,\gamma,\alpha,s}(p\|q)}{\partial q_j} = \frac{1}{\delta-1}\left\{\left[\sum_i\left[\alpha p_i^\delta+(1-\alpha)\,q_i^\delta\right]^{\frac{1}{\delta}}\right]^{\frac{s-\delta}{\delta-1}}\left[\alpha p_j^\delta+(1-\alpha)\,q_j^\delta\right]^{\frac{1-\delta}{\delta}}q_j^{\delta-1}\right.$$
$$\left.-\left[\sum_i\left[\alpha p_i^\gamma+(1-\alpha)\,q_i^\gamma\right]^{\frac{1}{\gamma}}\right]^{\frac{s-\delta}{\delta-1}}\left[\alpha p_j^\gamma+(1-\alpha)\,q_j^\gamma\right]^{\frac{1-\gamma}{\gamma}}q_j^{\gamma-1}\right\} \tag{4.60}$$

and

$$\frac{\partial LARSM_{\delta,\gamma,\alpha,s}(p\|q)}{\partial q_j} = \frac{1}{\delta-1}\left\{\left[\sum_i\left[\alpha p_i^\delta+(1-\alpha)\,q_i^\delta\right]^{\frac{1}{\delta}}\right]^{-1}\left[\alpha p_j^\delta+(1-\alpha)\,q_j^\delta\right]^{\frac{1-\delta}{\delta}}q_j^{\delta-1}\right.$$
$$\left.-\left[\sum_i\left[\alpha p_i^\gamma+(1-\alpha)\,q_i^\gamma\right]^{\frac{1}{\gamma}}\right]^{-1}\left[\alpha p_j^\gamma+(1-\alpha)\,q_j^\gamma\right]^{\frac{1-\gamma}{\gamma}}q_j^{\gamma-1}\right\} \tag{4.61}$$

This last expression can be obtained from (4.60) by simply doing $s=1$.
The components of these gradients will be zero for $p_j = q_j\ \forall j$.

4.6 Jensen divergences based on Entropies.

We use as basic convex functions, the opposite of the Shannon, Havrda-Charvat and Renyi Entropies.

4.6.1 Shannon's entropy.

On the basis of the simple convex function which is the opposite of Shannon's entropy, that is:

$$f(x) = x\log x \tag{4.62}$$

we treat the case of Jensen's divergence weighted by $0<\beta<1$, the classic case $\beta = 1/2$ is immediately inferred.
We have:

$$J_S(p\|q) = \beta\sum_i p_i\log p_i + (1-\beta)\sum_i q_i\log q_i$$
$$-\sum_i\left[\beta p_i+(1-\beta)\,q_i\right]\log\left[\beta p_i+(1-\beta)\,q_i\right] \tag{4.63}$$

The gradient with respect to "q" is expressed as:

$$\frac{\partial J_S\left(p\|q\right)}{\partial q_j} = \left(1-\beta\right)\left\{\log q_j - \log\left[\beta p_j + \left(1-\beta\right) q_j\right]\right\} \tag{4.64}$$

Of course these components will be zero for $p_j = q_j \ \forall j$.

Scale invariance with respect to "q"

For this divergence given by the relation (4.63), the invariance factor can not be calculated explicitly, so we use as invariance factor the expression $K^*(p,q) = \frac{\sum_j p_j}{\sum_j q_j}$.
In these conditions, the invariant divergence is written after simplifications:

$$\begin{aligned} JSI\left(p\|q\right) =& \beta \sum_i \bar{p}_i \log \bar{p}_i + \left(1-\beta\right) \sum_i \bar{q}_i \log \bar{q}_i \\ & - \sum_i \left[\beta \bar{p}_i + \left(1-\beta\right) \bar{q}_i\right] \log\left[\beta \bar{p}_i + \left(1-\beta\right) \bar{q}_i\right] \end{aligned} \tag{4.65}$$

The gradient with respect to "q" is expressed as:

$$\frac{\partial JSI\left(p\|q\right)}{\partial q_l} = \frac{1-\beta}{\sum_j q_j}\left[\log \frac{\bar{q}_l}{\beta \bar{p}_l + \left(1-\beta\right)\bar{q}_l} - \sum_i \bar{q}_i \log \frac{\bar{q}_i}{\beta \bar{p}_i + \left(1-\beta\right)\bar{q}_i}\right] \tag{4.66}$$

4.6.2 Havrda-Charvat entropy.

Here, the convex base function is the opposite of the Havrda-Charvat entropy; we treat the case of the Jensen divergence weighted by $0 < \beta < 1$, the classic case $\beta = 1/2$ is deduced immediately.
We have:

$$J_{HC}\left(p\|q\right) = \frac{1}{\alpha\left(\alpha-1\right)}\left\{\beta \sum_i p_i^\alpha + \left(1-\beta\right)\sum_i q_i^\alpha - \sum_i \left[\beta p_i + \left(1-\beta\right) q_i\right]^\alpha\right\} \tag{4.67}$$

The gradient with respect to "q" is written:

$$\frac{\partial J_{HC}\left(p\|q\right)}{\partial q_j} = \frac{1-\beta}{\alpha-1}\left\{q_j^{\alpha-1} - \left[\beta p_j + \left(1-\beta\right) q_j\right]^{\alpha-1}\right\} \tag{4.68}$$

Scale invariance with respect to "q".

For this divergence given by the relation (4.67), the invariance factor can not be calculated explicitly, so we use as invariance factor the expression $K^*(p,q) = \frac{\sum_j p_j}{\sum_j q_j}$.
Under these conditions, the invariant divergence is written after simplifications:

$$JHCI\left(p\|q\right) = \frac{1}{\alpha\left(\alpha-1\right)}\left\{\beta\sum_i \bar{p}_i^{\alpha} + \left(1-\beta\right)\sum_i \bar{q}_i^{\alpha} - \sum_i \left[\beta\bar{p}_i + \left(1-\beta\right)\bar{q}_i\right]^{\alpha}\right\} \tag{4.69}$$

The gradient with respect to "q" is expressed as:

$$\begin{aligned}\frac{\partial JHCI\left(p\|q\right)}{\partial q_l} =& \frac{1-\beta}{\left(\alpha-1\right)\sum_j q_j}\left\{\bar{q}_l^{\alpha-1} - \sum_i \bar{q}_i^{\alpha} - \left[\beta\bar{p}_l + \left(1-\beta\right)\bar{q}_l\right]^{\alpha-1}\right. \\ &\left. + \sum_i \bar{q}_i\left[\beta\bar{p}_i + \left(1-\beta\right)\bar{q}_i\right]^{\alpha-1}\right\}\end{aligned} \tag{4.70}$$

If, in the expression of the divergence (4.67), we replace the arithmetic mean by a generalized geometric mean, which will not change the sign of the expression, we do what O. Michel [70] has proposed for Renyi's divergence and on which we will come back later; the divergence thus constructed will be written:

$$J2HC\left(p\|q\right) = \frac{1}{\alpha\left(\alpha-1\right)}\left\{\beta\sum_i p_i^{\alpha} + \left(1-\beta\right)\sum_i q_i^{\alpha} - \sum_i \left[p_i^{\beta}q_i^{1-\beta}\right]^{\alpha}\right\} \tag{4.71}$$

The gradient with respect to "q" will be:

$$\frac{\partial J2HC\left(p\|q\right)}{\partial q_j} = \frac{1-\beta}{\alpha-1}\left\{q_j^{\alpha-1} - \left[p_j^{\beta}q_j^{1-\beta}\right]^{\alpha-1}\left(\frac{p_j}{q_j}\right)^{\beta}\right\} \tag{4.72}$$

Scale invariance with respect to "q".

For this extension proposed by O.Michel [70] given by (4.71), using the invariance factor $K^*\left(p,q\right)$ previously defined, we obtain after simplifications the following invariant divergence:

$$J2HCI\left(p\|q\right) = \frac{1}{\alpha\left(\alpha-1\right)}\left\{\beta\sum_i \bar{p}_i^{\alpha} + \left(1-\beta\right)\sum_i \bar{q}_i^{\alpha} - \sum_i \left[\bar{p}_i^{\beta}\bar{q}_i^{1-\beta}\right]^{\alpha}\right\} \tag{4.73}$$

The gradient with respect to "q" will be:

$$\frac{\partial J2_{HC}I\left(p\|q\right)}{\partial q_l}=\frac{1-\beta}{\left(\alpha-1\right)\sum_j q_j}\left\{\bar{q}_l^{\alpha-1}-\sum_i \bar{q}_i^{\alpha}-\left[\bar{p}_l^{\beta}\bar{q}_l^{1-\beta}\right]^{\alpha}\bar{p}_l^{\beta}\bar{q}_l^{-\beta}+\sum_i\left[\bar{p}_i^{\beta}\bar{q}_i^{1-\beta}\right]^{\alpha+1}\right\} \tag{4.74}$$

This being the case, we can also continue to follow the reasoning of O. Michel [70] in his work on Renyi's divergence, and introduce a Jensen-Havrda Charvat divergence between a generalized arithmetic mean of the 2 fields and one or the other of the fields, then, one can always invert the order of the arguments insofar as the divergences thus formed are not necessarily symmetrical.

4.6.3 Renyi's entropy.

Here, the convex base function is the opposite of Renyi's entropy, these leads to Arndt's ${}^1R_{\alpha}^1$ [6]. We treat the case of Jensen's divergence weighted by $0<\beta<1$, the classic case $\beta=1/2$ is inferred immediately.
The corresponding divergence is written as:

$$\begin{aligned}J_R^{\beta}\left(p\|q\right)=&\frac{1}{\alpha\left(\alpha-1\right)}\left\{\beta\log\sum_i p_i^{\alpha}+\left(1-\beta\right)\log\sum_i q_i^{\alpha}\right.\\&\left.-\log\sum_i\left[\beta p_i+\left(1-\beta\right)q_i\right]^{\alpha}\right\}\end{aligned} \tag{4.75}$$

The gradient with respect to "q" will be:

$$\frac{\partial J_R^{\beta}\left(p\|q\right)}{\partial q_j}=\frac{1-\beta}{\alpha-1}\left\{\frac{q_j^{\alpha-1}}{\sum_i q_i^{\alpha}}-\frac{\left[\beta p_j+\left(1-\beta\right)q_j\right]^{\alpha-1}}{\sum_i\left[\beta p_i+\left(1-\beta\right)q_i\right]^{\alpha}}\right\} \tag{4.76}$$

Scale invariance with respect to "q".

For such a divergence given by the relation (4.75), we use the $K^*(p,q)$ invariance factor already mentioned, which leads to the invariant divergence:

$$\begin{aligned}J_R^{\beta}I\left(p\|q\right)=&\frac{1}{\alpha\left(\alpha-1\right)}\left\{\beta\log\sum_i \bar{p}_i^{\alpha}+\left(1-\beta\right)\log\sum_i \bar{q}_i^{\alpha}\right.\\&\left.-\log\sum_i\left[\beta\bar{p}_i+\left(1-\beta\right)\bar{q}_i\right]^{\alpha}\right\}\end{aligned} \tag{4.77}$$

The gradient with respect to "q" is written as:

$$\frac{\partial J_R^{\beta} I\left(p\|q\right)}{\partial q_l} = \frac{1-\beta}{\left(\alpha-1\right)\sum_j q_j}\left\{\frac{\bar{q}_l^{\alpha-1}}{\sum_j q_j^{\alpha}} - 1 - \frac{\left[\beta\bar{p}_l + \left(1-\beta\right)\bar{q}_l\right]^{\alpha-1}}{\sum_i \left[\beta\bar{p}_i + \left(1-\beta\right)\bar{q}_i\right]^{\alpha}} + \frac{\sum_i \bar{q}_i \left[\beta\bar{p}_i + \left(1-\beta\right)\bar{q}_i\right]^{\alpha-1}}{\sum_i \left[\beta\bar{p}_i + \left(1-\beta\right)\bar{q}_i\right]^{\alpha}}\right\} \tag{4.78}$$

If in the expression (4.75) the generalized arithmetic mean is replaced by a generalized geometric mean, as proposed by O. Michel in [70], the divergence is written:

$$J2_R^{\beta}\left(p\|q\right) = \frac{1}{\alpha\left(\alpha-1\right)}\left\{\beta\log\sum_i p_i^{\alpha} + \left(1-\beta\right)\log\sum_i q_i^{\alpha} - \log\sum_i \left(p_i^{\beta} q_i^{1-\beta}\right)^{\alpha}\right\} \tag{4.79}$$

Whose gradient is:

$$\frac{\partial J2_R^{\beta}\left(p\|q\right)}{\partial q_j} = \frac{1-\beta}{\alpha-1}\left\{\frac{q_j^{\alpha-1}}{\sum_i q_i^{\alpha}} - \frac{\left[p_j^{\beta} q_j^{1-\beta}\right]^{\alpha-1}}{\sum_i \left[p_i^{\beta} q_i^{1-\beta}\right]^{\alpha}} \frac{p_j^{\beta}}{q_j^{\beta}}\right\} \tag{4.80}$$

Scale invariance with respect to "q".

From this modification proposed by O.Michel [70] given by (4.79); the corresponding scale invariant divergence is deduced by introducing the invariance factor $K^*(p,q)$:

$$J2_R^{\beta} I\left(p\|q\right) = \frac{1}{\alpha\left(\alpha-1\right)}\left\{\beta\log\sum_i \bar{p}_i^{\alpha} + \left(1-\beta\right)\log\sum_i \bar{q}_i^{\alpha} - \log\sum_i \left(\bar{p}_i^{\beta} \bar{q}_i^{1-\beta}\right)^{\alpha}\right\} \tag{4.81}$$

The gradient with respect to "q" is expressed as:

$$\frac{\partial J2_R^{\beta} I\left(p\|q\right)}{\partial q_l} = \frac{1-\beta}{\left(\alpha-1\right)\sum_j q_j}\left[\frac{\bar{q}_l^{\alpha-1}}{\sum_j \bar{q}_j^{\alpha}} - \frac{\left(\bar{p}_l^{\beta} \bar{q}_l^{1-\beta}\right)^{\alpha-1} \bar{p}_l^{\beta} \bar{q}_l^{-\beta}}{\sum_j \left(\bar{p}_j^{\beta} \bar{q}_j^{1-\beta}\right)^{\alpha}}\right] \tag{4.82}$$

In the same article [70], another variation of this type of divergence has been suggested. To fully understand how to obtain it, we'll proceed in successive steps.
First we write the dual divergence of $J2_R^{\beta}(p\|q)$ which gives:

$$J2_R^{\beta}(q\|p) = \frac{1}{\alpha(\alpha-1)}\left\{\beta\log\sum_i q_i^{\alpha} + (1-\beta)\log\sum_i p_i^{\alpha} - \log\sum_i\left(q_i^{\beta}p_i^{1-\beta}\right)^{\alpha}\right\} \tag{4.83}$$

Then we construct a divergence from the previous general form, but by replacing q by a linear combination of "p" and "q" with a weighting factor "γ", i.e. $\gamma p + (1-\gamma)q$, which gives:

$$J3_R^{\beta,\gamma}(p\|q) = \frac{1}{\alpha(\alpha-1)}\left\{\beta\log\sum_i\left[\gamma p_i + (1-\gamma)q_i\right]^{\alpha} + (1-\beta)\log\sum_i p_i^{\alpha} - \log\sum_i\left(\left[\gamma p_i + (1-\gamma)q_i\right]^{\beta}p_i^{1-\beta}\right)^{\alpha}\right\} \tag{4.84}$$

Of course, this little exercise can go on for a very long time and may not be of much interest...
We can still calculate the gradient relative to "q":

$$\frac{\partial J3_R^{\beta,\gamma}(p\|q)}{\partial q_j} = \beta\frac{1-\gamma}{\alpha-1}\left\{\frac{\left[\gamma p_j + (1-\gamma)q_j\right]^{\alpha-1}}{\left[\sum_i \gamma p_i + (1-\gamma)q_i\right]^{\alpha}} - \frac{\left(\left[\gamma p_j + (1-\gamma)q_j\right]^{\beta}p_j^{1-\beta}\right)^{\alpha}}{\left(\sum_i\left[\gamma p_i + (1-\gamma)q_i\right]^{\beta}p_i^{1-\beta}\right)^{\alpha}}\frac{1}{\gamma p_j + (1-\gamma)q_j}\right\} \tag{4.85}$$

chapter 5 - Alpha, Beta and Gamma divergences

5.1 Introduction.

In this chapter, we present the information corresponding to the divergences constructed on the basis of functions of the type:

$$f(x) = \frac{x^{\lambda}}{\lambda - 1} \tag{5.1}$$

The variable "x" is assumed to be positive and we use the parameter "λ" which hasn't been used yet and can become α, β or γ as needed.
This function is convex, for $\lambda > 0$; to make it convex regardless of "λ", we have to write:

$$f(x) = \frac{x^{\lambda}}{\lambda(\lambda - 1)} \tag{5.2}$$

To get moreover $f(1) = 0$, we have to turn it into:

$$f_1(x) = \frac{1}{\lambda(\lambda - 1)}\left(x^{\lambda} - 1\right) \tag{5.3}$$

This function is shown on the figure (5.1),
or else:

$$f_2(x) = \frac{1}{\lambda(\lambda - 1)}\left(x^{\lambda} - x\right) \tag{5.4}$$

which is represented on the figure: (5.2).

The functions $f_1(x)$ and $f_2(x)$ are *"simple convex functions"*.
If we want to have, in addition, a null derivative in $x = 1$, that is to say if we want to obtain a *"standard convex function"*, we obtain in both cases the same function $f_c(x)$ which is represented on the figure (5.3) and which is written:

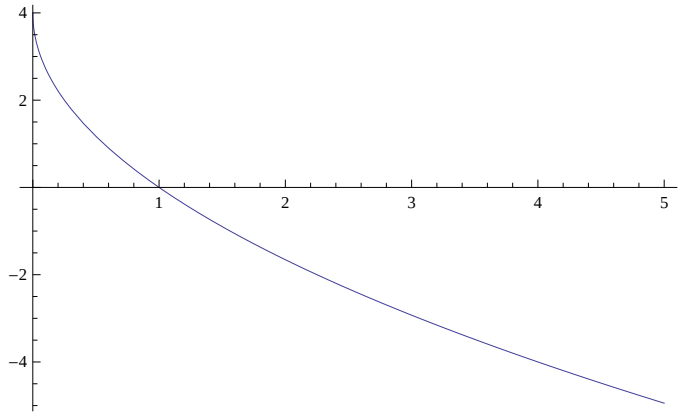

Figure 5.1: Function $f_1(x) = \frac{1}{\lambda(\lambda-1)}\left(x^\lambda - 1\right), \lambda = 0.5$

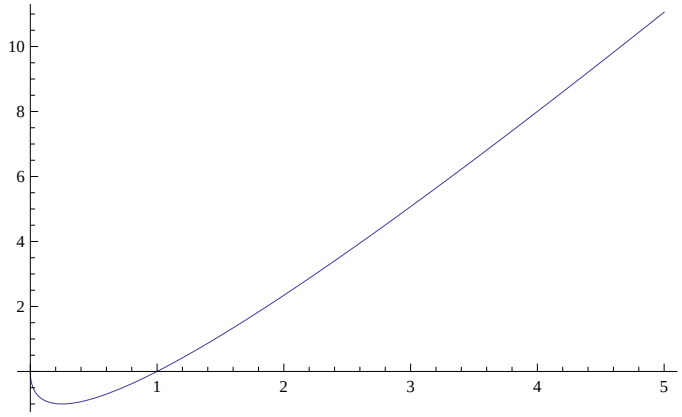

Figure 5.2: Function $f_2(x) = \frac{1}{\lambda(\lambda-1)}\left(x^\lambda - x\right), \lambda = 0.5$

$$f_c(x) = \frac{1}{\lambda(\lambda-1)}\left[x^\lambda - \lambda x - (1-\lambda)\right] \tag{5.5}$$

In all these figures, $\lambda = 0.5$.

An important point to keep in mind is that these functions are convex whatever "λ" is due to the "$1/\lambda$" factor that's been introduced, which may not always be found in the divergences proposed in the literature, but it will allow the $\lambda \to 0$ limit to be computed when necessary.
The question now is, what divergences can be built using such functions?

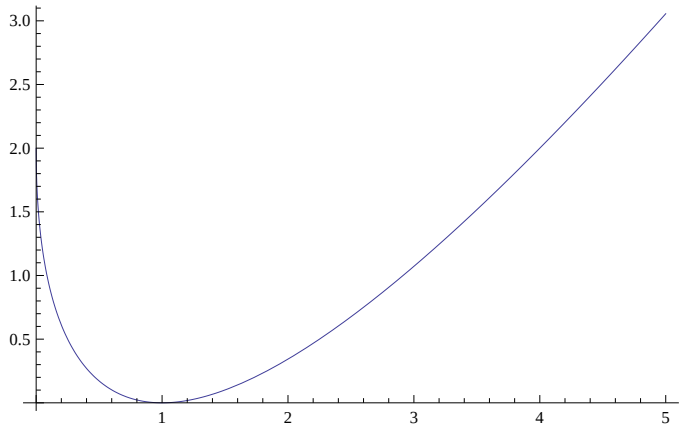

Figure 5.3: Function $f_c(x) = \frac{1}{\lambda(\lambda-1)}\left[x^\lambda - \lambda x - (1-\lambda)\right], \lambda = 0.5$

We will consider two constructive modes that have already been dealt with in the literature: Csiszär's divergences and Bregman's divergences.
In addition, we will also consider Jensen's divergences.

5.2 Csiszär's divergences - Alpha divergences.

They have been largely developed in the works of Amari [2] and Cichocki [23]; they are based on the functions $f_1(x)$, $f_2(x)$ and $f_c(x)$.

5.2.1 On $f_1(x)$.

In this case we obtain the divergence:

$$A_1(p\|q) = \frac{1}{\lambda(\lambda-1)} \sum_i \left(p_i^\lambda q_i^{1-\lambda} - q_i\right) \tag{5.6}$$

If we calculate the gradient with the intention to minimize with respect to "q", we get:

$$\frac{\partial A_1(p\|q)}{\partial q_j} = \frac{1}{\lambda(\lambda-1)}\left[(1-\lambda)\left(\frac{p_j}{q_j}\right)^\lambda - 1\right] \tag{5.7}$$

This seems suitable for building a descent algorithm, but the problem will be as always, that the gradient doesn't cancel out for $p_i = q_i \ \forall i$.
Moreover, we can notice that if we operate on probability densities ($\sum_i p_i =$

$\sum_i q_i = 1$), the divergence simplifies, but in this case the corresponding gradient will never be zero.
The dual divergence is built on the mirror function represented in figure (5.4)

$$\breve{f}_1(x) = \frac{1}{\lambda(\lambda-1)}\left(x^{1-\lambda} - x\right) \tag{5.8}$$

It is expressed as follows:

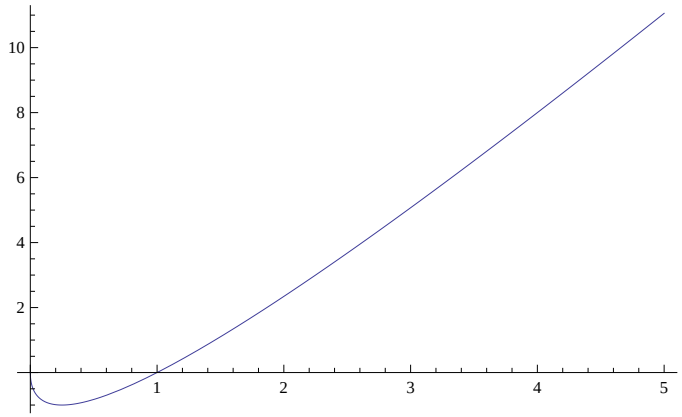

Figure 5.4: Function $\breve{f}_1(x) = \frac{1}{\lambda(\lambda-1)}\left(x^{1-\lambda} - x\right), \lambda = 0.5$

$$A_1(q\|p) = \frac{1}{\lambda(\lambda-1)}\sum_i \left(q_i^{\lambda} p_i^{1-\lambda} - p_i\right) \tag{5.9}$$

On this divergence, we can see that the gradient with respect to "q" never cancels out, hence a problem if we try to minimize.

5.2.2 On $f_2(x)$.

In this case we obtain the divergence:

$$A_2(p\|q) = \frac{1}{\lambda(\lambda-1)}\sum_i \left(p_i^{\lambda} q_i^{1-\lambda} - p_i\right) \tag{5.10}$$

The gradient will be:

$$\frac{\partial A_2(p\|q)}{\partial q_j} = -\frac{1}{\lambda}\left(\frac{p_j}{q_j}\right)^{\lambda} \tag{5.11}$$

We can see that the gradient with respect to "q" never cancels out.
If we operate on probability densities, the divergence becomes simpler, but the problem is again, the gradient will not cancel out.
The dual divergence will be built on the mirror function represented in figure (5.5):

$$\breve{f}_2(x) = \frac{1}{\lambda(\lambda-1)}\left(x^{1-\lambda}-1\right) \tag{5.12}$$

It is expressed as follows:

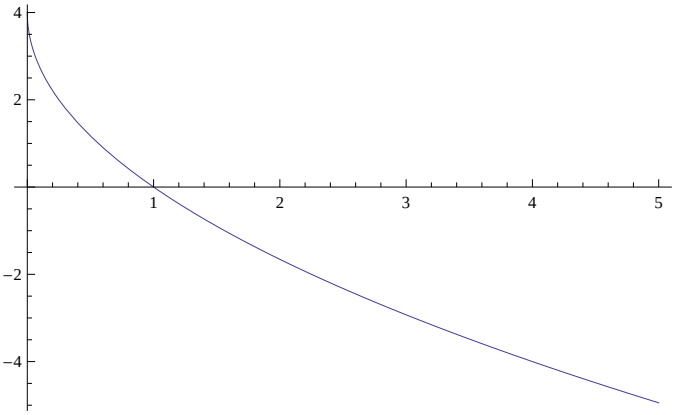

Figure 5.5: Function $\breve{f}_2(x) = \frac{1}{\lambda(\lambda-1)}\left(x^{1-\lambda}-1\right), \lambda = 0.5$

$$A_2(q\|p) = \frac{1}{\lambda(\lambda-1)}\sum_i \left(q_i^{\lambda} p_i^{1-\lambda} - q_i\right) \tag{5.13}$$

This divergence is made simpler if we operate on probability densities.
The gradient with respect to "q" is:

$$\frac{\partial A_2(q\|p)}{\partial q_j} = \frac{1}{\lambda(\lambda-1)}\left[\lambda\left(\frac{p_j}{q_j}\right)^{1-\lambda} - 1\right] \tag{5.14}$$

Of course, the problem is still that the gradient with respect to "q" will not be zero for $p_i = q_i$.

5.2.3 On $f_c(x)$.

In this case we obtain the divergence:

$$A(p\|q) = \frac{1}{\lambda(\lambda-1)}\sum_i \left(p_i^{\lambda} q_i^{1-\lambda} - \lambda p_i - (1-\lambda) q_i\right) \tag{5.15}$$

This divergence [42] can also be thought of as a arithmetic-geometric divergence [91], it is also the "Alpha divergence" of Amari [25]($\lambda \leftrightarrow \alpha$).
Its gradient with respect to "q" is written as follows:

$$\frac{\partial A\left(p\|q\right)}{\partial q_j} = \frac{1}{\lambda}\left[1 - \left(\frac{p_j}{q_j}\right)^{\lambda}\right] \tag{5.16}$$

It will be zero if $p_i = q_i \ \forall i$.
The dual divergence will be constructed on the mirror function shown in figure (5.6):

$$\breve{f}_c\left(x\right) = \frac{1}{\lambda\left(\lambda - 1\right)}\left[x^{1-\lambda} - \lambda - \left(1 - \lambda\right)x\right] \tag{5.17}$$

It will be written:

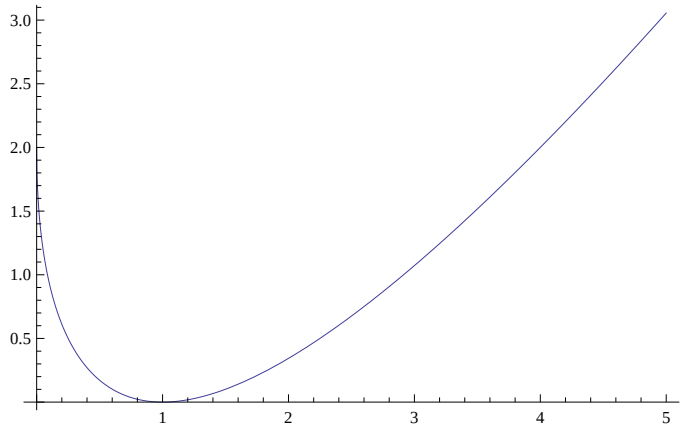

Figure 5.6: Function $\breve{f}_c\left(x\right) = \frac{1}{\lambda(\lambda-1)}\left[x^{1-\lambda} - \lambda - \left(1 - \lambda\right)x\right], \lambda = 0.5$

$$A\left(q\|p\right) = \frac{1}{\lambda\left(\lambda - 1\right)}\sum_i \left(q_i^{\lambda} p_i^{1-\lambda} - \lambda q_i - \left(1 - \lambda\right)p_i\right) \tag{5.18}$$

Its gradient with respect to "q" will be:

$$\frac{\partial A\left(q\|p\right)}{\partial q_j} = \frac{1}{\lambda - 1}\left[\left(\frac{p_j}{q_j}\right)^{1-\lambda} - 1\right] \tag{5.19}$$

It will be zero if $p_i = q_i \ \forall i$.

5.2.4 Some classical particular cases.

If $\lambda \to 1$ we have $A(p\|q) = KL(p\|q)$.
If $\lambda \to 0$ we have $A(p\|q) = KL(q\|p)$.
If $\lambda = 1/2$ we have $A(p\|q) = H(p\|q)$, that is $\sum_i \left(\sqrt{p_i} - \sqrt{q_i}\right)^2$ (Hellinger's divergence [43]).
If $\lambda = 2$ we have $A(p\|q) = \chi^2_N(p\|q)$, that is $\frac{1}{2}\sum_i \frac{(p_i - q_i)^2}{q_i}$ (Neyman's Chi2).
If $\lambda = -1$ we have $A(p\|q) = \chi^2_P(p\|q)$, that is $\frac{1}{2}\sum_i \frac{(p_i - q_i)^2}{p_i}$ (Pearson's Chi2).

5.3 Bregman's divergences - Beta divergences.

Since the Bregman divergences are convexity measures, it is obvious that the functions $f_1(x)$, $f_2(x)$ or $f_c(x)$ which differ from each other only by a linear function (i.e. which have the same second derivative) lead to the same divergence which can be written as follows:

$$B(p\|q) = \frac{1}{\lambda(\lambda-1)} \sum_i \left[p_i^\lambda - \lambda p_i q_i^{\lambda-1} - (1-\lambda) q_i^\lambda\right] \tag{5.20}$$

With $(\lambda \leftrightarrow \beta)$ we obtain the "Beta" divergence of Mihoko and Eguchi [71] or Uchida and Shioya [96]; we also find the divergence of BHHJ [11] by deleting in (5.20) the multiplicative factor "$1/\lambda$".
The gradient with respect to "q" will be:

$$\frac{\partial B(p\|q)}{\partial q_j} = q_j^{\lambda-2}(q_j - p_j) \tag{5.21}$$

This gradient will be zero if $p_i = q_i \ \forall i$.
From the relationship (2.10), the dual divergence is written:

$$B(q\|p) = \frac{1}{\lambda(\lambda-1)} \sum_i \left[q_i^\lambda - \lambda q_i p_i^{\lambda-1} - (1-\lambda) p_i^\lambda\right] \tag{5.22}$$

Its gradient relative to "q" will be:

$$\frac{\partial B(q\|p)}{\partial q_j} = \frac{1}{\lambda-1}\left(q_j^{\lambda-1} - p_j^{\lambda-1}\right) \tag{5.23}$$

Obviously, it will be zero if $p_i = q_i \ \forall i$.

5.3.1 Some classical particular cases.

If $\lambda \to 1$ we have $B(p\|q) = KL(p\|q)$.
This is the common point with "alpha divergences".
If $\lambda \to 0$ we have $B(p\|q) = IS(p\|q)$, it is the Itakura-Saito divergence [49].
If $\lambda = 2$ we have $B(p\|q) = \sum_i (p_i - q_i)^2$, it is the Mean square deviation (Euclidean distance).

5.4 Jensen's divergences.

Since Jensen's divergences are measures of convexity, it is obvious that the functions $f_1(x)$, $f_2(x)$ or $f_c(x)$ which differ from each other only by a linear function (i.e. which have the same second derivative) lead to the same divergence that is written:

$$J(p\|q) = \frac{1}{\lambda(\lambda-1)} \left\{ \sum_i \left[\alpha p_i^\lambda + (1-\alpha) q_i^\lambda \right] - \sum_i \left[\alpha p_i + (1-\alpha) q_i \right]^\lambda \right\} \tag{5.24}$$

The corresponding gradient with respect to "q" is written:

$$\frac{\partial J(p\|q)}{\partial q_j} = \frac{1-\alpha}{\lambda-1} \left\{ q_j^{\lambda-1} - \left[\alpha p_j + (1-\alpha) q_j \right]^{\lambda-1} \right\} \tag{5.25}$$

This gradient will be zero if $p_i = q_i \ \forall i$.

5.5 Invariance by change of scale.

We are applying the method defined in [34], in order to make invariant by change of scale on "q" the divergences of type "A", "B" and "J" of the preceding sections.

5.5.1 "*A*" divergences.

The starting divergence is given by (5.15):

$$A(p\|q) = \frac{1}{\lambda(\lambda-1)} \sum_i \left(p_i^\lambda q_i^{1-\lambda} - \lambda p_i - (1-\lambda) q_i \right) \tag{5.26}$$

The expression of the invariance factor K is obtained explicitly, so it is the nominal value that is written:

$$K_0 = \left(\frac{\sum_i p_i^{\lambda} q_i^{1-\lambda}}{\sum_i q_i} \right)^{\frac{1}{\lambda}} \tag{5.27}$$

Then the invariant divergence by change of scale on "q" is written:

$$AI\left(p\|q\right) = \frac{1}{\lambda - 1} \left[\left(\frac{\sum_i p_i^{\lambda} q_i^{1-\lambda}}{\sum_i q_i} \right)^{\frac{1}{\lambda}} \sum_i q_i - \sum_i p_i \right] \tag{5.28}$$

Which can also be written:

$$AI\left(p\|q\right) = \frac{1}{\lambda - 1} \left[\sum_i K_0 q_i - \sum_i p_i \right] \tag{5.29}$$

Example

If we consider the case $\lambda \to 1$, that is $KL\left(p\|q\right)$, we obtain:

$$K_0 = \frac{\sum_i p_i}{\sum_i q_i} \tag{5.30}$$

$$KLI\left(p\|q\right) = \sum_j p_j \left[\sum_i \bar{p}_i \log \frac{\bar{p}_i}{\bar{q}_i} + \bar{q}_i - \bar{p}_i \right] = \sum_j p_j \left[\sum_i \bar{p}_i \log \frac{\bar{p}_i}{\bar{q}_i} \right] \tag{5.31}$$

With: $\bar{p}_i = \frac{p_i}{\sum_j p_j}$ and $\bar{q}_i = \frac{q_i}{\sum_j q_j}$.
That's what we would get if we do directly $\lambda \to 1$ in the expression (5.28). Note that the passage to the limit $\lambda \to 1$ implies a specific calculation but effectively leads to the Kullback-Leibler divergence invariant with respect to "q".

The gradient of divergence (5.28) with respect to "q" is written:

$$\frac{\partial AI\left(p\|q\right)}{\partial q_j} = \frac{1}{\lambda} \underbrace{\left(\sum_i p_i^{\lambda} q_i^{1-\lambda} \right)^{\frac{1}{\lambda}} \left(\sum_i q_i \right)^{1-\frac{1}{\lambda}}}_{\mathbf{T}} \left[\frac{1}{\sum_i q_i} - \frac{p_j^{\lambda} q_j^{-\lambda}}{\sum_i p_i^{\lambda} q_i^{1-\lambda}} \right] \tag{5.32}$$

It is also possible to write this expression:

$$\frac{\partial AI\left(p\|q\right)}{\partial q_j} = \frac{1}{\lambda}\left[K_0 - K_0^{1-\lambda} p_j^{\lambda} q_j^{-\lambda}\right] \tag{5.33}$$

Remark: it can be observed that

$$\sum_j q_j \frac{\partial AI\left(p\|q\right)}{\partial q_j} = 0 \tag{5.34}$$

This relationship will be verified for all invariant divergences presented in this book.

By disregarding the constant multiplicative factor $\frac{1}{\lambda-1}$, the expression (5.28) is the difference of 2 terms; one can thus apply to each term of this difference the same increasing function, the "Generalized Logarithm" for example (see appendix 1), then the extreme form, the "Logarithm" which will lead to:

$$LAI\left(p\|q\right) = \frac{1}{\lambda}\log\sum_i q_i - \frac{1}{\lambda-1}\log\sum_i p_i + \frac{1}{\lambda\left(\lambda-1\right)}\log\sum_i p_i^{\lambda} q_i^{1-\lambda} \tag{5.35}$$

It can be seen, which is true in all cases, that the latter divergence is not only invariant by changing scale on "q" , but also by changing scale on "p". This is obviously related to the use of the Logarithm, (the Generalized Logarithm is not enough).
For Type B divergences, the analog of (5.35) will be the "Gamma divergence".
The gradient with respect to "q" is written:

$$\frac{\partial LAI\left(p\|q\right)}{\partial q_j} = \frac{1}{\lambda}\left[\frac{1}{\sum_i q_i} - \frac{p_j^{\lambda} q_j^{-\lambda}}{\sum_i p_i^{\lambda} q_i^{1-\lambda}}\right] \tag{5.36}$$

It is the expression(5.32) without the multiplicative factor "**T**", and we have as previously stated:

$$\sum_j q_j \frac{\partial LAI\left(p\|q\right)}{\partial q_j} = 0 \tag{5.37}$$

We can make the same work on the dual divergence (5.18):

$$A\left(q\|p\right) = \frac{1}{\lambda\left(\lambda-1\right)}\sum_i \left(q_i^{\lambda} p_i^{1-\lambda} - \lambda q_i - \left(1-\lambda\right) p_i\right) \tag{5.38}$$

By using the classical procedure for calculating the nominal invariance factor, the following results are obtained:

$$K_0 = \left(\frac{\sum_i p_i^{1-\lambda} q_i^{\lambda}}{\sum_i q_i} \right)^{\frac{1}{1-\lambda}} \tag{5.39}$$

The divergence made invariant will be written:

$$AI\left(q \| p\right) = \frac{1}{\lambda} \left[\sum_i p_i - \left(\frac{\sum_i p_i^{1-\lambda} q_i^{\lambda}}{\sum_i q_i} \right)^{\frac{1}{1-\lambda}} \sum_i q_i \right] \tag{5.40}$$

Or also, more simply:

$$AI\left(q \| p\right) = \frac{1}{\lambda} \left[\sum_i p_i - \sum_i K_0 q_i \right] \tag{5.41}$$

One can then apply an increasing function on each of the terms the divergence (5.40), for example the Generalized Logarithm function, and go further to the Logarithm to obtain:

$$\begin{aligned} LAI\left(q \| p\right) = & \frac{1}{\lambda} \left[\log \sum_i p_i - \frac{1}{1-\lambda} \log \sum_i p_i^{1-\lambda} q_i^{\lambda} \right. \\ & \left. + \frac{\lambda}{1-\lambda} \log \sum_i q_i \right] \end{aligned} \tag{5.42}$$

Note that this divergence is, of course, invariant with respect to "q", but it is also invariant with respect to "p".
A symmetrical divergence invariant with respect to "p" and "q" can be obtained by summing the (half) logarithmic invariant divergences $LAI\left(p \| q\right)$ and $LAI\left(q \| p\right)$.
Such a divergence is written:

$$LAI\left(p, q\right) = \frac{1}{\lambda(1-\lambda)} \left[\log \frac{\sum_i p_i \sum_i q_i}{\sum_i p_i^{\lambda} q_i^{1-\lambda} \sum_i q_i^{\lambda} p_i^{1-\lambda}} \right] \tag{5.43}$$

For type B divergences, the analogue of (5.43) will be the "symmetrical Gamma divergence ".

5.5.2 "*B*" divergences.

"Beta" divergence.

The initial divergence was (5.20):

$$B\left(p\|q\right)=\frac{1}{\lambda\left(\lambda-1\right)}\sum_{i}\left[p_{i}^{\lambda}-\lambda p_{i}q_{i}^{\lambda-1}-\left(1-\lambda\right)q_{i}^{\lambda}\right] \tag{5.44}$$

The expression of the nominal invariance factor K_0 is calculated explicitly; it is written as:

$$K_{0}=\frac{\sum_{i}p_{i}q_{i}^{\lambda-1}}{\sum_{i}q_{i}^{\lambda}} \tag{5.45}$$

This leads to the divergence:

$$BI\left(p\|q\right)=\frac{1}{\lambda\left(\lambda-1\right)}\left[\sum_{i}p_{i}^{\lambda}-\left(\frac{\sum_{i}p_{i}q_{i}^{\lambda-1}}{\sum_{i}q_{i}^{\lambda}}\right)^{\lambda}\sum_{i}q_{i}^{\lambda}\right] \tag{5.46}$$

Or also:

$$BI\left(p\|q\right)=\frac{1}{\lambda\left(\lambda-1\right)}\left[\sum_{i}p_{i}^{\lambda}-\sum_{i}K_{0}^{\lambda}q_{i}^{\lambda}\right] \tag{5.47}$$

To recover the divergence of Basu et al. [11] it is then necessary, in (5.46), to remove the multiplicative factor $1/\lambda$ and make the parameter change $\lambda=1+\beta$.
The gradient of divergence (5.46) with respect to "q" is written:

$$\frac{\partial BI\left(p\|q\right)}{\partial q_{j}}=\underbrace{\left(\sum_{i}p_{i}q_{i}^{\lambda-1}\right)^{\lambda}\left(\sum_{i}q_{i}^{\lambda}\right)^{1-\lambda}}_{\mathbf{S}}\left[\frac{q_{j}^{\lambda-1}}{\sum_{i}q_{i}^{\lambda}}-\frac{p_{j}q_{j}^{\lambda-2}}{\sum_{i}p_{i}q_{i}^{\lambda-1}}\right] \tag{5.48}$$

Or, in a simplified form:

$$\frac{\partial BI\left(p\|q\right)}{\partial q_{j}}=K_{0}^{\lambda}\left[q_{j}^{\lambda-1}-\left(K_{0}\right)^{-1}p_{j}q_{j}^{\lambda-2}\right] \tag{5.49}$$

It can be noted that:

$$\sum_{j}q_{j}\frac{\partial BI\left(p\|q\right)}{\partial q_{j}}=0 \tag{5.50}$$

"Gamma" divergence.

At this point, following the usual procedure, we can apply the "Generalized Logarithm" on each of the terms in the brackets in (5.46), and then eventually we can move to the Logarithm which leads to the "Gamma divergence" of Fujisawa and Eguchi [36]. ($\lambda \leftrightarrow \gamma$).:

$$LBI\left(p\|q\right) = \frac{1}{\lambda\left(\lambda-1\right)}\left[\log\sum_i p_i^\lambda - \lambda\log\sum_i p_i q_i^{\lambda-1} + \left(\lambda-1\right)\log\sum_i q_i^\lambda\right] \tag{5.51}$$

This divergence is invariant by scale change not only on "q", but also on "p". The gradient with respect to "q" is written:

$$\frac{\partial LBI\left(p\|q\right)}{\partial q_j} = \frac{q_j^{\lambda-1}}{\sum_i q_i^\lambda} - \frac{p_j q_j^{\lambda-2}}{\sum_i p_i q_i^{\lambda-1}} \tag{5.52}$$

This expression is analogous to (5.48) except for the multiplicative factor "**S**" in (5.48).
We can see that as previously stated:

$$\sum_j q_j \frac{\partial LBI\left(p\|q\right)}{\partial q_j} = 0 \tag{5.53}$$

Dual "Beta" divergence.

We can then come back to the dual divergence (5.22):

$$B\left(q\|p\right) = \frac{1}{\lambda\left(\lambda-1\right)}\sum_i\left[q_i^\lambda - \lambda q_i p_i^{\lambda-1} - \left(1-\lambda\right)p_i^\lambda\right] \tag{5.54}$$

The expression of K_0 allowing to get the invariance is computed explicitly; we have:

$$K_0 = \left(\frac{\sum_i p_i^{\lambda-1} q_i}{\sum_i q_i^\lambda}\right)^{\frac{1}{\lambda-1}} \tag{5.55}$$

The corresponding invariant divergence will be written:

$$BI\left(q\|p\right) = \frac{1}{\lambda}\left[\sum_i p_i^\lambda - \left(\frac{\sum_i q_i p_i^{\lambda-1}}{\sum_i q_i^\lambda}\right)^{\frac{\lambda}{\lambda-1}}\sum_i q_i^\lambda\right] \tag{5.56}$$

Or, more simply:

$$BI\left(q\|p\right)=\frac{1}{\lambda}\left[\sum_{i}p_{i}^{\lambda}-\sum_{i}K_{0}^{\lambda}q_{i}^{\lambda}\right] \tag{5.57}$$

Dual "Gamma" divergence.

From the previous expression, we obtain the Logarithmic form, i.e. the dual Gamma divergence:

$$LBI\left(q\|p\right)=\frac{1}{\lambda}\left[\log\sum_{i}p_{i}^{\lambda}-\frac{1}{1-\lambda}\log\sum_{i}q_{i}^{\lambda}+\frac{\lambda}{1-\lambda}\log\sum_{i}q_{i}p_{i}^{\lambda-1}\right] \tag{5.58}$$

Its gradient with respect to "q" is written:

$$\frac{\partial LBI\left(q\|p\right)}{\partial q_{j}}=\frac{1}{\lambda-1}\left[\frac{q_{j}^{\lambda-1}}{\sum_{i}q_{i}^{\lambda}}-\frac{p_{j}^{\lambda-1}}{\sum_{i}q_{i}p_{i}^{\lambda-1}}\right] \tag{5.59}$$

Symmetrical "Gamma" divergence.

It is immediately obtained as the (half) sum of the Gamma divergence and the dual Gamma divergence; it is written as follows:

$$LBI\left(p,q\right)=\frac{1}{\lambda-1}\log\frac{\sum_{i}p_{i}^{\lambda}\sum_{i}q_{i}^{\lambda}}{\sum_{i}p_{i}q_{i}^{\lambda-1}\sum_{i}q_{i}p_{i}^{\lambda-1}} \tag{5.60}$$

Its gradient with respect to "q" is written:

$$\frac{\partial LBI\left(p,q\right)}{\partial q_{j}}=\frac{\lambda}{\lambda-1}\frac{q_{j}^{\lambda-1}}{\sum_{i}q_{i}^{\lambda}}-\frac{p_{j}q_{j}^{\lambda-2}}{\sum_{i}p_{i}q_{i}^{\lambda-1}}-\frac{1}{\lambda-1}\frac{p_{j}^{\lambda-1}}{\sum_{i}q_{i}p_{i}^{\lambda-1}} \tag{5.61}$$

5.5.3 "J" divergences.

We write:

$$\begin{aligned}J\left(p\|Kq\right)=&\frac{1}{\lambda\left(\lambda-1\right)}\left\{\sum_{i}\left[\alpha p_{i}^{\lambda}+\left(1-\alpha\right)K^{\lambda}q_{i}^{\lambda}\right]\right.\\&\left.-\sum_{i}\left[\alpha p_{i}+\left(1-\alpha\right)Kq_{i}\right]^{\lambda}\right\}\end{aligned} \tag{5.62}$$

We deduce that:

$$\frac{\partial J\left(p\|Kq\right)}{\partial K}=\frac{1-\alpha}{\lambda-1}\sum_{i}q_{i}\left\{K^{\lambda-1}q_{i}^{\lambda-1}-\left[\alpha p_{i}+\left(1-\alpha\right)Kq_{i}\right]^{\lambda-1}\right\} \tag{5.63}$$

The method proposed in **Chapter 3**, which consists in obtaining the invariance factor K by solving the equation:

$$\frac{\partial J\left(p\|Kq\right)}{\partial K}=0 \tag{5.64}$$

as proposed in (3.17), does not lead to an explicit solution.
On the other hand, if we choose to use the invariance factor:

$$K=\frac{\sum_{i}p_{i}}{\sum_{i}q_{i}} \tag{5.65}$$

We obtain an invariant divergence which is written as follows:

$$\begin{aligned}JI\left(p\|q\right)=&\frac{\left(\sum_{j}p_{j}\right)^{\lambda}}{\lambda\left(\lambda-1\right)}\left\{\sum_{i}\left[\alpha\overline{p}_{i}^{\lambda}+\left(1-\alpha\right)\overline{q}_{i}^{\lambda}\right]\right.\\&\left.-\sum_{i}\left[\alpha\overline{p}_{i}+\left(1-\alpha\right)\overline{q}_{i}\right]^{\lambda}\right\}\end{aligned} \tag{5.66}$$

with: $\bar{p}_{i}=\frac{p_{i}}{\sum_{j}p_{j}}$ and $\bar{q}_{i}=\frac{q_{i}}{\sum_{j}q_{j}}$.

Its gradient with respect to "q" is written as follows:

$$\begin{aligned}\frac{\partial JI\left(p\|q\right)}{\partial q_{j}}=&\frac{\left(1-\alpha\right)}{\left(\lambda-1\right)}\frac{\left(\sum_{l}p_{l}^{\lambda}\right)}{\sum_{l}q_{l}}\left\{\left[\sum_{i}\bar{q}_{i}\left[\alpha\bar{p}_{i}+\left(1-\alpha\right)\bar{q}_{i}\right]^{\lambda-1}-\bar{q}_{i}^{\lambda}\right]\right.\\&\left.-\left[\left[\alpha\bar{p}_{j}+\left(1-\alpha\right)\bar{q}_{j}\right]^{\lambda-1}-\bar{q}_{j}^{\lambda-1}\right]\right\}\end{aligned} \tag{5.67}$$

Of course, we still have the fundamental relationship:

$$\sum_{j}q_{j}\frac{\partial JI\left(p\|q\right)}{\partial q_{j}}=0 \tag{5.68}$$

5.5.4 Important note on invariance.

In the $AI\left(p\|q\right)$, $BI\left(p\|q\right)$ and $JI\left(p\|q\right)$ divergences, the invariance was established with respect to the variable "q"; moreover, it was noted that these

divergences are in the form of a difference of 2 positive terms.
Based on this observation, an increasing function can be applied to each of the terms of the difference without changing the sign of the divergence (positive in this case); moreover, since both terms are positive, the use of the Logarithm function is possible, the use of the Generalized Logarithm being only an intermediate step.
We can now make an additional remark if we are considering the invariance with respect to "p": we can observe that if "p" is multiplied by K in $AI(p\|q)$, the 2 terms of the divergence are multiplied by K, and only the application of a logarithmic function on each of the terms of the difference will allow the factor K to vanish.
Similarly, if in $BI(p\|q)$, we multiply "p" by K, each of the terms of the difference will be multiplied by K^{λ} and, again, only the application of the logarithmic function on each of the 2 terms of the difference will result in the invariance with respect to "p", removing the K factor.

5.5.5 A particular case of the invariance factor.

As reported in **Chapter 3**, for certain divergences, the method for calculating the invariance factor $K(p,q)$ by resolution of (3.35) does not lead to an explicit solution, however, an expression of the form:

$$K(p,q) = \frac{\sum_j p_j}{\sum_j q_j} \tag{5.69}$$

allows to obtain a scale change invariant form of any divergence.
As previously proposed for Jensen's divergences, we will introduce this factor into the "Alpha" and "Beta" divergences (5.15) (5.20) and compare the resulting expressions with the forms proposed in (5.28) and (5.46).
If we put (5.69) in (5.15), with the notations $\bar{p}_i = \frac{p_i}{\sum_j p_j}$ and $\bar{q}_i = \frac{q_i}{\sum_j q_j}$,, we obtain:

$$AI_{bis}(p\|q) = \frac{\sum_j p_j}{\lambda(\lambda-1)} \sum_i \left(\bar{p}_i^{\lambda} \bar{q}_i^{1-\lambda} - \lambda \bar{p}_i - (1-\lambda)\bar{q}_i \right) \tag{5.70}$$

Which can be simplified in the following way:

$$AI_{bis}(p\|q) = \frac{\sum_j p_j}{\lambda(\lambda-1)} \left[\left(\sum_i \bar{p}_i^{\lambda} \bar{q}_i^{1-\lambda} \right) - 1 \right] \tag{5.71}$$

Similarly, if we put (5.69) in (5.20), we obtain:

$$BI_{bis}(p\|q) = \frac{\left(\sum_j p_j\right)^{\lambda}}{\lambda(\lambda-1)} \sum_i \left[\bar{p}_i^{\lambda} - \lambda \bar{p}_i \bar{q}_i^{\lambda-1} - (1-\lambda)\,\bar{q}_i^{\lambda}\right] \tag{5.72}$$

These 2 divergences are invariant by change of scale on "q", moreover, if we disregard the multiplicative factor which depends only on "$\sum_j p_j$", they are also invariant with respect to "p".
Their respective gradients with respect to "q" are written as:

$$\frac{\partial AI_{bis}(p\|q)}{\partial q_j} = \frac{\sum_l p_l}{\lambda \sum_l q_l} \left[\left(\sum_i \bar{p}_i^{\lambda} \bar{q}_i^{1-\lambda}\right) - \bar{p}_j^{\lambda} \bar{q}_j^{\lambda}\right] \tag{5.73}$$

and:

$$\frac{\partial BI_{bis}(p\|q)}{\partial q_j} = \frac{\left(\sum_l p_l\right)^{\lambda}}{\sum_l q_l} \left[\left(\sum_i \bar{p}_i \bar{q}_i^{\lambda-1} - \bar{q}_i^{\lambda}\right) - \left(\bar{p}_j \bar{q}_j^{\lambda-2} - \bar{q}_j^{\lambda-1}\right)\right] \tag{5.74}$$

It can be observed that we have always the fundamental relationship typical of scale invariant divergences:

$$\sum_j q_j \frac{\partial AI_{bis}(p\|q)}{\partial q_j} = 0 \tag{5.75}$$

and

$$\sum_j q_j \frac{\partial BI_{bis}(p\|q)}{\partial q_j} = 0 \tag{5.76}$$

5.6 Generalization in the sense of Cichocki and Ghosh.

5.6.1 Generalized divergence.

In [24], Cichocki proposes a general writing which allows to have a unique expression for the "Alpha divergences " and for the "Beta divergences "; of course, even if the writing is unique, the two remain distinct as we will see. We first introduce the two parameters $a \equiv \lambda_\alpha$ and $b \equiv \lambda_\beta$ specific to the two types of divergence and the expression of the generalized divergence will be:

$$AB\left(p\|q\right) = -\frac{1}{a\left(b-1\right)}\left[\sum_i p_i^a q_i^{b-1} - \frac{a}{a+b-1}\sum_i p_i^{a+b-1} - \frac{b-1}{a+b-1}\sum_i q_i^{a+b-1}\right] \tag{5.77}$$

So the Alpha and Beta divergences can be recovered in the following way:
- If $a+b-1=1$, that is $b-1=1-a$, we have the "Alpha divergence".
- If $a=1$, that only leaves b, we get the "Beta divergence".
To recover Cichocki's writing exactly, we have to do $a \equiv \alpha$ and $b \equiv \beta + 1$. Ghosh [38] suggests something similar. To retrieve our result, we just have to introduce in Ghosh's text, a multiplicative factor $\frac{1}{A+B}$, then to make the replacement: $A \equiv a$ and $B \equiv b-1$ (see appendix 3).

5.6.2 Invariance by change of scale on "q".

To make this generalized divergence invariant by change of scale, we apply on the expression (5.77) the method allowing to determine the factor K_0 and we obtain:

$$K_0 = \left[\frac{\sum_i p_i^a q_i^{b-1}}{\sum_i q_i^{a+b-1}}\right]^{\frac{1}{a}} \tag{5.78}$$

The particular cases corresponding to "Alpha" and "Beta" divergences are immediately revealed:
If $a+b-1=1$, that is if $b-1=1-a$, we have:

$$K_{0,a} = \left[\frac{\sum_i p_i^a q_i^{1-a}}{\sum_i q_i}\right]^{\frac{1}{a}} \tag{5.79}$$

If $a=1$ we obtain:

$$K_{0,b} = \frac{\sum_i p_i q_i^{b-1}}{\sum_i q_i^b} \tag{5.80}$$

This being so, the generalized divergence made invariant to the change of scale is then written down, all calculations done:

$$ABI\left(p\|q\right) = \frac{1}{\left(b-1\right)\left(a+b-1\right)}\left\{\sum_i p_i^{a+b-1} - \left(\sum_i p_i^a q_i^{b-1}\right)^{\frac{a+b-1}{a}}\left(\sum_i q_i^{a+b-1}\right)^{\frac{1-b}{a}}\right\} \tag{5.81}$$

The calculation of the gradient with respect to "q" leads to:

$$-\frac{\partial ABI\left(p\|q\right)}{\partial q_j}=\frac{1}{a}\frac{\left(\sum_i p_i^a q_i^{b-1}\right)^{\frac{a+b-1}{a}}}{\left(\sum_i q_i^{a+b-1}\right)^{\frac{b-1}{a}}}\left[\frac{p_j^a q_j^{b-2}}{\sum_i p_i^a q_i^{b-1}}-\frac{q_j^{a+b-2}}{\sum_i q_i^{a+b-1}}\right] \tag{5.82}$$

We still have the same relationship:

$$\sum_j q_j\frac{\partial ABI\left(p\|q\right)}{\partial q_j}=0 \tag{5.83}$$

Once again, the two special cases are recovered.

5.6.3 Logarithmic form.

Disregarding the multiplicative factor, we have always a difference of two positive terms, therefore the applying the generalized logarithm or the logarithm on each of these terms leads to:

$$\begin{aligned}LABI\left(p\|q\right)=&\frac{1}{\left(b-1\right)\left(a+b-1\right)}\left\{\log\sum_i p_i^{a+b-1}-\right.\\&\frac{a+b-1}{a}\log\sum_i p_i^a q_i^{b-1}-\\&\left.\frac{1-b}{a}\log\sum_i q_i^{a+b-1}\right\}\end{aligned} \tag{5.84}$$

If we compute the gradient with respect to "q", we obtain:

$$-\frac{\partial LABI\left(p\|q\right)}{\partial q_j}=\frac{1}{a}\left[\frac{p_j^a q_j^{b-2}}{\sum_i p_i^a q_i^{b-1}}-\frac{q_j^{a+b-2}}{\sum_i q_i^{a+b-1}}\right] \tag{5.85}$$

As it is always the case with invariant divergences, we have:

$$\sum_j q_j\frac{\partial LABI\left(p\|q\right)}{\partial q_j}=0 \tag{5.86}$$

The two special cases are retrieved:
- If $a+b-1=1$, we have:

$$-\frac{\partial LABI\left(p\|q\right)}{\partial q_j}=\frac{1}{a}\left[\frac{p_j^a q_j^{-a}}{\sum_i p_i^a q_i^{1-a}}-\frac{1}{\sum_i q_i}\right]\equiv-\frac{\partial LAI\left(p\|q\right)}{\partial q_j} \tag{5.87}$$

- If $a = 1$, we have:

$$-\frac{\partial LABI\left(p\|q\right)}{\partial q_j} = \frac{p_j q_j^{b-2}}{\sum_i p_i q_i^{b-1}} - \frac{q_j^{b-1}}{\sum_i q_i^b} \equiv -\frac{\partial LBI\left(p\|q\right)}{\partial q_j} \tag{5.88}$$

Another way to implement invariance.

From the generalized divergence (5.77), we introduce the invariance factor $K\left(p,q\right) = \frac{\sum_j p_j}{\sum_j q_j}$ and we reach another form of the generalized divergence invariant by a scale change which is written:

$$\begin{aligned} ABI_{bis}\left(p\|q\right) = &-\frac{\left(\sum_l p_l\right)^{a+b-1}}{a\left(b-1\right)}\left[\sum_i \bar{p}_i^{(a)}\bar{q}_i^{(b-1)} - \frac{a}{a+b-1}\sum_i \bar{p}_i^{(a+b-1)}\right. \\ &\left. -\frac{b-1}{a+b-1}\sum_i \bar{q}_i^{(a+b-1)}\right] \end{aligned} \tag{5.89}$$

Its gradient with respect to "q" is written:

$$\begin{aligned} \frac{\partial ABI_{bis}\left(p\|q\right)}{\partial q_j} = &\frac{\left(\sum_l p_l\right)^{a+b-1}}{a\sum_l q_l}\left(\sum_i \left(\bar{p}_i\right)^a\left(\bar{q}_i\right)^{b-1} - \sum_i \left(\bar{q}_i\right)^{a+b-1}\right. \\ &\left. +\left(\bar{q}_j\right)^{a+b-2} - \left(\bar{p}_j\right)^a\left(\bar{q}_j\right)^{b-2}\right) \end{aligned} \tag{5.90}$$

We can verify that:

$$\sum_j q_j \frac{\partial ABI_{bis}\left(p\|q\right)}{\partial q_j} = 0 \tag{5.91}$$

5.6.4 Transition to Renyi's divergence.

The divergence (5.84) can be written in an alternative way:

$$\begin{aligned} LABI\left(p\|q\right) = &\frac{1}{a+b-1}\left\{\frac{1}{b-1}\log\sum_i p_i^{a+b-1} - \right. \\ &\frac{a+b-1}{a\left(b-1\right)}\log\sum_i p_i^a q_i^{b-1} + \\ &\left.\frac{1}{a}\log\sum_i q_i^{a+b-1}\right\} \end{aligned} \tag{5.92}$$

To recover Renyi's divergence, one must first make $a+b-1=1$, i.e. $b-1=1-a$, with $a \equiv \alpha$ to recover Renyi's usual notation, and one obtains something which is the invariant version of the "α divergence" in logarithmic form:

$$LAI_{\alpha}(p\|q) = \frac{1}{1-\alpha}\log\sum_i p_i - \frac{1}{\alpha(1-\alpha)}\log\sum_i p_i^{\alpha} q_i^{1-\alpha} + \frac{1}{\alpha}\log\sum_i q_i \tag{5.93}$$

This expression is of course also valid when the proportionality factor K is equal to 1; we can remark that the relation (5.93) can also be written:

$$LABI(p\|q) = \frac{1}{\alpha(\alpha-1)}\log\sum_i (\bar{p}_i)^{\alpha}(\bar{q}_i)^{1-\alpha} \tag{5.94}$$

With $\bar{p}_i = \frac{p_i}{\sum_j p_j}$ and $\bar{q}_i = \frac{q_i}{\sum_j q_j}$.
This is a Renyi divergence with normalized variables.
Now, if we assume we're dealing with probability densities, $\sum_i p_i = \sum_i q_i = 1$ we get exactly the Renyi form:

$$R(p\|q) = LABIS(p\|q) = \frac{1}{\alpha(\alpha-1)}\log\sum_i p_i^{\alpha} q_i^{1-\alpha} \tag{5.95}$$

Once we have considered that $\sum_i p_i = \sum_i q_i$ it is of course senseless to consider a possibility of invariance, even if these sums are different from 1.

5.6.5 Dual Generalized Divergence.

From the expression (5.77), the generalized dual divergence is written:

$$AB(q\|p) = -\frac{1}{a(b-1)}\left[\sum_i q_i^a p_i^{b-1} - \frac{a}{a+b-1}\sum_i q_i^{a+b-1} - \frac{b-1}{a+b-1}\sum_i p_i^{a+b-1}\right] \tag{5.96}$$

From there, and according to the usual method, one can derive the invariance factor $K_0(p,q)$ to be inserted in $AB(K_0 q\|p)$ to obtain an invariant form of this divergence; one thus obtains:

$$K_0(p,q) = \left[\frac{\sum_i q_i^a p_i^{b-1}}{\sum_i q_i^{a+b-1}}\right]^{\frac{1}{b-1}} \tag{5.97}$$

After inserting this expression in $AB\left(K_0 q \| p\right)$, one obtains the invariant dual generalized divergence that one can write down all simplifications made:

$$ABI\left(q\|p\right) = \frac{1}{a\left(a+b-1\right)}\left\{\sum_i p_i^{a+b-1} - \left[\frac{\sum_i q_i^a p_i^{b-1}}{\sum_i q_i^{a+b-1}}\right]^{\frac{a}{b-1}} \sum_i q_i^a p_i^{b-1}\right\} \tag{5.98}$$

Which can also be written in synthetic form:

$$ABI\left(q\|p\right) = \frac{1}{a\left(a+b-1\right)}\left\{\sum_i p_i^{a+b-1} - K_0^a \sum_i q_i^a p_i^{b-1}\right\} \tag{5.99}$$

We can verify that this divergence is invariant with respect to "q". The calculation of the gradient leads to:

$$\begin{aligned}\frac{\partial ABI\left(q\|p\right)}{\partial q_j} = -\frac{1}{b-1}\Bigg\{&\left[\frac{\sum_i q_i^a p_i^{b-1}}{\sum_i q_i^{a+b-1}}\right]^{\frac{a}{b-1}} q_j^{a-1} p_j^{b-1} \\ &- \left[\frac{\sum_i q_i^a p_i^{b-1}}{\sum_i q_i^{a+b-1}}\right]^{\frac{a+b-1}{b-1}} q_j^{a+b-2}\Bigg\}\end{aligned} \tag{5.100}$$

This expression can also be written:

$$\begin{aligned}\frac{\partial ABI\left(q\|p\right)}{\partial q_j} = -\frac{1}{b-1}\frac{\left(\sum_i q_i^a p_i^{b-1}\right)^{\frac{a+b-1}{b-1}}}{\left(\sum_i q_i^{a+b-1}\right)^{\frac{a}{b-1}}}\Bigg[&\frac{1}{\sum_i q_i^a p_i^{b-1}} q_j^{a-1} p_j^{b-1} \\ &- \frac{1}{\sum_i q_i^{a+b-1}} q_j^{a+b-2}\Bigg]\end{aligned} \tag{5.101}$$

We can note that we still have:

$$\sum_j q_j \frac{\partial ABI\left(q\|p\right)}{\partial q_j} = 0 \tag{5.102}$$

Logarithmic form.

By applying the logarithm on the two terms of the difference that appears in (5.100), we obtain the corresponding logarithmic form which is written

for example:

$$LABI\left(q\|p\right)=\frac{1}{a\left(a+b-1\right)}\left\{\log\sum_{i}p_{i}^{a+b-1}+\frac{a}{b-1}\log\sum_{i}q_{i}^{a+b-1}\right.$$
$$\left.-\frac{a+b-1}{b-1}\log\sum_{i}q_{i}^{a}p_{i}^{b-1}\right\} \tag{5.103}$$

This divergence is invariant not only with respect to "q" but also with respect to "p."
The calculation of the gradient leads to:

$$\frac{\partial LABI\left(q\|p\right)}{\partial q_{j}}=-\frac{1}{b-1}\left\{\frac{1}{\sum_{i}q_{i}^{a}p_{i}^{b-1}}q_{j}^{a-1}p_{j}^{b-1}-\frac{1}{\sum_{i}q_{i}^{a+b-1}}q_{j}^{a+b-2}\right\} \tag{5.104}$$

It can be seen that this expression is immediately deduced from (5.101) by deleting in the latter, the constant multiplicative factor:

$$\frac{\left(\sum_{i}q_{i}^{a}p_{i}^{b-1}\right)^{\frac{a+b-1}{b-1}}}{\left(\sum_{i}q_{i}^{a+b-1}\right)^{\frac{a}{b-1}}} \tag{5.105}$$

That observation is always true.
It is worth noting that we still have:

$$\sum_{j}q_{j}\frac{\partial LABI\left(q\|p\right)}{\partial q_{j}}=0 \tag{5.106}$$

5.6.6 Remark.

The results of the previous section can be found from the work of Ghosh et al.[38].
Indeed, we can consider that the divergence analyzed in [38] is a generalization of that of BHHJ [11]; it is therefore in fact the Beta divergence (type B divergence)(5.20).
The details for making these connections are given in **Appendix 3**.

chapter 6 - Other classical divergences.

6.1 CHI2 (χ^2) divergences.

6.1.1 Neyman's CHI2 divergence.

This divergence is also called W Kagan's divergence [8], [10], [9]; it is a Csiszär divergence based on the standard convex function v.fig.(6.1):

$$f_c(x) = (x-1)^2 \tag{6.1}$$

One obtains immediately the "χ^2_N" divergence of Neymann [73]:

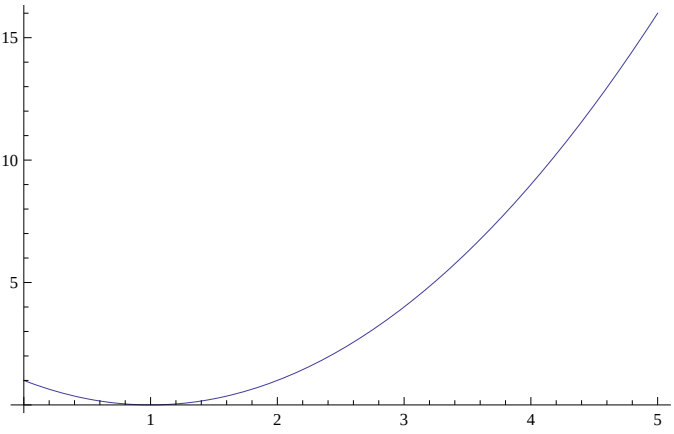

Figure 6.1: Function $f_c(x) = (x-1)^2$

$$\chi^2_N(p\|q) = \sum_i \frac{(p_i - q_i)^2}{q_i} \tag{6.2}$$

The gradient with respect to "q" is written:

$$\frac{\partial \chi^2_N(p\|q)}{\partial q_j} = 1 - \frac{p_j^2}{q_j^2} \tag{6.3}$$

If now we introduce in (6.2) a first simplification: $\sum_i p_i = \sum_i q_i$, we can obtain the divergence:

$$\chi^2_N(p\|q) = \sum_i \frac{p_i^2}{q_i} - p_i \tag{6.4}$$

It is derived from the simple convex function see fig.(6.2):

$$f_1(x) = x^2 - x \tag{6.5}$$

But we can also obtain the divergence:

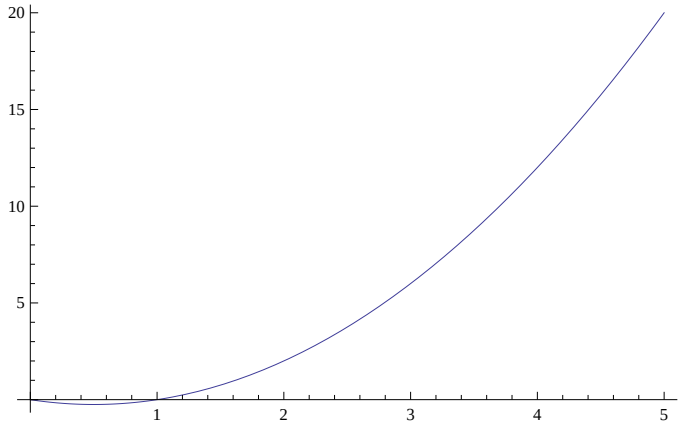

Figure 6.2: Function $f_1(x) = x^2 - x$

$$\chi^2_N(p\|q) = \sum_i \frac{p_i^2}{q_i} - q_i \tag{6.6}$$

which is based on the simple convex function:

$$f_2(x) = x^2 - 1 \tag{6.7}$$

In both cases, an additional simplification $\sum_i p_i = \sum_i q_i = 1$ allows to obtain:

$$\chi^2_N(p\|q) = \left(\sum_i \frac{p_i^2}{q_i}\right) - 1 \tag{6.8}$$

An invariant form of the divergence (6.2) can be obtained; indeed, the derivation of the nominal invariance factor leads to the explicit solution:

$$K_0(p,q) = \sqrt{\frac{\sum_i p_i^2 q_i^{-1}}{\sum_i q_i}} \tag{6.9}$$

By introducing this expression into (6.2), we obtain the invariant divergence with respect to "q":

$$\chi^2_N I(p\|q) = 2\left[\left(\sum_i \frac{p_i^2}{q_i}\sum_i q_i\right)^{\frac{1}{2}} - \sum_i p_i\right] = 2\left[K_0\sum_i q_i - \sum_i p_i\right] \tag{6.10}$$

The gradient with respect to "q" is written as:

$$\frac{\partial \chi^2_N I(p\|q)}{\partial q_j} = \left(\frac{\sum_i \frac{p_i^2}{q_i}}{\sum_i q_i}\right)^{\frac{1}{2}}\left[1 - \frac{\sum_i q_i \frac{p_j^2}{q_j^2}}{\sum_i \frac{p_i^2}{q_i}}\right] \tag{6.11}$$

Starting from the expression (6.10), we obtain the logarithmic form:

$$L\chi^2_N I(p\|q) = \log\sum_i \frac{p_i^2}{q_i} - \log\frac{\left(\sum_i p_i\right)^2}{\left(\sum_i q_i\right)} \tag{6.12}$$

Note that this divergence is not only invariant with respect to "q", but also with respect to "p". The corresponding gradient with respect to "q" is written:

$$\frac{\partial L\chi^2_N I(p\|q)}{\partial q_j} = \left(\frac{1}{\sum_i q_i}\right)\left[1 - \frac{\sum_i q_i \frac{p_j^2}{q_j^2}}{\sum_i \frac{p_i^2}{q_i}}\right] \tag{6.13}$$

6.1.2 Pearson's Chi2 divergence.

The dual divergence $\chi^2_N(q\|p)$ is based on the standard convex function "$\breve{f}_c$" see fig.(6.3), the mirror function of "f_c" (6.1):

$$\breve{f}_c(x) = \frac{(x-1)^2}{x} \tag{6.14}$$

This function diverges at its origin.
The divergence we obtained is the Pearson's χ^2_P written:

$$\chi^2_P(p\|q) = \chi^2_N(q\|p) = \sum_i \frac{(q_i - p_i)^2}{p_i} \tag{6.15}$$

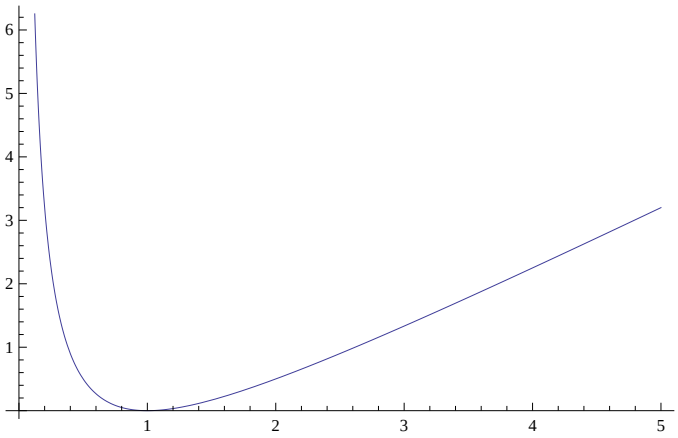

Figure 6.3: Function $\breve{f}_c(x) = \frac{(x-1)^2}{x}$

The gradient with respect to "q" is written as:

$$\frac{\partial \chi_P^2(p\|q)}{\partial q_j} = 2\left(\frac{q_j}{p_j} - 1\right) \tag{6.16}$$

If now we introduce a first simplification: $\sum_i p_i = \sum_i q_i$, we can obtain the divergence:

$$\chi_P^2(p\|q) = \sum_i \frac{q_i^2}{p_i} - q_i \tag{6.17}$$

This divergence is built in the sense of Csiszär on the simple convex function:

$$f(x) = \frac{1}{x} - 1 \tag{6.18}$$

The extreme simplification $\sum_i p_i = \sum_i q_i = 1$, yields:

$$\chi_P^2(p\|q) = \left(\sum_i \frac{q_i^2}{p_i}\right) - 1 \tag{6.19}$$

An invariant form of the divergence (6.15) can be obtained; indeed, the calculation of the nominal invariance factor leads to the explicit solution:

$$K_0(p,q) = \frac{\sum_i q_i}{\sum_i q_i^2 p_i^{-1}} \tag{6.20}$$

By introducing this expression into (6.15), we obtain the invariant divergence with respect to "q":

$$\chi_P^2 I\left(p\|q\right) = \sum_i p_i - \frac{\left(\sum_i q_i\right)^2}{\sum_i \frac{q_i^2}{p_i}} = \sum_i p_i - K_0 \sum_i q_i \tag{6.21}$$

The gradient with respect to "q" is written:

$$\frac{\partial \chi_P^2 I\left(p\|q\right)}{\partial q_j} = \frac{\sum_i q_i}{\sum_i \frac{q_i^2}{p_i}} \left(\frac{\sum_i q_i}{\sum_i \frac{q_i^2}{p_i}} \frac{q_j}{p_j} - 1 \right) \tag{6.22}$$

We derive the logarithmic form of the divergence (6.21):

$$L\chi_P^2 I\left(p\|q\right) = \log \sum_i \frac{q_i^2}{p_i} - \log \frac{\left(\sum_i q_i\right)^2}{\sum_i p_i} \tag{6.23}$$

Note that this divergence is not only invariant with respect to "q", but also with respect to "p".
Its gradient with respect to "q" is written:

$$\frac{\partial L\chi_P^2 I\left(p\|q\right)}{\partial q_j} = \frac{1}{\sum_i q_i} \left(\frac{\sum_i q_i}{\sum_i \frac{q_i^2}{p_i}} \frac{q_j}{p_j} - 1 \right) \tag{6.24}$$

6.1.3 Extensions of Chi2 divergences.

An extension of the χ^2 divergences is mentioned in [25] and [77]. It makes it possible, by action on the single parameter "α" to move gradually from Neyman's χ_N^2 divergence ($\alpha = 1$) to Pearson's χ_P^2 divergence ($\alpha = 0$).
This divergence that the authors attribute to Rukhin [86] is written:

$$RU_\alpha\left(p\|q\right) = \sum_i \frac{\left(p_i - q_i\right)^2}{\alpha q_i + \left(1-\alpha\right) p_i} \tag{6.25}$$

It is built in the Csiszär sense on the standard convex function:

$$f_c\left(x\right) = \frac{\left(x-1\right)^2}{\alpha + \left(1-\alpha\right) x} \tag{6.26}$$

The gradient of $RU_\alpha\left(p\|q\right)$, (6.25), with respect to "q" is written as follows:

$$\frac{\partial RU_\alpha\left(p\|q\right)}{\partial q_j} = \frac{\left(p_j - q_j\right)\left(\alpha p_j - \alpha q_j - 2p_j\right)}{\left[\alpha q_j + \left(1-\alpha\right) p_j\right]^2} \tag{6.27}$$

The particular cases corresponding to the χ^2 divergences of Neyman and Pearson can be found immediately.

6.1.4 Chi2 relative divergences.

Some ambiguity exists in the literature as to the denomination of the χ^2 relative divergences; in any case, it is a matter of replacing "p" or "q" by the weighted sum $\alpha p + (1-\alpha) q$, $0 \leq \alpha \leq 1$ in the χ^2 divergences of the preceding sections.
For example, if in the Neyman χ^2_N Divergence (6.2), we replace "q" by $\alpha p + (1-\alpha) q$, we obtain an expression that is referred to in the literature as the relative Pearson Divergence; this is only a detail.

6.2 Hellinger's divergence.

It is a divergence built in the sense of Csiszär on the standard convex function:

$$f_c(x) = \left(\sqrt{x} - 1\right)^2 \tag{6.28}$$

The divergence obtained is symmetrical; it is written as follows:

$$H(p\|q) = \sum_i \left(\sqrt{p_i} - \sqrt{q_i}\right)^2 \tag{6.29}$$

Its gradient with respect to "q" is given by:

$$\frac{\partial H(p\|q)}{\partial q_j} = 1 - \sqrt{\frac{p_j}{q_j}} \tag{6.30}$$

An invariant form (with respect to "q") of this divergence can be obtained by using the nominal invariance factor which is expressed as follows:

$$K_0(p,q) = \left[\frac{\sum_i \sqrt{p_i q_i}}{\sum_i q_i}\right]^2 \tag{6.31}$$

The invariant divergence that results is written as follows:

$$HI(p\|q) = \sum_i p_i - \frac{\left(\sum_i \sqrt{p_i q_i}\right)^2}{\sum_i q_i} \tag{6.32}$$

Its gradient with respect to "q" is given by:

$$\frac{\partial HI(p\|q)}{\partial q_j} = \left[\frac{\sum_i \sqrt{p_i q_i}}{\sum_i q_i}\right]^2 - \frac{\sum_i \sqrt{p_i q_i}}{\sum_i q_i} \frac{p_j}{\sqrt{p_j q_j}} \tag{6.33}$$

6.3 Triangular discrimination.

This divergence mentioned by Tanéja [91] is built in the sense of Csiszär on the standard convex function:

$$f_c(x) = \frac{(x-1)^2}{x+1} \tag{6.34}$$

This gives the symmetrical divergence:

$$DT(p\|q) = \sum_i \frac{(p_i - q_i)^2}{p_i + q_i} \tag{6.35}$$

Its gradient with respect to "q" is given by:

$$\frac{\partial DT(p\|q)}{\partial q_j} = 1 - \frac{4p_j^2}{(p_j + q_j)^2} \tag{6.36}$$

It is equal to zéro if $p_j = q_j \ \forall j$.
This divergence is to be linked to the one discussed in the following section.

6.4 Harmonic mean divergence of Toussaint.

She's quoted in Basseville, [8],[10],[9]; it is close to what we would call "M_{AH}" [91] in Chapter 7.
It's a Csiszär divergence based on the simple convex function see.fig.(6.4):

$$f(x) = x - \frac{2x}{1+x} \tag{6.37}$$

We obtain the divergence:

$$T(p\|q) = \sum_i \left(p_i - \frac{2p_i q_i}{p_i + q_i} \right) \tag{6.38}$$

Its gradient with respect to "q" is given by:

$$\frac{\partial T(p\|q)}{\partial q_j} = -2 \left(\frac{p_j}{p_j + q_j} \right)^2 \tag{6.39}$$

This gradient is never equal to zero.
The dual divergence is written as follows:

$$T(q\|p) = \sum_i \left(q_i - \frac{2p_i q_i}{p_i + q_i} \right) \tag{6.40}$$

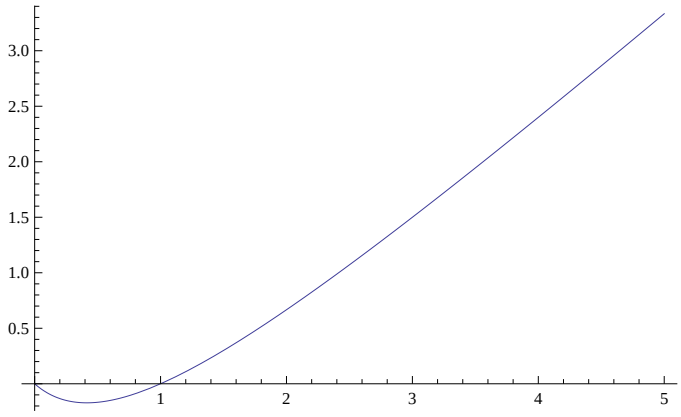

Figure 6.4: Function $f(x) = x - \frac{2x}{1+x}$

It is based on the mirror function of the previous one, i.e. on the simple convex function see.fig.(6.5):

$$\breve{f}(x) = 1 - \frac{2x}{1+x} \tag{6.41}$$

Its gradient with respect to "q" is given by:

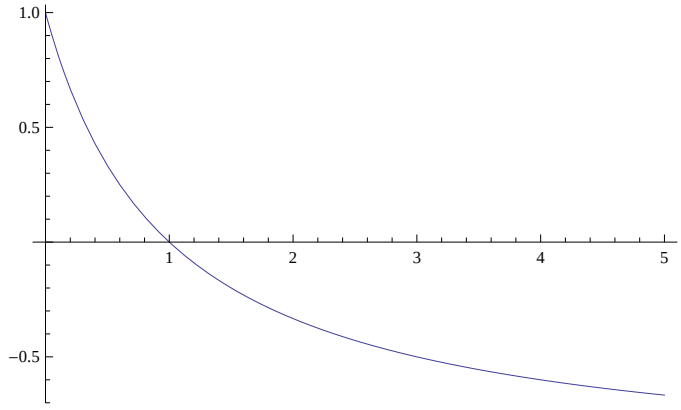

Figure 6.5: Function $\breve{f}(x) = 1 - \frac{2x}{1+x}$

$$\frac{\partial T(q\|p)}{\partial q_j} = 1 - 2\left(\frac{p_j}{p_j + q_j}\right)^2 \tag{6.42}$$

He is not zero for $p_j = q_j \ \forall j$, hence the problem previously mentioned.

If we build the standard convex function from the function $f(x)$ or the mirror function $\breve{f}(x)$, we obtain one and the same function see.fig.(6.6):

$$f_c(x) = \frac{1}{2}\frac{(x-1)^2}{x+1} \tag{6.43}$$

This result is indicative of the fact that the Csiszär divergence built on this

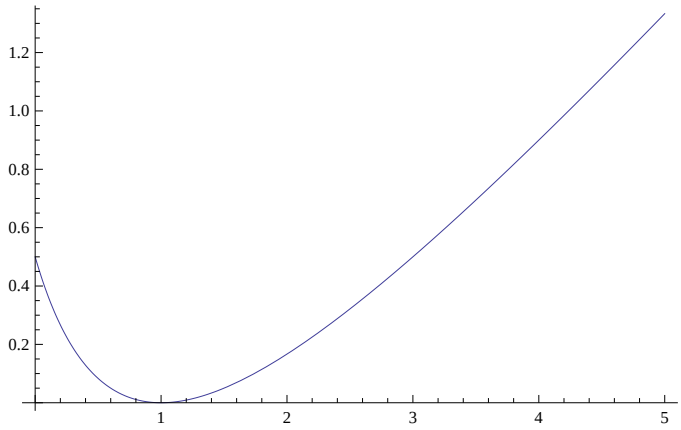

Figure 6.6: Function $f_c(x) = \frac{1}{2}\frac{(x-1)^2}{x+1}$

function will be symmetrical.
It is up to a multiplicative factor, the function that allows us to construct the Triangular Discrimination in the previous section and that will enable us to build the arithmetic-harmonic mean M_{AH} divergence. [91].

6.5 Henze-Penrose divergence.

This divergence proposed in [44] is used in [72] with β <1, in the form:

$$HP(p\|q) = \sum_i \frac{\beta^2 p_i^2 + (1-\beta)^2 q_i^2}{\beta p_i + (1-\beta) q_i} \tag{6.44}$$

This form is oversimplified because it is not zero if $p_i = q_i \ \forall i$, even if one uses variables summed to 1.
The first modification consists in making it correct from this point of view, that is to say equal to zero when $p_i = q_i \ \forall i$; we obtain 2 possible forms:

$$HP1(p\|q) = \sum_i \frac{\beta^2 p_i^2 + (1-\beta)^2 q_i^2}{\beta p_i + (1-\beta) q_i} - \sum_i \left(\beta^2 + (1-\beta)^2\right) q_i \tag{6.45}$$

and

$$HP2\left(p\|q\right)=\sum_{i}\frac{\beta^{2}p_{i}^{2}+\left(1-\beta\right)^{2}q_{i}^{2}}{\beta p_{i}+\left(1-\beta\right)q_{i}}-\sum_{i}\left(\beta^{2}+\left(1-\beta\right)^{2}\right)p_{i} \tag{6.46}$$

The first expression is a Csiszär divergence built on the simple convex function:

$$f_{1}\left(x\right)=\frac{\beta^{2}x^{2}+\left(1-\beta\right)^{2}}{\beta x+\left(1-\beta\right)}-\left[\beta^{2}+\left(1-\beta\right)^{2}\right] \tag{6.47}$$

illustrated in the figure (6.7), for $\beta=0.5$.

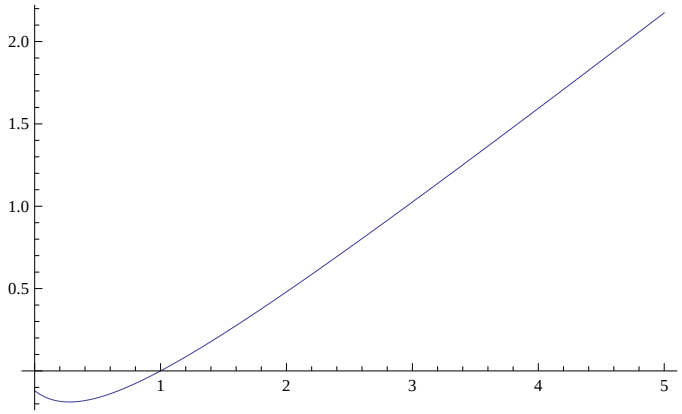

Figure 6.7: Function $f_{1}\left(x\right)=\frac{\beta^{2}x^{2}+\left(1-\beta\right)^{2}}{\beta x+\left(1-\beta\right)}-\left[\beta^{2}+\left(1-\beta\right)^{2}\right]$, $\beta=0.5$

The second is a Csiszär divergence built on the simple convex function:

$$f_{2}\left(x\right)=\frac{\beta^{2}x^{2}+\left(1-\beta\right)^{2}}{\beta x+\left(1-\beta\right)}-\left[\beta^{2}+\left(1-\beta\right)^{2}\right]x \tag{6.48}$$

shown in the figure (6.8), for $\beta=0.5$.

If we construct the standard convex function from f_{1} or f_{2}, we obtain the same function which is written:

$$\begin{aligned}f_{c}\left(x\right)=&\frac{\beta^{2}x^{2}+\left(1-\beta\right)^{2}}{\beta x+\left(1-\beta\right)}-\left[\beta^{2}+\left(1-\beta\right)^{2}\right]\\&+\left(x-1\right)\left[2\beta^{3}-4\beta^{2}+\beta\right]\end{aligned} \tag{6.49}$$

illustrated in the figure (6.9), for $\beta=0.5$.

By using the notation:

$$A=\beta^{2}+\left(1-\beta\right)^{2};\ B=2\beta^{3}-4\beta^{2}+\beta \tag{6.50}$$

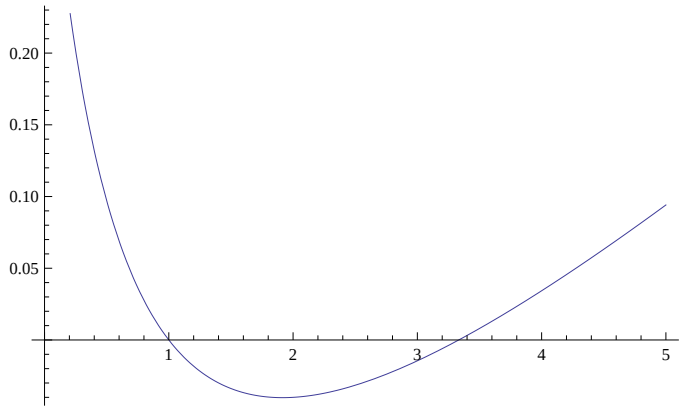

Figure 6.8: Function $f_2(x) = \frac{\beta^2 x^2 + (1-\beta)^2}{\beta x + (1-\beta)} - \left[\beta^2 + (1-\beta)^2\right] x$, $\beta = 0.5$

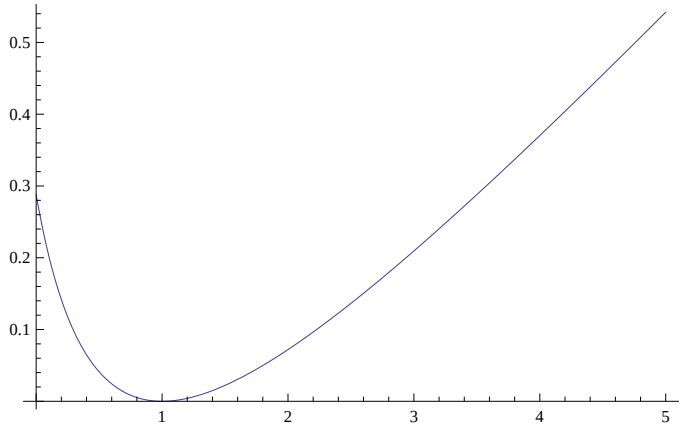

Figure 6.9: Function $f_c(x) = \frac{\beta^2 x^2 + (1-\beta)^2}{\beta x + (1-\beta)} - \left[\beta^2 + (1-\beta)^2\right] + (x-1)\left[2\beta^3 - 4\beta^2 + \beta\right]$, $\beta = 0.5$

the Csiszär divergence built on the standard convex function (6.49) is written:

$$HPC(p\|q) = \sum_i \frac{\beta^2 p_i^2 + (1-\beta)^2 q_i^2}{\beta p_i + (1-\beta) q_i} + \sum_i B p_i - \sum_i (A+B) q_i \quad (6.51)$$

Its gradient with respect to "q" is written as:

$$\frac{\partial HPC\left(p\|q\right)}{\partial q_j} = \left(1-\beta\right)\left\{1-\frac{2\beta^2 p_j^2}{\left[\beta p_j+\left(1-\beta\right)q_j\right]^2}\right\} - \left(2\beta^3-2\beta^2-\beta+1\right) \tag{6.52}$$

Which can also be written more simply:

$$\frac{\partial HPC\left(p\|q\right)}{\partial q_j} = 2\beta^2\left(1-\beta\right)\left\{1-\frac{p_j^2}{\left[\beta p_j+\left(1-\beta\right)q_j\right]^2}\right\} \tag{6.53}$$

The dual divergence, built on the function $\breve{f}_c$ mirroring f_c, is written:

$$HPC\left(q\|p\right) = \sum_i \frac{\beta^2 q_i^2+\left(1-\beta\right)^2 p_i^2}{\beta q_i+\left(1-\beta\right)p_i} + \sum_i Bq_i - \sum_i \left(A+B\right)p_i \tag{6.54}$$

Its gradient with respect to "q" is written as:

$$\frac{\partial HPC\left(q\|p\right)}{\partial q_j} = 2\beta\left(1-\beta\right)^2\left\{1-\frac{p_j^2}{\left[\beta q_j+\left(1-\beta\right)p_j\right]^2}\right\} \tag{6.55}$$

These gradients are of course zero for $p_j = q_j\ \forall j$.

6.6 Polya information divergence.

Here we consider the "Polya information divergence" [41] quoted in [40].

6.6.1 Some general observations.

On the basis of the simplified Kullback divergence or Kullback Information expressed as:

$$IKL(a\|b) = \sum_i a_i \log\left(a_i/b_i\right) \tag{6.56}$$

With $\gamma \geq 0$, Grendar and Niven [40] are proposing the divergence:

$$I_\gamma(p\|q) = IKL(p\|q+\gamma p) + \frac{1}{\gamma}IKL(q\|q+\gamma p) + \frac{1+\gamma}{\gamma}\log(1+\gamma) \tag{6.57}$$

which leads to:

$$I_\gamma(p\|q) = \sum_i p_i \log\frac{p_i}{q_i+\gamma p_i} + \frac{1}{\gamma}\sum_i q_i \log\frac{q_i}{q_i+\gamma p_i} + \frac{1+\gamma}{\gamma}\log(1+\gamma) \tag{6.58}$$

This divergence can be used essentially with data of sum equal to 1, i.e. typically probability densities. Indeed, if in the preceding expression we do $p_i = q_i \ \forall i$, we will obtain zero only if $\sum_i p_i = \sum_i q_i = 1$.
This means that the basic convex function that allows us to construct this divergence in the Csiszär sense can have 2 forms:

$$f_1(x) = x \log \frac{x}{1+\gamma x} + \frac{1}{\gamma} \log \frac{1}{1+\gamma x} + \left\{ \frac{1+\gamma}{\gamma} \log(1+\gamma) \right\} x \tag{6.59}$$

shown in the figure (6.10), for $\gamma = 0.8$,

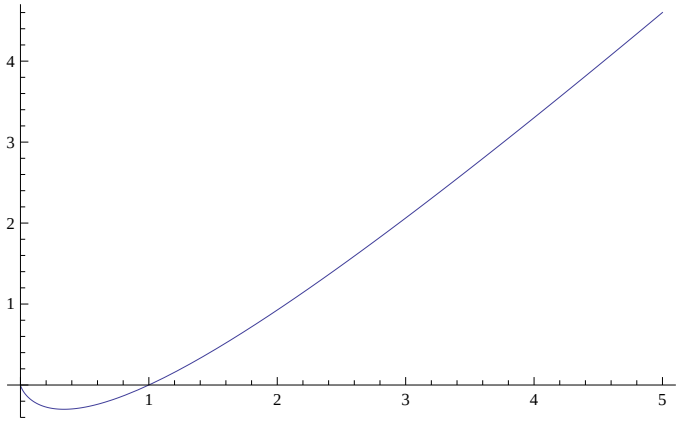

Figure 6.10: Function $f_1(x) = x \log \frac{x}{1+\gamma x} + \frac{1}{\gamma} \log \frac{1}{1+\gamma x} + \frac{(1+\gamma)x}{\gamma} \log(1+\gamma)$, $\gamma = 0.8$

or:

$$f_2(x) = x \log \frac{x}{1+\gamma x} + \frac{1}{\gamma} \log \frac{1}{1+\gamma x} + \frac{1+\gamma}{\gamma} \log(1+\gamma) \tag{6.60}$$

shown in the figure (6.11), for $\gamma = 0.8$,
This being stated, these 2 functions are not standard forms because:

$$f_1'(x) = \log \frac{x}{1+\gamma x} + \frac{1+\gamma}{\gamma} \log(1+\gamma) \ \Rightarrow \ f_1'(1) = \frac{1}{\gamma} \log(1+\gamma) \tag{6.61}$$

$$f_2'(x) = \log \frac{x}{1+\gamma x} \ \Rightarrow \ f_2'(1) = \log \frac{1}{1+\gamma} \tag{6.62}$$

If we construct the corresponding standard convex function, we obtain from $f_1(x)$ or from $f_2(x)$, the same expression:

$$f_c(x) = x \log \frac{x}{1+\gamma x} + \frac{1}{\gamma} \log \frac{1}{1+\gamma x} + \frac{1+\gamma x}{\gamma} \log(1+\gamma) \tag{6.63}$$

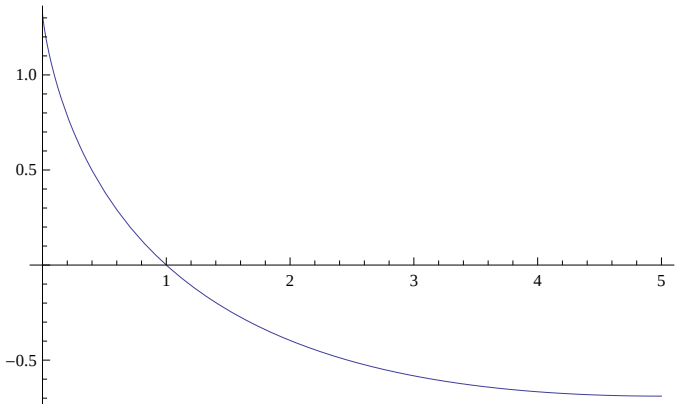

Figure 6.11: Function $f_2(x) = x \log \frac{x}{1+\gamma x} + \frac{1}{\gamma} \log \frac{1}{1+\gamma x} + \frac{1+\gamma}{\gamma} \log(1+\gamma)$, $\gamma = 0.8$

illustrated in the figure (6.12), for $\gamma = 0.8$
This function has all of the desired properties: $f_c(1) = 0$, $f_c'(1) = 0$,

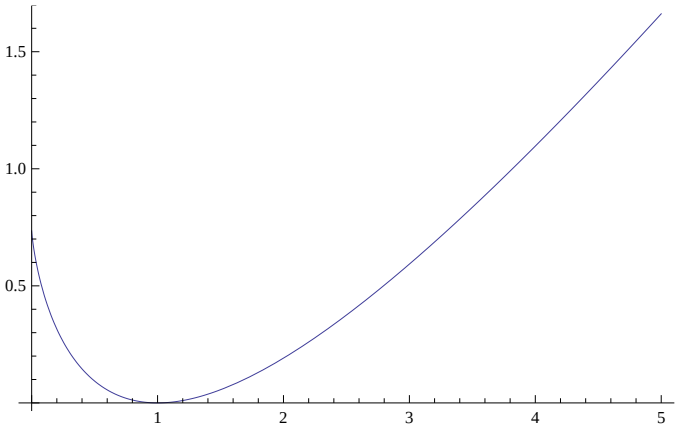

Figure 6.12: Function $f_c(x) = x \log \frac{x}{1+\gamma x} + \frac{1}{\gamma} \log \frac{1}{1+\gamma x} + \frac{1+\gamma x}{\gamma} \log(1+\gamma)$, $\gamma = 0.8$

which makes it possible to construct a Csiszär divergence that is usable in

all generality:

$$IC_{\gamma}\left(p\|q\right)=\sum_{i}p_{i}\log\frac{p_{i}}{q_{i}+\gamma p_{i}}+\frac{1}{\gamma}\sum_{i}q_{i}\log\frac{q_{i}}{q_{i}+\gamma p_{i}}+\frac{\log(1+\gamma)}{\gamma}\sum_{i}q_{i}+\gamma p_{i} \tag{6.64}$$

The gradient with respect to "q" is expressed as:

$$\frac{\partial IC_{\gamma}\left(p\|q\right)}{\partial q_{j}}=\frac{1}{\gamma}\log\frac{q_{j}+\gamma q_{j}}{q_{j}+\gamma p_{j}} \tag{6.65}$$

If we are now interested in the dual divergence, it is built, in the sense of Csiszär, on the mirror function of (6.63),

$$\breve{f}_{c}\left(x\right)=\log\frac{1}{x+\gamma}+\frac{x}{\gamma}\log\frac{x}{x+\gamma}+\frac{x+\gamma}{\gamma}\log\left(1+\gamma\right) \tag{6.66}$$

shown in the figure (6.13), for $\gamma=0.8$:

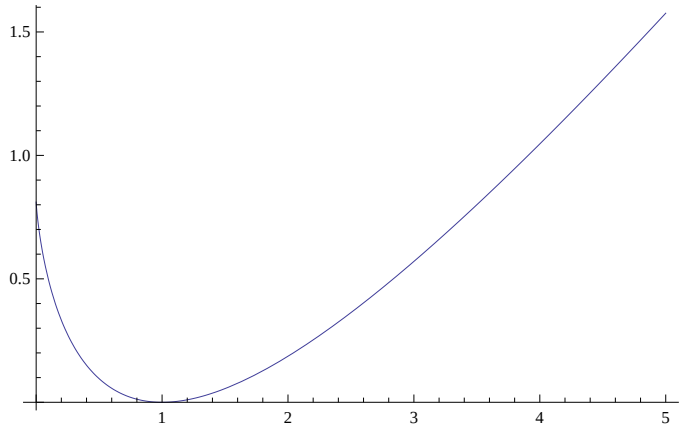

Figure 6.13: Function $\breve{f}_{c}\left(x\right)=\log\frac{1}{x+\gamma}+\frac{x}{\gamma}\log\frac{x}{x+\gamma}+\frac{x+\gamma}{\gamma}\log\left(1+\gamma\right)$, $\gamma=0.8$

This divergence is written:

$$IC_{\gamma}\left(q\|p\right)=\sum_{i}q_{i}\log\frac{q_{i}}{p_{i}+\gamma q_{i}}+\frac{1}{\gamma}\sum_{i}p_{i}\log\frac{p_{i}}{p_{i}+\gamma q_{i}}+\frac{\log(1+\gamma)}{\gamma}\sum_{i}p_{i}+\gamma q_{i} \tag{6.67}$$

The gradient with respect to "q" is expressed as:

$$\frac{\partial IC_{\gamma}\left(q\|p\right)}{\partial q_j} = \log \frac{q_j + \gamma q_j}{p_j + \gamma q_j} \tag{6.68}$$

The symmetrical divergence can be obtained in the Jeffreys sense by making the (half) sum of the 2 preceding conjugated divergences; the divergence obtained is of course always a Csiszär divergence built on a convex function which is the (half) sum of the conjugated functions $f_c\left(x\right)$ and $\breve{f}_c\left(x\right)$.

6.6.2 Invariance by change of scale.

We can try to construct a scale-invariant version of the divergence (6.64) using the standard process.
First, we write:

$$\begin{aligned} IC_{\gamma}\left(p\|Kq\right) = &\sum_i p_i \log \frac{p_i}{Kq_i + \gamma p_i} + \frac{1}{\gamma}\sum_i Kq_i \log \frac{Kq_i}{Kq_i + \gamma p_i} \\ &+ \frac{\log(1+\gamma)}{\gamma}\sum_i Kq_i + \gamma p_i \end{aligned} \tag{6.69}$$

We then calculate the derivative of this expression with respect to K, and in order to determine the expression of K, we write that this derivative is zero.
Which leads us to have to solve:

$$\sum_i q_i \log\left(1 + \frac{\gamma p_i}{Kq_i}\right) = \sum_i q_i \log\left(1+\gamma\right) \tag{6.70}$$

that is:

$$\sum_i q_i \log \frac{\left(1 + \frac{\gamma p_i}{Kq_i}\right)}{(1+\gamma)} = 0 \tag{6.71}$$

There is no explicit K solution. Strictly speaking, this scalar product will be zero if the 2 vectors are orthogonal, or if one of the vectors is zero; however, to illustrate the remark mentioned in Ch.3, section 3.5, we can use as invariance factor:

$$K^* = \frac{\sum_i p_i}{\sum_i q_i} \tag{6.72}$$

By inserting this expression of K^* in (6.69), it comes, with all simplifications made:

$$ICINV_\gamma(p\|q) = \sum_j p_j \left\{ \sum_i \overline{p}_i \log \frac{\overline{p}_i}{\overline{q}_i + \gamma\overline{p}_i} + \frac{1}{\gamma} \sum_i \overline{q}_i \log \frac{\overline{q}_i}{\overline{q}_i + \gamma\overline{p}_i} + \frac{\log(1+\gamma)}{\gamma} \sum_i \overline{q}_i + \gamma\overline{p}_i \right\} \tag{6.73}$$

So, if we calculate the gradient of (6.73) with respect to "q", we obtain all the calculations done:

$$\frac{\partial ICINV_\gamma(p\|q)}{\partial q_j} = \frac{\sum_l p_l}{\gamma \sum_l q_l} \left[\log \frac{\overline{q}_j}{\overline{q}_j + \gamma\overline{p}_j} - \sum_i \overline{q}_i \log \frac{\overline{q}_i}{\overline{q}_i + \gamma\overline{p}_i} \right] \tag{6.74}$$

And we can observe that, as with all the scale invariant divergences , we have:

$$\sum_j q_j \frac{\partial ICINV_\gamma(p\|q)}{\partial q_j} = 0 \tag{6.75}$$

6.6.3 Remark.

We can exhibit a Bregman divergence based on the function $f_c(x)$, which is written:

$$BIC_\gamma(p\|q) = \sum_i p_i \log \frac{p_i}{q_i} - \frac{1}{\gamma}(1+\gamma p_i) \log \frac{1+\gamma p_i}{1+\gamma q_i} \tag{6.76}$$

But, of course, there is no evidence of the convexity of this divergence.

6.7 Bose-Einstein and Fermi-Dirac divergences.

6.7.1 Bose-Einstein divergence.

Such divergence mentioned in Basseville [8, 10, 9], and in the references cited [56], is a Csiszär divergence based on the standard convex function shown in the figure (6.14), for $\alpha = 0.8$:

$$f_{c,\alpha}(x) = x \log x + (\alpha + x) \log \frac{\alpha + 1}{\alpha + x} \qquad \alpha > 0 \tag{6.77}$$

It is written:

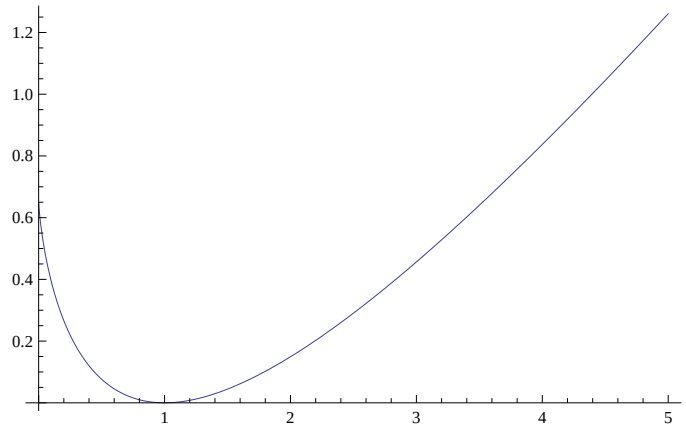

Figure 6.14: Function $f_{c,\alpha}(x) = x \log x + (\alpha + x) \log \frac{\alpha+1}{\alpha+x}$, $\alpha = 0.8$

$$BE1_\alpha(p\|q) = \sum_i p_i \log \frac{p_i}{q_i} + (\alpha q_i + p_i) \log \frac{(\alpha q_i + q_i)}{(\alpha q_i + p_i)} \tag{6.78}$$

However, another divergence appears under the same name in [37]; it is written:

$$BE2(p\|q) = \sum_i p_i \log \frac{p_i}{q_i} + (1 + p_i) \log \frac{(1 + q_i)}{(1 + p_i)} \tag{6.79}$$

It is not a Csiszar divergence.
It is deduced from a work by Furuichi [37] quoted in referring to Kapur [54].
It is, in fact, a Bregman divergence built on the function (6.77) which is written:

$$BE2_\alpha(p\|q) = \sum_i p_i \log \frac{p_i}{q_i} + (\alpha + p_i) \log \frac{(\alpha + q_i)}{(\alpha + p_i)} \tag{6.80}$$

Note that except for a change in parameter, this divergence is similar to (6.76).
If, in the expression (6.80) we introduce the simplification $\alpha = 1$, we recover the form proposed in [37] [54] [55].
From these expressions, we can introduce the Generalized Logarithm as suggested in [37] by referring to Tsallis' entropy [95].
The introduction of Tsallis entropy consists in replacing the Logarithm function by the Generalized Logarithm function denoted "Log_d" (see Appendix 1), which yields:

$$BE1_{\alpha,d}(p\|q) = \sum_i p_i \log_d \frac{p_i}{q_i} + (\alpha q_i + p_i) \log_d \frac{(\alpha q_i + q_i)}{(\alpha q_i + p_i)} \tag{6.81}$$

That is:

$$BE1_{\alpha,d}\left(p\|q\right)=\sum_{i}p_{i}\frac{\left(\frac{p_{i}}{q_{i}}\right)^{1-d}-1}{1-d}+\left(\alpha q_{i}+p_{i}\right)\frac{\left(\frac{\alpha q_{i}+q_{i}}{\alpha q_{i}+p_{i}}\right)^{1-d}-1}{1-d} \tag{6.82}$$

The calculation of the gradient with respect to "q" leads to:

$$\begin{aligned}\frac{\partial BE1_{\alpha,d}\left(p\|q\right)}{\partial q_{j}}=&-\left(\frac{p_{j}}{q_{j}}\right)^{2-d}+\frac{\alpha d}{1-d}\left(\frac{\alpha q_{j}+q_{j}}{\alpha q_{j}+p_{j}}\right)^{1-d}\\&+\left(\alpha+1\right)\left(\frac{\alpha q_{j}+q_{j}}{\alpha q_{j}+p_{j}}\right)^{-d}-\frac{\alpha}{1-d}\end{aligned} \tag{6.83}$$

The limit $d\rightarrow 1$ leads to:

$$\frac{\partial BE1_{\alpha,1}\left(p\|q\right)}{\partial q_{j}}=\alpha\log\frac{\alpha q_{j}+q_{j}}{\alpha q_{j}+p_{j}} \tag{6.84}$$

Which is the gradient of the initial Bose-Einstein divergence $BE1_{\alpha}\left(p\|q\right)$.
If we now consider the form $BE2_{\alpha}\left(p\|q\right)$, we obtain:

$$BE2_{\alpha,d}\left(p\|q\right)=\sum_{i}p_{i}\log_{d}\frac{p_{i}}{q_{i}}+\left(\alpha+p_{i}\right)\log_{d}\frac{\left(\alpha+q_{i}\right)}{\left(\alpha+p_{i}\right)} \tag{6.85}$$

That is:

$$BE2_{\alpha,d}\left(p\|q\right)=\sum_{i}p_{i}\frac{\left(\frac{p_{i}}{q_{i}}\right)^{1-d}-1}{1-d}+\left(\alpha+p_{i}\right)\frac{\left(\frac{\alpha+q_{i}}{\alpha+p_{i}}\right)^{1-d}-1}{1-d} \tag{6.86}$$

For $\alpha=1$, this divergence is referred to as the Bose-Einstein-Tsallis divergence in [37].
The gradient with respect to "q" is written:

$$\frac{\partial BE2_{\alpha,d}\left(p\|q\right)}{\partial q_{j}}=-\left(\frac{p_{j}}{q_{j}}\right)^{2-d}+\left(\frac{\alpha+p_{j}}{\alpha+q_{j}}\right)^{d} \tag{6.87}$$

The limit $d\rightarrow 1$ leads to:

$$\frac{\partial BE2_{\alpha,1}\left(p\|q\right)}{\partial q_{j}}=-\frac{p_{j}}{q_{j}}+\frac{\alpha+p_{j}}{\alpha+q_{j}} \tag{6.88}$$

Which is the gradient of $BE2_{\alpha}\left(p\|q\right)$.

Scale change invariance.

From (6.78) one can write:

$$BE1_{\alpha}\left(p\|Kq\right)=\sum_{i}p_{i}\log\frac{p_{i}}{Kq_{i}}+\left(\alpha Kq_{i}+p_{i}\right)\log\frac{\left(\alpha Kq_{i}+Kq_{i}\right)}{\left(\alpha Kq_{i}+p_{i}\right)} \tag{6.89}$$

and we have:

$$\frac{\partial BE1_{\alpha}\left(p\|Kq\right)}{\partial K}=\alpha\left[\sum_{i}q_{i}\log\frac{\alpha Kq_{i}+Kq_{i}}{\alpha Kq_{i}+p_{i}}\right] \tag{6.90}$$

To cancel this derivative, it is necessary to solve:

$$\sum_{i}q_{i}\log\frac{\alpha Kq_{i}+Kq_{i}}{\alpha Kq_{i}+p_{i}}=0 \tag{6.91}$$

The zero scalar product implies that the corresponding vectors are orthogonal or that one of the vectors is zero. There is no explicit solution, however, we can use:

$$K^{*}=\frac{\sum_{i}p_{i}}{\sum_{i}q_{i}} \tag{6.92}$$

If we introduce this expression of K^{*} in the relation (6.89), we obtain:

$$BE1_{\alpha}I\left(p\|q\right)=\sum_{i}\bar{p}_{i}\log\frac{\bar{p}_{i}}{\bar{q}_{i}}+\left(\alpha\bar{q}_{i}+\bar{p}_{i}\right)\log\frac{\left(\alpha+1\right)\bar{q}_{i}}{\left(\alpha\bar{q}_{i}+\bar{p}_{i}\right)} \tag{6.93}$$

The gradient with respect to "q" expresses as:

$$\frac{\partial BE1_{\alpha}I\left(p\|q\right)}{\partial q_{l}}=\frac{\alpha}{\sum_{j}q_{j}}\left[\log\frac{\left(\alpha+1\right)\bar{q}_{l}}{\left(\alpha\bar{q}_{l}+\bar{p}_{l}\right)}-\sum_{i}\bar{q}_{i}\log\frac{\left(\alpha+1\right)\bar{q}_{i}}{\left(\alpha\bar{q}_{i}+\bar{p}_{i}\right)}\right] \tag{6.94}$$

6.7.2 Fermi-Dirac divergence.

This divergence reported in Basseville [8, 10, 9], and in the mentioned references is a Csiszär divergence based on the standard convex function shown in the figure (6.15), for $\beta=3$:

$$f_{c,\beta}\left(x\right)=x\log x+\left(\beta-x\right)\log\frac{\beta-x}{\beta-1}\qquad\beta>1 \tag{6.95}$$

It is written as:

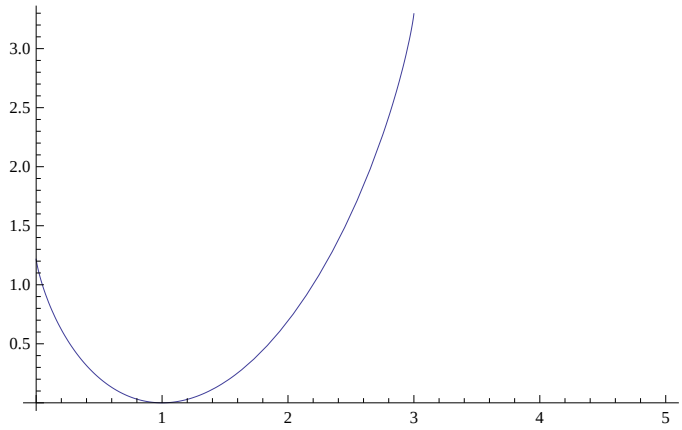

Figure 6.15: Function $f_{c,\beta}(x) = x\log x + (\beta - x)\log\frac{\beta-x}{\beta-1}$, $\beta = 3$

$$FD1_\beta(p\|q) = \sum_i p_i \log\frac{p_i}{q_i} + \sum_i (\beta q_i - p_i)\log\frac{\beta q_i - p_i}{\beta q_i - q_i} \tag{6.96}$$

However, another divergence appears under the same name in [37] [55]; it is written:

$$FD2(p\|q) = \sum_i p_i \log\frac{p_i}{q_i} + (1 - p_i)\log\frac{(1-p_i)}{(1-q_i)} \tag{6.97}$$

This is not a Csiszâr divergence, it is in fact a Bregman divergence built on the simple convex function:

$$f_\beta(x) = x\log x + (\beta - x)\log(\beta - x); \quad x < \beta \tag{6.98}$$

This divergence is written:

$$FD2_\beta(p\|q) = \sum_i p_i \log\frac{p_i}{q_i} + \sum_i (\beta - p_i)\log\frac{\beta - p_i}{\beta - q_i} \tag{6.99}$$

If in this expression, one makes $\beta = 1$, one finds the expression proposed in [37]; we will note that the fact of taking $\beta = 1$ is equivalent to considering that we are dealing with probability densities, i.e. $p_i < 1\ \forall i, q_i < 1\ \forall i$.
From these expressions, we can introduce the Generalized logarithm as suggested in [37] as inspired by the entropy of Tsallis [95].
Using $FD1_\beta(p\|q)$, we obtain:

$$FD1_{\beta,d}(p\|q) = \sum_i p_i \log_d\frac{p_i}{q_i} + \sum_i (\beta q_i - p_i)\log_d\frac{\beta q_i - p_i}{\beta q_i - q_i} \tag{6.100}$$

That is:

$$FD1_{\beta,d}\left(p\|q\right)=\sum_{i}p_{i}\frac{\left(\frac{p_{i}}{q_{i}}\right)^{1-d}-1}{1-d}+\left(\beta q_{i}-p_{i}\right)\frac{\left(\frac{\beta q_{i}-p_{i}}{\beta q_{i}-q_{i}}\right)^{1-d}-1}{1-d} \tag{6.101}$$

The computation of the gradient with respect to "q" leads to:

$$\begin{aligned}\frac{\partial FD1_{\beta,d}\left(p\|q\right)}{\partial q_{j}}=&-\frac{p_{j}}{q_{j}}\left[\left(\frac{p_{j}}{q_{j}}\right)^{1-d}-\left(\frac{\beta q_{j}-p_{j}}{\beta q_{j}-q_{j}}\right)^{1-d}\right]\\&+\beta\frac{\left(\frac{\beta q_{j}-p_{j}}{\beta q_{j}-q_{j}}\right)^{1-d}-1}{1-d}\end{aligned} \tag{6.102}$$

The last term of the second member is the expression of $\beta\log_{d}\left(\frac{\beta q_{i}-p_{i}}{\beta q_{i}-q_{i}}\right)$, so that when $d\rightarrow 1$, we obtain:

$$\frac{\partial FD1_{\beta,1}\left(p\|q\right)}{\partial q_{j}}=\beta\log\frac{\beta q_{j}-p_{j}}{\beta q_{j}-q_{j}} \tag{6.103}$$

Which is equal to $\frac{\partial FD1_{\beta}(p\|q)}{\partial q_{j}}$.
If now, we consider the divergence $FD2_{\beta}\left(p\|q\right)$ with $\beta=1$ and using the generalized logarithm, we have:

$$FD2_{1,d}\left(p\|q\right)=\sum_{i}p_{i}\log_{d}\frac{p_{i}}{q_{i}}+\left(1-p_{i}\right)\log_{d}\frac{\left(1-p_{i}\right)}{\left(1-q_{i}\right)} \tag{6.104}$$

That is:

$$FD2_{1,d}\left(p\|q\right)=\sum_{i}p_{i}\frac{\left(\frac{p_{i}}{q_{i}}\right)^{1-d}-1}{1-d}+\left(1-p_{i}\right)\frac{\left(\frac{1-p_{i}}{1-q_{i}}\right)^{1-d}-1}{1-d} \tag{6.105}$$

This divergence is referred to as the Fermi-Dirac-Tsallis divergence in [37].
The calculation of the gradient with respect to "q" leads to:

$$\frac{\partial FD2_{1,d}\left(p\|q\right)}{\partial q_{j}}=-\left(\frac{p_{j}}{q_{j}}\right)^{2-d}+\left(\frac{1-p_{j}}{1-q_{j}}\right)^{2-d} \tag{6.106}$$

Obviously, if $d\rightarrow 1$, we obtain:

$$\frac{\partial FD2_{1,1}\left(p\|q\right)}{\partial q_{j}}=-\left(\frac{p_{j}}{q_{j}}\right)+\left(\frac{1-p_{j}}{1-q_{j}}\right) \tag{6.107}$$

Which is equal to $\frac{\partial FD2_{\beta}(p\|q)}{\partial q_{j}}$, with $\beta=1$.

Invariance by change of scale.

From (6.96), we have:

$$FD1_\beta(p\|Kq) = \sum_i p_i \log \frac{p_i}{Kq_i} + \sum_i (\beta Kq_i - p_i) \log \frac{\beta Kq_i - p_i}{\beta Kq_i - Kq_i} \quad (6.108)$$

Consequently:

$$\frac{\partial FD1_\beta(p\|Kq)}{\partial K} = \beta \sum_i q_i \log \frac{\beta Kq_i - p_i}{\beta Kq_i - Kq_i} \quad (6.109)$$

To cancel this derivative, there is no explicit solution, an alternative is to use K^*, which leads to the simplified invariant divergence:

$$FD1I_\beta(p\|q) = \sum_i \bar{p}_i \log \frac{\bar{p}_i}{\bar{q}_i} + \sum_i (\beta\bar{q}_i - \bar{p}_i) \log \frac{\beta\bar{q}_i - \bar{p}_i}{(\beta-1)\bar{q}_i} \quad (6.110)$$

The gradient with respect to "q" is expressed as:

$$\begin{aligned}\frac{\partial FD1_\beta I(p\|q)}{\partial q_l} =& \frac{1}{\sum_j q_j}\left[2\frac{\bar{p}_l}{\bar{q}_l} - 2\right. \\ &\left. +\beta \log \frac{\beta\bar{q}_l - \bar{p}_l}{(\beta-1)\bar{q}_l} - \beta \sum_i \bar{q}_i \log \frac{\beta\bar{q}_i - \bar{p}_i}{(\beta-1)\bar{q}_i}\right]\end{aligned} \quad (6.111)$$

6.8 Divergence of P.K.Bathia and S.Singh.

This divergence proposed in [18] is basically a Csiszär divergence, but in this work, there are some mistakes: the basic convex function, represented on the figure (6.16), for $\alpha = 0.5$, is not a standard function, indeed, the authors write it:

$$f(x) = \frac{x \sinh(\alpha \log x)}{\sinh \alpha} \quad (6.112)$$

We can note that the multiplicative factor $\frac{1}{\sinh\alpha}$ is not really helpful. For this function, we do have convexity and $f(1) = 0$, but we don't have $f'(1) = 0$, contrary to what is stated in [18].
If we still use this function to construct a Csiszär divergence, we obtain:

$$SB_\alpha(p\|q) = \sum_i p_i \frac{\sinh\left(\alpha \log \frac{p_i}{q_i}\right)}{\sinh \alpha} \quad (6.113)$$

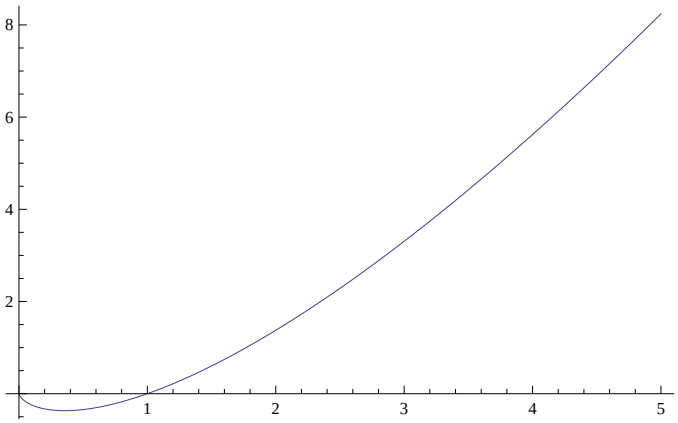

Figure 6.16: Function $f(x) = \frac{x \sinh(\alpha \log x)}{\sinh \alpha}$, $\alpha = 0.5$

The problem with this expression is classical: the gradient with respect to "q" does not cancel for $p_i = q_i \ \forall i$.
To eliminate this kind of problem, it is necessary to introduce the corresponding standard convex function, represented in the figure (6.17), for $\alpha = 0.5$, which is written as follows:

$$f_c(x) = \frac{x \sinh(\alpha \log x) - \alpha x + \alpha}{\sinh \alpha} \tag{6.114}$$

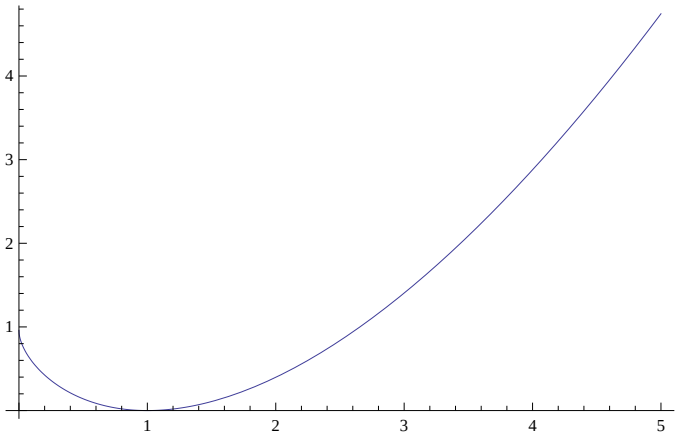

Figure 6.17: Function $f_c(x) = \frac{x \sinh(\alpha \log x) - \alpha x + \alpha}{\sinh \alpha}$, $\alpha = 0.5$

We are thus led to the divergence:

$$SB_{c,\alpha}\left(p\|q\right)=\frac{1}{\sinh\alpha}\left[\sum_{i}p_{i}\sinh\left(\alpha\log\frac{p_{i}}{q_{i}}\right)+\alpha\left(q_{i}-p_{i}\right]\right] \tag{6.115}$$

Whose gradient with respect to "q" is:

$$\frac{\partial SB_{c,\alpha}\left(p\|q\right)}{\partial q_{j}}\approx-\frac{p_{j}}{q_{j}}\cosh\left(\alpha\log\frac{p_{j}}{q_{j}}\right)+1 \tag{6.116}$$

The dual divergence built on the mirror function of (6.114) will be written as follows:

$$SB_{c,\alpha}\left(q\|p\right)=\frac{1}{\sinh\alpha}\left[\sum_{i}q_{i}\sinh\left(\alpha\log\frac{q_{i}}{p_{i}}\right)+\alpha\left(p_{i}-q_{i}\right]\right] \tag{6.117}$$

and its gradient with respect to "q" is:

$$\frac{\partial SB_{c,\alpha}\left(q\|p\right)}{\partial q_{j}}=\frac{1}{\sinh\alpha}\left[\sinh\left(\alpha\log\frac{q_{j}}{p_{j}}\right)+\alpha\cosh\left(\alpha\log\frac{q_{j}}{p_{j}}\right)-\alpha\right] \tag{6.118}$$

6.8.1 Invariance by scale change on "q".

For this divergence given by the expression (6.115), the nominal invariance factor cannot be calculated, so we use the expression K^{*}, which leads to the simplified invariant divergence:

$$SB_{c,\alpha}\left(p\|q\right)I=\frac{1}{\sinh\alpha}\left[\sum_{i}\bar{p}_{i}\sinh\left(\alpha\log\frac{\bar{p}_{i}}{\bar{q}_{i}}\right)\right] \tag{6.119}$$

Its gradient with respect to "q" is:

$$\frac{\partial SB_{c,\alpha}\left(p\|q\right)I}{\partial q_{l}}=\frac{\alpha\coth\alpha}{\sum_{j}q_{j}}\left[\sum_{i}\bar{p}_{i}\log\frac{\bar{p}_{i}}{\bar{q}_{i}}-\frac{\bar{p}_{l}}{\bar{q}_{l}}\log\frac{\bar{p}_{l}}{\bar{q}_{l}}\right] \tag{6.120}$$

chapter 7 - Divergences between means

The major part of this chapter can be found in the book by I.J.Taneja (online) [91] and in the article by Ben-Tal et al. [12]. The construction mode of these divergences is twofold: either one relies on the well-known inequalities between the classical means, or one constructs them in the sense of Csiszär by relying on the properties of the convex functions.
We will make this double aspect more explicit.
In addition, in this chapter, we will present the logarithmic forms of these divergences.
The invariant forms of these divergences will be developed by introducing the invariance factor $K^*(p,q) = \frac{\sum_j p_j}{\sum_j q_j}$ which is not necessarily the nominal invariance factor.
The particular case of M_{AG} for which the nominal invariance factor is explicitly calculated has been presented in the chapter dealing with "Alpha Divergences".
Some basic recalls on means are summarized in Appendix 2.

7.1 Square root (Quadratic) - Arithmetic mean divergence.

This divergence is founded on the inequality: $M_S \geq M_A$.

7.1.1 Unweighted version.

The most classical form of this divergence is constructed in the sense of Csiszär on the basis of the standard convex function:

$$f_c(x) = \sqrt{\frac{x^2+1}{2}} - \frac{x+1}{2} \tag{7.1}$$

We immediately obtain:

$$M_{SA}(p\|q) = \sum_i \sqrt{\frac{p_i^2 + q_i^2}{2}} - \sum_i \frac{p_i + q_i}{2} \tag{7.2}$$

This divergence is symmetrical.
Of course, if one is in a situation where one knows that $\sum_i p_i = \sum_i q_i$, (but not necessarily $= 1$), this divergence becomes simpler, it becomes:

$$M1_{SA}(p\|q) = \sum_i \sqrt{\frac{p_i^2 + q_i^2}{2}} - \sum_i q_i \tag{7.3}$$

That is to say, it is constructed in the sense of Csiszär on the simple convex function:

$$f_1(x) = \sqrt{\frac{x^2+1}{2}} - 1 \tag{7.4}$$

But it can also become:

$$M2_{SA}(p\|q) = \sum_i \sqrt{\frac{p_i^2 + q_i^2}{2}} - \sum_i p_i \tag{7.5}$$

It is then built on the simple convex function:

$$f_2(x) = \sqrt{\frac{x^2+1}{2}} - x \tag{7.6}$$

We can exhibit a family of simple convex functions $f(x)$ that will give simplified divergences. Knowing $f_c(x)$, these functions are deduced from each other by modifying the value of $f'(1)$ in the expression:

$$f(x) = f_c(x) + (x-1) f'(1) \tag{7.7}$$

If we now consider the situation where $\sum_i p_i = \sum_i q_i = 1$ we obtain in all cases an oversimplified form which of course, we are not always allowed to use; this form is written as follows:

$$MS_{SA}(p\|q) = \sum_i \sqrt{\frac{p_i^2 + q_i^2}{2}} - 1 \tag{7.8}$$

It should be observed that this divergence cannot be constructed directly as a Csiszär divergence because the simplification $\sum_i p_i = \sum_i q_i = 1$ cannot be taken into account until the divergence is exhibited.
Finally, it is worth noting that:

$$f_c(x) = \frac{f1(x) + f2(x)}{2} \tag{7.9}$$

This explains the symmetry of the divergence M_{SA}.

7.1.2 Weighted version.

The standard convex function allowing to construct this divergence is shown in the figure (7.1) for $\alpha = 0.5$.

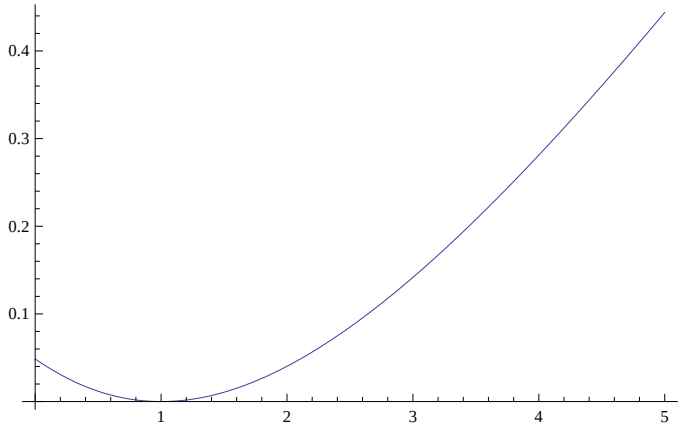

Figure 7.1: Function $f_{c\alpha}(x) = \sqrt{\alpha x^2 + 1 - \alpha} - (\alpha x + 1 - \alpha)$; $\alpha = 0.5$

It is written with $0 < \alpha < 1$:

$$f_{c\alpha}(x) = \sqrt{\alpha x^2 + 1 - \alpha} - (\alpha x + 1 - \alpha) \tag{7.10}$$

This results in:

$$M_{SA\alpha}(p\|q) = \sum_i \sqrt{\alpha p_i^2 + (1-\alpha) q_i^2} - \sum_i [\alpha p_i + (1-\alpha) q_i] \tag{7.11}$$

This divergence is not symmetrical because of the weighting (unless $\alpha = 1/2$).
Its gradient with respect to "q" is written:

$$\frac{\partial M_{SA\alpha}}{\partial q_j} = (1-\alpha) \left[\frac{q_j}{\sqrt{\alpha p_j^2 + (1-\alpha) q_j^2}} - 1 \right] \tag{7.12}$$

Scale invariant version with respect to "q".

For this divergence, the nominal invariance factor is not obtained explicitly, so we use as the invariance factor the expression $K^*(p,q)$; thus we obtain

the expression for the corresponding simplified invariant divergence, which is written as:

$$M_{SA\alpha}I\left(p\|q\right)=\sum_{j}p_{j}\left\{\sum_{i}\sqrt{\alpha\bar{p}_{i}^{2}+\left(1-\alpha\right)\bar{q}_{i}^{2}}-\sum_{i}\left[\alpha\bar{p}_{i}+\left(1-\alpha\right)\bar{q}_{i}\right]\right\} \tag{7.13}$$

After simplification and deleting the multiplying factor, we will have:

$$M_{SA\alpha}I\left(p\|q\right)=\sum_{i}\sqrt{\alpha\bar{p}_{i}^{2}+\left(1-\alpha\right)\bar{q}_{i}^{2}}-1 \tag{7.14}$$

The gradient with respect to "q" is given by:

$$\frac{\partial M_{SA\alpha}I\left(p\|q\right)}{\partial q_{l}}=\frac{1-\alpha}{\sum_{j}q_{j}}\left\{\left[\alpha\bar{p}_{l}^{2}+\left(1-\alpha\right)\bar{q}_{l}^{2}\right]^{-\frac{1}{2}}\bar{q}_{l}-\sum_{i}\left[\alpha\bar{p}_{i}^{2}+\left(1-\alpha\right)\bar{q}_{i}^{2}\right]^{-\frac{1}{2}}\bar{q}_{i}^{2}\right\} \tag{7.15}$$

7.1.3 Logarithmic form.

Since the divergence (7.11) appears as the difference of 2 positive quantities, one can apply on each term an increasing function without changing the sign of the expression, for example the function "Generalized Logarithm" and at the limit the "Logarithm", which allows to write:

$$LM_{SA\alpha}=\log\left[\sum_{i}\sqrt{\alpha p_{i}^{2}+\left(1-\alpha\right)q_{i}^{2}}\right]-\log\left[\sum_{i}\alpha p_{i}+\left(1-\alpha\right)q_{i}\right] \tag{7.16}$$

With such a divergence, nothing is clear about the convexity, so nothing is guaranteed; however, we can calculate the gradient with respect to "q":

$$\begin{aligned}\frac{\partial LM_{SA\alpha}}{\partial q_{j}}=&\frac{\left(1-\alpha\right)}{\sum_{i}\sqrt{\alpha p_{i}^{2}+\left(1-\alpha\right)q_{i}^{2}}}\frac{q_{j}}{\sqrt{\alpha p_{j}^{2}+\left(1-\alpha\right)q_{j}^{2}}}\\&-\frac{\left(1-\alpha\right)}{\sum_{i}\alpha p_{i}+\left(1-\alpha\right)q_{i}}\end{aligned} \tag{7.17}$$

This expression is zero if $p_i = q_i \ \forall i$.

Logarithmic form invariant with respect to "q".

To obtain such a Logarithmic form we apply the function "Log" separately on each term of (7.14) and we obtain:

$$LM_{SA\alpha}I\left(p\|q\right)=\log\sum_{i}\sqrt{\alpha\bar{p}_{i}^{2}+\left(1-\alpha\right)\bar{q}^{2i}} \tag{7.18}$$

which gradient with respect to "q" is:

$$\frac{\partial LM_{SA\alpha}I\left(p\|q\right)}{\partial q_l}=\frac{1}{\sum_i\sqrt{\alpha\bar{p}_i^2+\left(1-\alpha\right)\bar{q}_i^2}}\frac{\partial M_{SA\alpha}I\left(p\|q\right)}{\partial q_l} \tag{7.19}$$

7.2 Square root (Quadratic)- Géométric mean divergence.

This divergence is based on the inequality: $M_S\geq M_G$.

7.2.1 Unweighted version.

It is a Csiszär divergence built on the standard convex function:

$$f_c(x)=\sqrt{\frac{x^2+1}{2}}-\sqrt{x} \tag{7.20}$$

This results in the divergence:

$$M_{SG}\left(p\|q\right)=\sum_i\sqrt{\frac{p_i^2+q_i^2}{2}}-\sum_i\sqrt{p_iq_i} \tag{7.21}$$

7.2.2 Weighted version.

With $0\leq\alpha\leq1$, this divergence is based on the standard convex function:

$$f_c\left(x\right)=\sqrt{\alpha x^2+\left(1-\alpha\right)}-x^\alpha \tag{7.22}$$

shown in the figure (7.2) for $\alpha=0.5$.
The corresponding divergence will be written:

$$M_{SG\alpha}\left(p\|q\right)=\sum_i\sqrt{\alpha p_i^2+\left(1-\alpha\right)q_i^2}-\sum_i p_i^\alpha q_i^{1-\alpha} \tag{7.23}$$

we deduce the gradient with respect to "q":

$$\frac{\partial M_{SG\alpha}}{\partial q_j}=\left(1-\alpha\right)\left[\frac{q_j}{\sqrt{\alpha p_j^2+\left(1-\alpha\right)q_j^2}}-p_j^\alpha q_j^{-\alpha}\right] \tag{7.24}$$

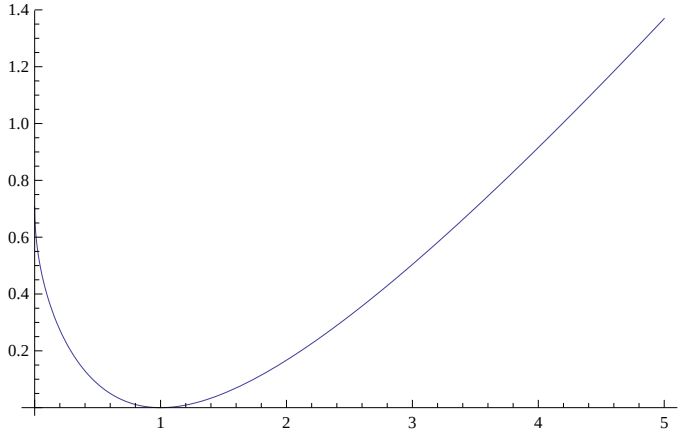

Figure 7.2: Function $f_c(x) = \sqrt{\alpha x^2 + (1-\alpha)} - x^\alpha$; $\alpha = 0.5$

Scale invariant form with respect to "q".

For this divergence, the nominal invariance factor cannot be obtained explicitly, so we use the expression $K^*(p,q)$ as the invariance factor; thus, after simplification, we obtain the expression of the corresponding invariant divergence:

$$M_{SG\alpha}I(p\|q) = \sum_i \sqrt{\alpha\bar{p}_i^2 + (1-\alpha)\,\bar{q}_i^2} - \sum_i \bar{p}_i^\alpha \bar{q}_i^{1-\alpha} \tag{7.25}$$

which has a gradient with respect to "q" given by:

$$\frac{\partial M_{SG\alpha}I(p\|q)}{\partial q_l} = \frac{1-\alpha}{\sum_j q_j}\left\{\frac{\bar{q}_l}{\sqrt{\alpha\bar{p}_l^2 + (1-\alpha)\,\bar{q}_l^2}} - \sum_i \frac{\bar{q}_i^2}{\sqrt{\alpha\bar{p}_i^2 + (1-\alpha)\,\bar{q}_i^2}} - \bar{p}_l^\alpha \bar{q}_l^{-\alpha} + \sum_i \bar{p}_i^\alpha \bar{q}_i^{1-\alpha}\right\} \tag{7.26}$$

7.2.3 Logarithmic form.

After applying an increasing function (Log), on each of the terms of the difference appearing in (7.23), the corresponding divergence will be written:

$$LM_{SG\alpha}(p\|q) = \log\sum_i \sqrt{\alpha p_i^2 + (1-\alpha)\,q_i^2} - \log\sum_i p_i^\alpha q_i^{1-\alpha} \tag{7.27}$$

The gradient with respect to "q" is given by:

$$\frac{\partial LM_{SG\alpha}}{\partial q_j} = (1-\alpha)\left[\frac{1}{\sum_i \sqrt{\alpha p_i^2 + (1-\alpha)\, q_i^2}} \frac{q_j}{\sqrt{\alpha p_j^2 + (1-\alpha)\, q_j^2}} - \frac{p_j^{\alpha} q_j^{-\alpha}}{\sum_i p_i^{\alpha} q_i^{1-\alpha}}\right] \tag{7.28}$$

This expression is zero if $p_i = q_i \ \forall i$.

Logarithmic form - Invariance with respect to "q".

The logarithmic version invariant with respect to "q" is obtained by applying the function "Log" separately to each term of (7.25) and one obtains:

$$LM_{SG\alpha} I\left(p \| q\right) = \log \sum_i \sqrt{\alpha \bar{p}_i^2 + (1-\alpha)\, \bar{q}_i^2} - \log \sum_i \bar{p}_i^{\alpha} \bar{q}_i^{1-\alpha} \tag{7.29}$$

whose gradient is written as:

$$\frac{\partial LM_{SG\alpha} I\left(p \| q\right)}{\partial q_l} = \frac{1-\alpha}{\sum_j q_j} \left\{\frac{1}{\sum_i \sqrt{\alpha \bar{p}_i^2 + (1-\alpha)\, \bar{q}_i^2}} \left[\frac{\bar{q}_l}{\sqrt{\alpha \bar{p}_l^2 + (1-\alpha)\, \bar{q}_l^2}} - \sum_i \frac{\bar{q}_i^2}{\sqrt{\alpha \bar{p}_i^2 + (1-\alpha)\, \bar{q}_i^2}}\right] + 1 - \frac{\bar{p}_l^{\alpha} \bar{q}_l^{-\alpha}}{\sum_i \bar{p}_i^{\alpha} \bar{q}_i^{1-\alpha}}\right\} \tag{7.30}$$

7.3 Square root (Quadratic) - Harmonic mean divergence.

This divergence is based on the inequality: $M_S \geq M_H$.

7.3.1 Unweighted version.

It is a Csiszär divergence based on the standard convex function:

$$f_c\left(x\right) = \sqrt{\frac{x^2+1}{2}} - \frac{2x}{1+x} \tag{7.31}$$

Which leads to the divergence:

$$M_{SH}(p\|q) = \sum_i \sqrt{\frac{p_i^2 + q_i^2}{2}} - \sum_i \frac{2p_i q_i}{p_i + q_i} \tag{7.32}$$

7.3.2 Weighted version.

With $0 \leq \alpha \leq 1$, it is built in the Csiszär sense using the standard convex function :

$$f_{c\alpha}(x) = \sqrt{\alpha x^2 + (1-\alpha)} - \frac{x}{(1-\alpha)x + \alpha} \tag{7.33}$$

It is shown in the figure (7.3) for $\alpha = 0.5$.

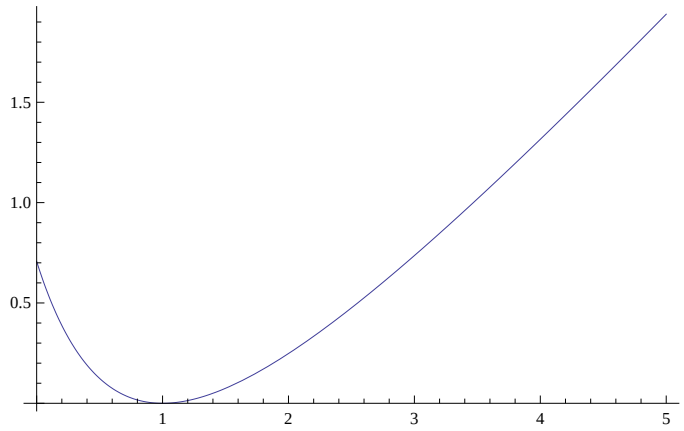

Figure 7.3: Function $f_{c\alpha}(x) = \sqrt{\alpha x^2 + (1-\alpha)} - \frac{x}{(1-\alpha)x+\alpha}$; $\alpha = 0.5$

This results in the divergence:

$$M_{SH\alpha}(p\|q) = \sum_i \sqrt{\alpha p_i^2 + (1-\alpha) q_i^2} - \sum_i \frac{p_i q_i}{(1-\alpha) p_i + \alpha q_i} \tag{7.34}$$

The gradient with respect to "q" is given by:

$$\frac{\partial M_{SH\alpha}}{\partial q_j} = (1-\alpha) \left\{ \frac{q_j}{\sqrt{\alpha p_j^2 + (1-\alpha) q_j^2}} - \frac{p_j^2}{\left[(1-\alpha) p_j + \alpha q_j\right]^2} \right\} \tag{7.35}$$

This expression is zero if $p_i = q_i \ \forall i$.

Scale invariant version with respect to "q".

For this divergence, the nominal invariance factor is not obtained explicitly, so we use as the invariance factor the expression $K^*(p,q)$; thus, after simplification, we obtain the form of the invariant corresponding divergence:

$$M_{SH\alpha}I\left(p\|q\right)=\sum_i\sqrt{\alpha\bar{p}_i^2+(1-\alpha)\bar{q}_i^2}-\sum_i\frac{\bar{p}_i\bar{q}_i}{(1-\alpha)\,\bar{p}_i+\alpha\bar{q}_i} \tag{7.36}$$

The gradient with respect to "q" is given by:

$$\begin{aligned}\frac{\partial M_{SH\alpha}I\left(p\|q\right)}{\partial q_l}=&\frac{(1-\alpha)}{\sum_j q_j}\\&\left\{\frac{\bar{q}_l}{\sqrt{\alpha\bar{p}_l^2+(1-\alpha)\bar{q}_l^2}}-\sum_i\frac{\bar{q}_i^2}{\sqrt{\alpha\bar{p}_i^2+(1-\alpha)\bar{q}_i^2}}\right.\\&\left.-\frac{\bar{p}_l^2}{\left[(1-\alpha)\,\bar{p}_l+\alpha\bar{q}_l\right]^2}+\sum_i\frac{\bar{p}_i^2\bar{q}_i}{\left[(1-\alpha)\,\bar{p}_i+\alpha\bar{q}_i\right]^2}\right\}\end{aligned} \tag{7.37}$$

7.3.3 Logarithmic version.

We can write directly from (7.34):

$$LM_{SH}\left(p\|q\right)=\log\sum_i\sqrt{\alpha p_i^2+(1-\alpha)\,q_i^2}-\log\sum_i\frac{p_iq_i}{(1-\alpha)\,p_i+\alpha q_i} \tag{7.38}$$

Its gradient with respect to "q" is written as follows:

$$\begin{aligned}\frac{\partial LM_{SH\alpha}}{\partial q_j}=(1-\alpha)&\left\{\frac{1}{\sum_i\sqrt{\alpha p_i^2+(1-\alpha)\,q_i^2}}\frac{q_j}{\sqrt{\alpha p_j^2+(1-\alpha)\,q_j^2}}\right.\\&\left.-\frac{1}{\sum_i\frac{p_iq_i}{(1-\alpha)p_i+\alpha q_i}}\frac{p_j^2}{\left[(1-\alpha)\,p_j+\alpha q_j\right]^2}\right\}\end{aligned} \tag{7.39}$$

This expression is zero if $p_i=q_i\ \forall i$.

Logarithmic form - Invariance with respect to "q".

The logarithmic invariant version is obtained by applying the function "Log" separately to each term of (7.36) and one obtains:

$$LM_{SH\alpha}I\left(p\|q\right)=\log\sum_i\sqrt{\alpha\bar{p}_i^2+(1-\alpha)\bar{q}_i^2}-\log\sum_i\frac{\bar{p}_i\bar{q}_i}{(1-\alpha)\,\bar{p}_i+\alpha\bar{q}_i} \tag{7.40}$$

Its gradient with respect to "q" is given by:

$$\frac{\partial LM_{SH\alpha}I\left(p\|q\right)}{\partial q_l} = \frac{(1-\alpha)}{\sum_j q_j}\left\{\frac{1}{\sum_i \sqrt{\alpha\bar{p}_i^2+(1-\alpha)\bar{q}_i^2}}\right. \left\{\frac{\bar{q}_l}{\sqrt{\alpha\bar{p}_l^2+(1-\alpha)\bar{q}_l^2}} - \sum_i \frac{\bar{q}_i^2}{\sqrt{\alpha\bar{p}_i^2+(1-\alpha)\bar{q}_i^2}}\right\} - \frac{1}{\sum_i \frac{\bar{p}_i\bar{q}_i}{(1-\alpha)\bar{p}_i+\alpha\bar{q}_i}}\left.\left\{\frac{\bar{p}_l^2}{\left[(1-\alpha)\,\bar{p}_l+\alpha\bar{q}_l\right]^2} + \sum_i \frac{\bar{p}_i^2\bar{q}_i}{\left[(1-\alpha)\,\bar{p}_i+\alpha\bar{q}_i\right]^2}\right\}\right\} \tag{7.41}$$

7.4 Arithmetic - Geometric mean divergence

. This divergence is at the basis of the "Alpha divergences" of Amari [2]; here we develop it from an elementary version. An extended version has been considered in the chapter dealing specifically with "Alpha divergences". This divergence is based on the inequality: $M_A \geq M_G$.

7.4.1 Basis unweighted version.

In the simplest version, it is a Csiszär divergence built on the standard convex function:

$$f_c\left(x\right) = \frac{1}{2}\left(\sqrt{x}-1\right)^2 \tag{7.42}$$

We easily obtain:

$$M_{AG}\left(p\|q\right) = \frac{1}{2}\sum_i \left(\sqrt{p_i}-\sqrt{q_i}\right)^2 \tag{7.43}$$

This symmetrical divergence is also known as Hellinger's divergence [13], but it can also be written:

$$M_{AG}\left(p\|q\right) = \sum_i \frac{p_i+q_i}{2} - \sum_i \sqrt{p_i q_i} \tag{7.44}$$

So we better understand the name Arithmetic-Geometric mean divergence. The simplification $\sum_i p_i = \sum_i q_i$ allows us to obtain 2 simplified divergences:

$$MS1_{AG}\left(p\|q\right) = \sum_i q_i - \sum_i \sqrt{p_i q_i} \tag{7.45}$$

Which is deduced in the sense of Csiszär from the simple convex function: $f_1(x) = 1 - \sqrt{x}$, and:

$$MS2_{AG}(p\|q) = \sum_i p_i - \sum_i \sqrt{p_i q_i} \tag{7.46}$$

deduced in the sense of Csiszär from the simple convex function: $f_2(x) = x - \sqrt{x}$.
The supplementary simplification $\sum_i p_i = \sum_i q_i = 1$, leads to:

$$MS_{AG}(p\|q) = 1 - \sum_i \sqrt{p_i q_i} \tag{7.47}$$

This expression is not a divergence of Csiszär; to obtain it, one must first construct a Csiszär divergence by using one of the simple convex functions $f_1(x) = 1 - \sqrt{x}$ or $f_2(x) = x - \sqrt{x}$ (functions mirroring each other) deduced from $f_c(x)$, then introduce "a posteriori" the simplification $\sum_i p_i = 1$ or $\sum_i q_i = 1$.

7.4.2 Weighted version.

If one does not introduce a multiplicative factor whose sign depends on the weighting factor $0 \leq \alpha \leq 1$, the basic standard convex basic function will be written:

$$f_{c\alpha}(x) = \alpha x + (1-\alpha) - x^{\alpha} \tag{7.48}$$

It is represented in the figure (7.4) for $\alpha = 0.5$.
The corresponding divergence is written:

$$M_{AG\alpha}(p\|q) = \sum_i \alpha p_i + (1-\alpha) q_i - \sum_i p_i^{\alpha} q_i^{1-\alpha} \tag{7.49}$$

To broaden the range of validity of this divergence to all positive values of "α", we will use:

$$M_{AG\alpha}(p\|q) = \frac{1}{1-\alpha}\left[\sum_i \alpha p_i + (1-\alpha) q_i - \sum_i p_i^{\alpha} q_i^{1-\alpha}\right] \tag{7.50}$$

This is the Havrda-Charvat divergence [42], and this is exactly what is written for the "α" divergence.
There's no point in commenting on these things that are developed elsewhere.
The gradient with respect to "q" is written:

$$\frac{\partial M_{AG\alpha}}{\partial q_j} = 1 - \left(\frac{p_j}{q_j}\right)^{\alpha} \tag{7.51}$$

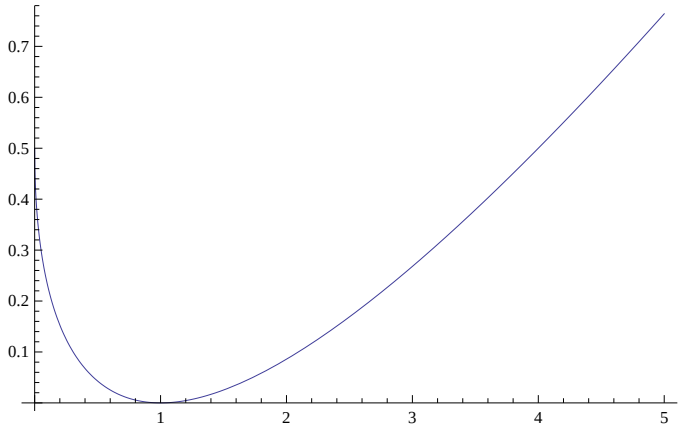

Figure 7.4: Function $f_{c\alpha}(x) = \alpha x + (1-\alpha) - x^{\alpha}$, $\alpha = 0.5$

Invariant form with respect to "q".

For such a divergence, the nominal invariance factor is obtained explicitly, the corresponding invariant form has been developed in the chapter dealing with "Alpha Divergence".
For the sake of homogeneity, we present here the case of an invariance factor of the form $K^*(p,q)$; we thus obtain the following expression for the corresponding simplified invariant divergence:

$$M_{AG\alpha}I(p\|q) = \frac{1}{1-\alpha}\left[\sum_i \left[\alpha\bar{p}_i + (1-\alpha)\,\bar{q}_i\right] - \sum_i \bar{p}_i^{\alpha}\bar{q}_i^{1-\alpha}\right] \tag{7.52}$$

Which simplifies into:

$$M_{AG\alpha}I(p\|q) = \frac{1}{1-\alpha}\left[1 - \sum_i \bar{p}_i^{\alpha}\bar{q}_i^{1-\alpha}\right] \tag{7.53}$$

The gradient with respect to "q" is written:

$$\frac{\partial M_{AG\alpha}I(p\|q)}{\partial q_l} = \frac{1}{\sum_j q_j}\left(\bar{p}_l^{\alpha}\bar{q}_l^{-\alpha} - \sum_i \bar{p}_i^{\alpha}\bar{q}_i^{1-\alpha}\right) \tag{7.54}$$

7.4.3 Logarithmic form.

The logarithmic form of the divergence (7.50) can be written as follows:

$$LM_{AG\alpha}(p\|q) = \frac{1}{\alpha-1}\left[\log\sum_i p_i^{\alpha}q_i^{1-\alpha} - \log\sum_i \alpha p_i + (1-\alpha)\,q_i\right] \tag{7.55}$$

The gradient with respect to "q" is written:

$$\frac{\partial LM_{AG\alpha}}{\partial q_j} = \frac{1}{\sum_i \alpha p_i + (1-\alpha)\, q_i} - \frac{p_j^{\alpha} q_j^{-\alpha}}{\sum_i p_i^{\alpha} q_i^{1-\alpha}} \tag{7.56}$$

If we consider that the data fields involved are probability densities, that is to say, if $\sum_i p_i = \sum_i q_i = 1$, we obtain from (7.55) a simplified expression which is the classical Renyi divergence [82]:

$$R_{\alpha}\left(p \| q\right) = \frac{1}{\alpha - 1}\left[\log \sum_i p_i^{\alpha} q_i^{1-\alpha}\right] \tag{7.57}$$

Logarithmic form - Invariance with respect to "q".

The invariant logarithmic version is obtained by applying the function "Log" separately to each term of (7.53) and we have:

$$LM_{AG\alpha} I\left(p \| q\right) = \frac{1}{\alpha - 1} \log \sum_i \bar{p}_i^{\alpha} \bar{q}_i^{1-\alpha} \tag{7.58}$$

His gradient with respect to q is expressed as:

$$\frac{\partial LM_{AG\alpha} I\left(p \| q\right)}{\partial q_l} = \frac{1}{\sum_j q_j}\left(1 - \frac{\bar{p}_l^{\alpha} \bar{q}_l^{-\alpha}}{\sum_i \bar{p}_i^{\alpha} \bar{q}_i^{1-\alpha}}\right) \tag{7.59}$$

7.5 Arithmetic - Harmonic mean divergence

This divergence is founded on the inequality: $M_A \geq M_H$.

7.5.1 Unweighted version.

This divergence is constructed in the sense of Csiszär on the standard convex function:

$$f_c\left(x\right) = \frac{1}{2}\frac{\left(x-1\right)^2}{x+1} \tag{7.60}$$

The resulting divergence is written:

$$M_{AH}\left(p \| q\right) = \frac{1}{2}\sum_i \frac{\left(p_i - q_i\right)^2}{p_i + q_i} \tag{7.61}$$

or, more clearly:

$$M_{AH}(p\|q) = \left[\sum_i \frac{p_i + q_i}{2} - \sum_i \frac{2p_i q_i}{p_i + q_i}\right] \tag{7.62}$$

If we consider data fields with a sum explicitly equal to 1, we have the simplified form:

$$MS_{AH}(p\|q) = \left[1 - \sum_i \frac{2p_i q_i}{p_i + q_i}\right] \tag{7.63}$$

7.5.2 Weighted version.

With $0 \leq \alpha \leq 1$, the standard convex function shown in the figure (7.5) for $\alpha = 0.5$, is written as follows:

$$f_{c\alpha}(x) = \alpha x + (1-\alpha) - \frac{x}{(1-\alpha)x + \alpha} \tag{7.64}$$

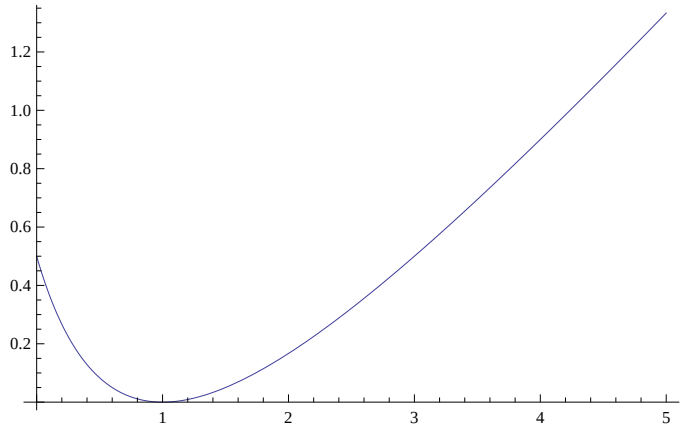

Figure 7.5: Function $f_{c\alpha}(x) = \alpha x + (1-\alpha) - \frac{x}{(1-\alpha)x+\alpha}$, $\alpha = 0.5$

It leads to the divergence:

$$M_{AH\alpha}(p\|q) = \sum_i \alpha p_i + (1-\alpha) q_i - \sum_i \frac{p_i q_i}{\alpha q_i + (1-\alpha) p_i} \tag{7.65}$$

The gradient with respect to "q" is expressed as:

$$\frac{\partial M_{AH\alpha}}{\partial q_j} = (1-\alpha)\left\{1 - \frac{p_j^2}{[\alpha q_j + (1-\alpha) p_j]^2}\right\} \tag{7.66}$$

Scale invariant version with respect to "q".

For this divergence, the nominal invariance factor is not obtained explicitly, so as invariance factor we use the expression $K^{*}(p,q) = \frac{\sum_j p_j}{\sum_j q_j}$; we then obtain the corresponding simplified invariant divergence:

$$M_{AH\alpha}I\left(p\|q\right) = \sum_i \left[\alpha\bar{p}_i + (1-\alpha)\,\bar{q}_i\right] - \sum_i \frac{\bar{p}_i\bar{q}_i}{(1-\alpha)\,\bar{p}_i + \alpha\bar{q}_i} \tag{7.67}$$

Or in a simplified form:

$$M_{AH\alpha}I\left(p\|q\right) = 1 - \sum_i \frac{\bar{p}_i\bar{q}_i}{(1-\alpha)\,\bar{p}_i + \alpha\bar{q}_i} \tag{7.68}$$

The gradient with respect to "q" is given by:

$$\frac{\partial M_{AH\alpha}I\left(p\|q\right)}{\partial q_l} = \frac{(\alpha-1)}{\sum_j q_j}\left\{\frac{\bar{p}_l^2}{\left[(1-\alpha)\,\bar{p}_l + \alpha\bar{q}_l\right]^2} - \sum_i \frac{\bar{p}_i^2\bar{q}_i}{\left[(1-\alpha)\,\bar{p}_i + \alpha\bar{q}_i\right]^2}\right\} \tag{7.69}$$

7.5.3 Logarithmic form.

It is written immediately from the weighted version (7.65):

$$LM_{AH\alpha}\left(p\|q\right) = \log\sum_i \alpha p_i + (1-\alpha)\,q_i - \log\sum_i \frac{p_i q_i}{\alpha q_i + (1-\alpha)\,p_i} \tag{7.70}$$

The gradient with respect to "q" is written as:

$$\frac{\partial LM_{AH\alpha}}{\partial q_j} = (1-\alpha)\left\{\frac{1}{\sum_i \alpha p_i + (1-\alpha)\,q_i} - \frac{1}{\sum_i \frac{p_i q_i}{\alpha q_i + (1-\alpha)p_i}}\frac{p_j^2}{\left[\alpha q_j + (1-\alpha)\,p_j\right]^2}\right\} \tag{7.71}$$

Logarithmc form - Invariance with respect to "q".

The logarithmic invariant version is obtained by applying the "Log" function separately to each term of (7.68) and we obtain:

$$LM_{AH\alpha}I\left(p\|q\right) = -\log\sum_i \frac{\bar{p}_i\bar{q}_i}{(1-\alpha)\,\bar{p}_i + \alpha\bar{q}_i} \tag{7.72}$$

Whose gradient with respect to "q" is given by:

$$\frac{\partial M_{AH\alpha} I\left(p\|q\right)}{\partial q_l} = \frac{\left(\alpha-1\right)}{\sum_i \frac{\bar{p}_i \bar{q}_i}{(1-\alpha)\bar{p}_i + \alpha\bar{q}_i} \sum_j q_j} \left[\frac{\bar{p}_l^2}{\left[(1-\alpha)\,\bar{p}_l + \alpha\bar{q}_l\right]^2} - \sum_i \frac{\bar{p}_i^2 \bar{q}_i}{\left[(1-\alpha)\,\bar{p}_i + \alpha\bar{q}_i\right]^2} \right] \tag{7.73}$$

chapter 8 - Divergences between mixed fields.

In this chapter, the main idea is to analyze the difference between 2 data fields "p" and "q" by expressing the difference in a broad sense between "p" or "q" and the weighted sum of "p" and "q"; $0 \leq \alpha \leq 1$ will be the corresponding weighting factor. Thus, whatever the basic divergence, one can go from a zero divergence to a divergence between the elementary fields by varying the parameter "α".
The gap between "p" or "q" and their weighted sum is expressed by a Kullback-Leibler divergence [57] or, to generalize, by a Havrda-Charvat divergence [42], but nothing prohibits the use of other basic divergences.
Moreover, to broaden the picture, we will envisage taking into account in the expression of the divergences, the order of the arguments, which is understandable inasmuch as the divergences used are not generally symmetrical.
This argumentation describes a particular constructive process for these divergences, however, in each case examined, we will always come back to Csiszär's constructive process and the associated convex function.
Of course, in the literature, these studies have been proposed mainly to quantify the gap between two probability densities; here we try to leave this restrictive context and extend the applications to more general data fields.

8.1 F divergence (Jensen-Shannon relative divergence).

8.1.1 Direct form.

This divergence is reported by Taneja [91].
The constructive process can be described by:

$$F(p\|q) = KL(p\|\alpha p + (1-\alpha)q) \tag{8.1}$$

If we use as expression of K.L.'s divergence, the general form usable for any (non-normalized) variables, we obtain:

$$F(p\|q) = \sum_i \left[p_i \log \frac{p_i}{\alpha p_i + (1-\alpha) q_i} + (1-\alpha)(q_i - p_i) \right] \tag{8.2}$$

With $0 \leq \alpha \leq 1$, this divergence is constructed in the sense of Csiszär on the standard convex function shown in the figure (8.1) for $\alpha = 0.5$, it is written as:

$$f_c(x) = x \log \frac{x}{\alpha x + (1-\alpha)} + (1-\alpha)(1-x) \tag{8.3}$$

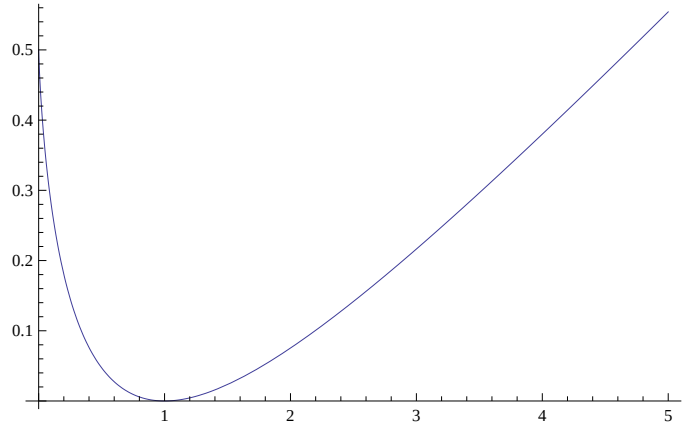

Figure 8.1: Function $f_c(x) = x \log \frac{x}{\alpha x + (1-\alpha)} + (1-\alpha)(1-x)$, $\alpha = 0.5$

The gradient with respect to "q" is expressed as:

$$\frac{\partial F(p\|q)}{\partial q_j} = (1-\alpha)\left(1 - \frac{p_j}{\alpha p_j + (1-\alpha) q_j}\right) \tag{8.4}$$

In the case where we can consider that $\sum_i p_i = \sum_i q_i$, not necessarily equal to 1, we have a simplified form, which is written as follows:

$$FS\left(p\|q\right) = \sum_i p_i \log \frac{p_i}{\alpha p_i + \left(1-\alpha\right) q_i} \tag{8.5}$$

With $\alpha = 1/2$, this is the K divergence of Lin [68].
With $0 \leq \alpha \leq 1$, it is a Csiszär divergence built on the simple convex function, shown in the figure (8.2) for $\alpha = 0.5$, which is expressed as follows:

$$f\left(x\right) = x \log \frac{x}{\alpha x + \left(1-\alpha\right)} \tag{8.6}$$

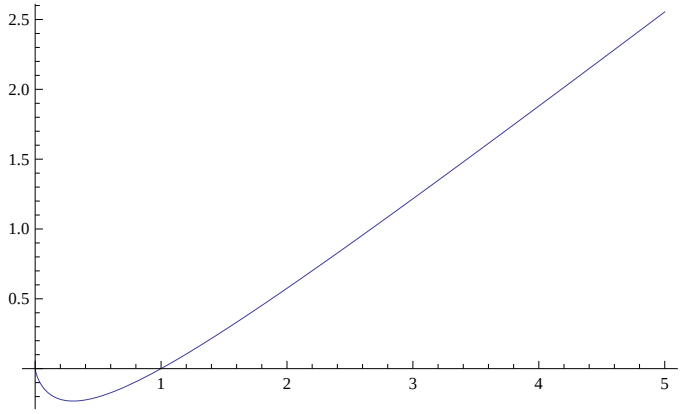

Figure 8.2: Function $f\left(x\right) = x \log \frac{x}{\alpha x + (1-\alpha)}$, $\alpha = 0.5$

8.1.2 Dual form.

The constructive process can be described by:

$$F\left(q\|p\right) = KL\left(q\|\alpha q + \left(1-\alpha\right) p\right) \tag{8.7}$$

If we use as expression of the K.L. divergence, the general form usable for non-normalized variables, we obtain:

$$F\left(q\|p\right) = \sum_i \left[q_i \log \frac{q_i}{\alpha q_i + \left(1-\alpha\right) p_i} + \left(1-\alpha\right)\left(p_i - q_i\right) \right] \tag{8.8}$$

With $0 \leq \alpha \leq 1$, it is a Csiszär divergence built on the standard convex function, represented in the figure (8.3) for $\alpha = 0.5$, which is expressed as

follows:

$$\breve{f}_c(x) = \log \frac{1}{\alpha + (1-\alpha)\, x} + (1-\alpha)\,(x-1) \tag{8.9}$$

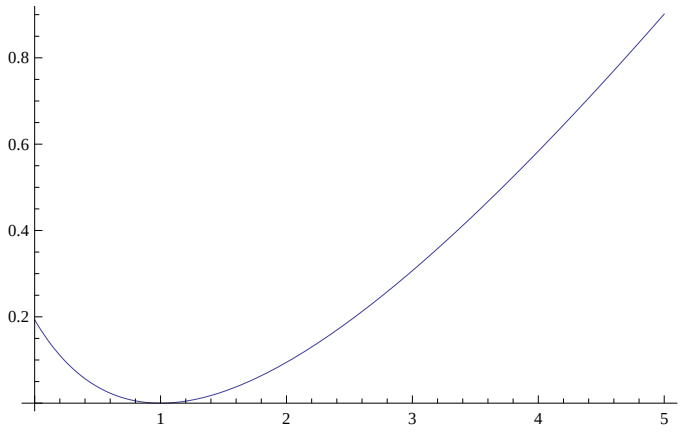

Figure 8.3: Function $\breve{f}_c(x) = \log \frac{1}{\alpha+(1-\alpha)x} + (1-\alpha)\,(x-1)$, $\alpha = 0.5$

Its gradient with respect to "q" is written as follows:

$$\frac{\partial F(q\|p)}{\partial q_j} = \log \frac{q_j}{\alpha q_j + (1-\alpha)\, p_j} - \frac{\alpha q_j}{\alpha q_j + (1-\alpha)\, p_j} + \alpha \tag{8.10}$$

In the case where we can consider that $\sum_i p_i = \sum_i q_i$, not necessarily equal to 1, we have a simplified form which is written as:

$$FS(q\|p) = \sum_i q_i \log \frac{q_i}{\alpha q_i + (1-\alpha)\, p_i} \tag{8.11}$$

It is a Csiszär divergence built on the simple convex function represented on the figure (8.4) for $\alpha = 0.5$, written as follows:

$$\breve{f}(x) = \log \frac{1}{\alpha + (1-\alpha)\, x} \tag{8.12}$$

8.1.3 Symmetrical form.

It is written in the sense of Jeffreys [52] as the (half) sum of the two previous ones:

$$F(p,q) = F(p\|q) + F(q\|p) \tag{8.13}$$

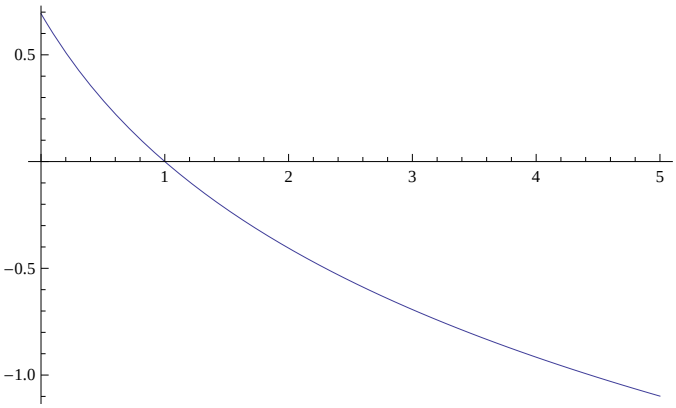

Figure 8.4: Function $\breve{f}(x) = \log \frac{1}{\alpha+(1-\alpha)x}$, $\alpha = 0.5$

That is:

$$F(p,q) = \sum_i p_i \log \frac{p_i}{\alpha p_i + (1-\alpha) q_i} + \sum_i q_i \log \frac{q_i}{\alpha q_i + (1-\alpha) p_i} \tag{8.14}$$

Whose gradient with respect to "q" is expressed as:

$$\begin{aligned} \frac{\partial F(q,p)}{\partial q_j} =& 1 - (1-\alpha) \frac{p_j}{\alpha p_j + (1-\alpha) q_j} + \log \frac{q_j}{\alpha q_j + (1-\alpha) p_j} \\ & - \frac{\alpha q_j}{\alpha q_j + (1-\alpha) p_j} \end{aligned} \tag{8.15}$$

8.1.4 Invariance by change of scale on "q".

The classical method of computing the $K(p,q)$ invariance factor corresponding to the divergence $F(p\|q)$ (equation (8.2)) involves solving in K the equation:

$$\frac{\partial F(p\|Kq)}{\partial K} = 0 \tag{8.16}$$

That is, to solve with respect to K, the equation:

$$\sum_i q_i \left[1 - \frac{p_i}{\alpha p_i + (1-\alpha) K q_i} \right] = 0 \tag{8.17}$$

There's no explicit solution to this problem.
So we use as an invariance factor the expression $K^*(p,q) = \frac{\sum_j p_j}{\sum_j q_j}$.

In these conditions, the invariant divergence is written after simplifications with the normalized variables $\bar{p}$ and $\bar{q}$:

$$FI\left(p\|q\right)=\sum_{i}\bar{p}_{i}\log\frac{\bar{p}_{i}}{\alpha\bar{p}_{i}+\left(1-\alpha\right)\bar{q}_{i}} \tag{8.18}$$

Hence the expression of the gradient:

$$\frac{\partial FI\left(p\|q\right)}{\partial q_{l}}=\frac{1-\alpha}{\sum_{j}q_{j}}\left[\sum_{i}\frac{\bar{p}_{i}\bar{q}_{i}}{\alpha\bar{p}_{i}+\left(1-\alpha\right)\bar{q}_{i}}-\frac{\bar{p}_{l}}{\alpha\bar{p}_{l}+\left(1-\alpha\right)\bar{q}_{l}}\right] \tag{8.19}$$

8.2 G divergence (Arithmetic-Geometric relative divergence).

These divergences are discussed by Taneja [91], they are deduced from the previous ones by inverting the order of the compared fields; only the weighted forms are presented here.

8.2.1 Direct form.

The constructive process is of the following form:

$$G\left(p\|q\right)=KL\left(\alpha p+\left(1-\alpha\right)q\|p\right) \tag{8.20}$$

If we use the general form of the Kullback-Leibler divergence that can be used for non-normalized fields, we obtain:

$$\begin{aligned}G\left(p\|q\right)=&\sum_{i}\left[\alpha p_{i}+\left(1-\alpha\right)q_{i}\right]\log\frac{\alpha p_{i}+\left(1-\alpha\right)q_{i}}{p_{i}}\\&+\left(1-\alpha\right)\left(p_{i}-q_{i}\right)\end{aligned} \tag{8.21}$$

It is a Csiszär's divergence built on the standard convex function, represented in the figure (8.5) for $\alpha=0.5$, which is written as follows:

$$f_{c}\left(x\right)=\left[\alpha x+\left(1-\alpha\right)\right]\log\frac{\alpha x+\left(1-\alpha\right)}{x}+\left(1-\alpha\right)\left(x-1\right) \tag{8.22}$$

Its gradient with respect to "q" is expressed as:

$$\frac{\partial G\left(p\|q\right)}{\partial q_{j}}=\left(1-\alpha\right)\log\frac{\alpha p_{j}+\left(1-\alpha\right)q_{j}}{p_{j}} \tag{8.23}$$

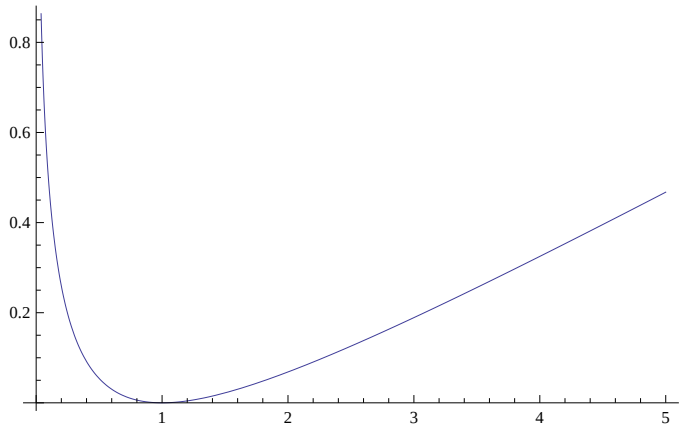

Figure 8.5: Function $f_c(x) = [\alpha x + (1-\alpha)] \log \frac{\alpha x + (1-\alpha)}{x} + (1-\alpha)(x-1)$, $\alpha = 0.5$

When we can consider that $\sum_i p_i = \sum_i q_i$, not necessarily equal to 1, a simplified form exists:

$$GS(p\|q) = \sum_i [\alpha p_i + (1-\alpha) q_i] \log \frac{\alpha p_i + (1-\alpha) q_i}{p_i} \tag{8.24}$$

It is a Csiszär divergence built on the simple convex function shown in the figure (8.6) for $\alpha = 0.5$, which is written:

$$f(x) = [\alpha x + (1-\alpha)] \log \frac{\alpha x + (1-\alpha)}{x} \tag{8.25}$$

8.2.2 Dual form.

The constructive process can be described by the relation:

$$G(q\|p) = KL(\alpha q + (1-\alpha) p\|q) \tag{8.26}$$

If we take the general form of the Kullback-Leibler divergence that can be used for non-normalized fields, we obtain:

$$\begin{aligned} G(q\|p) = &\sum_i [\alpha q_i + (1-\alpha) p_i] \log \frac{\alpha q_i + (1-\alpha) p_i}{q_i} \\ &+ (1-\alpha)(q_i - p_i) \end{aligned} \tag{8.27}$$

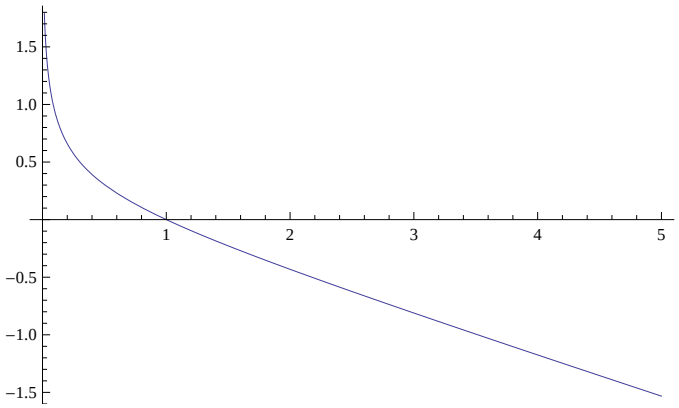

Figure 8.6: Function $f(x) = [\alpha x + (1-\alpha)] \log \frac{\alpha x + (1-\alpha)}{x}$, $\alpha = 0.5$

It is a Csiszär divergence built on the standard convex function shown in the figure (8.7) for $\alpha = 0.5$, which is written:

$$\breve{f}_c(x) = [\alpha + (1-\alpha)\, x] \log [\alpha + (1-\alpha)\, x] + (1-\alpha)(1-x) \tag{8.28}$$

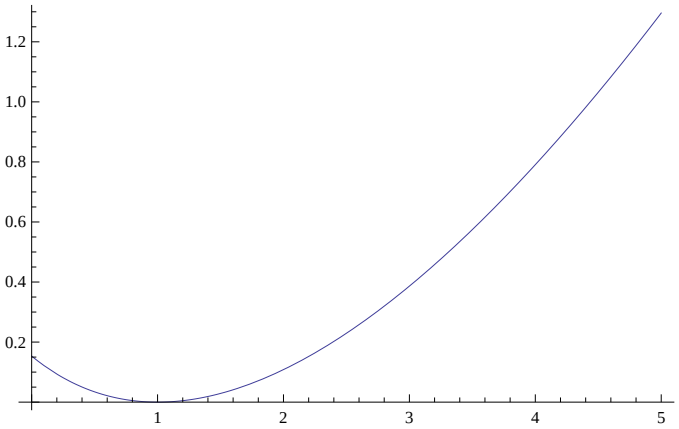

Figure 8.7: Function $\breve{f}_c(x) = [\alpha + (1-\alpha)\, x] \log [\alpha + (1-\alpha)\, x] + (1-\alpha)(1-x)$, $\alpha = 0.5$

The gradient with respect to "q" is given by:

$$\frac{\partial G(q\|p)}{\partial q_j} = \alpha \log \frac{\alpha q_j + (1-\alpha)\, p_j}{q_j} + (1-\alpha)\left(1 - \frac{p_j}{q_j}\right) \tag{8.29}$$

When $\sum_i p_i = \sum_i q_i$, not necessarily equal to 1, a simplified form exists, it is written as follows:

$$GS(q\|p) = \sum_i [\alpha q_i + (1-\alpha) p_i] \log \frac{\alpha q_i + (1-\alpha) p_i}{q_i} \tag{8.30}$$

It is a Csiszär divergence built on the simple convex function represented on the figure (8.8) for $\alpha = 0.5$, which is written:

$$\breve{f}(x) = [\alpha + (1-\alpha) x] \log [\alpha + (1-\alpha) x] \tag{8.31}$$

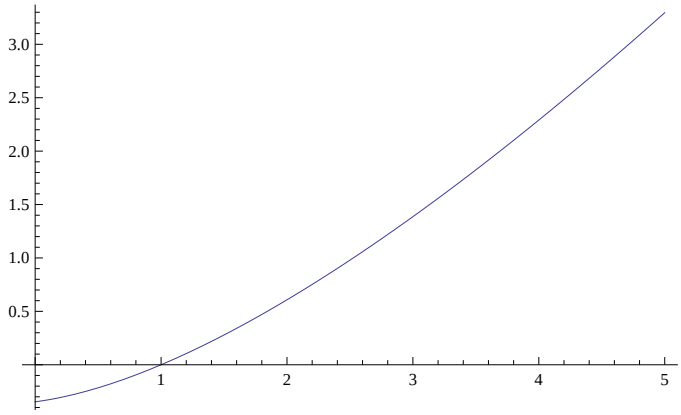

Figure 8.8: Function $\breve{f}(x) = [\alpha + (1-\alpha) x] \log [\alpha + (1-\alpha) x]$, $\alpha = 0.5$

8.2.3 T symmetrical divergence.

It is classically constructed in the form of:

$$T(p,q) = G(p\|q) + G(q\|p) \tag{8.32}$$

That is:

$$\begin{aligned} T(p,q) &= \sum_i [\alpha p_i + (1-\alpha) q_i] \log \frac{\alpha p_i + (1-\alpha) q_i}{p_i} \\ &+ \sum_i [\alpha q_i + (1-\alpha) p_i] \log \frac{\alpha q_i + (1-\alpha) p_i}{q_i} \end{aligned} \tag{8.33}$$

The corresponding gradient with respect to "q" is immediately deduced from the gradients of the functions $G(p\|q)$ and $G(q\|p)$.

8.2.4 Invariance by change of scale on "q".

Regarding the $G(p\|q)$ divergence given by the equation (8.21), the classical method does not explicitly allow to calculate a value for the invariance factor K.
By introducing the non-nominal invariance factor K^*, in the expression (8.21), we obtain the corresponding invariant divergence which is written after simplifications:

$$GI(p\|q) = \sum_i \left[\alpha\bar{p}_i + (1-\alpha)\,\bar{q}_i\right] \log \frac{\alpha\bar{p}_i + (1-\alpha)\,\bar{q}_i}{\bar{p}_i} \tag{8.34}$$

Its gradient with respect to "q" is given by:

$$\frac{\partial GI(p\|q)}{\partial q_l} = \frac{1-\alpha}{\sum_j q_j}\left[\log \frac{\alpha\bar{p}_l + (1-\alpha)\,\bar{q}_l}{\bar{p}_l} - \sum_i \bar{q}_i \log \frac{\alpha\bar{p}_i + (1-\alpha)\,\bar{q}_i}{\bar{p}_i}\right] \tag{8.35}$$

8.3 Generalization of the F and G divergences.

To proceed with this generalization, the same work is done as in the previous sections, but replacing the Kullback-Leibler divergence used systematically in the constructive process, by the Havrda-Charvat divergence, i.e. something that is an "alpha (s) divergence", and relying on the fact that the K .L. divergence can be seen as a particular case of the H.C. divergence by variation of the parameter "s" ($0 \leq s \leq 1$) (we no longer use "α" which is already used to weight the two fields "p" and "q").
Indeed, the H.C. divergence is written as follows:

$$HC_s(a\|b) = \frac{1}{s(s-1)} \sum_i a_i^s b_i^{1-s} - sa_i - (1-s)\,b_i \tag{8.36}$$

This expression tends towards $KL(a\|b)$ when $s \to 1$ and towards $KL(b\|a)$ when $s \to 0$.
As in the previous sections, only the weighted forms will be developed; in this context the data fields that will be used are "p", "q", and the weighted combination "$\alpha p + (1-\alpha)\,q$".

8.3.1 Direct form.

We rely on the following constructive approach:

$$FG_{s,\alpha}(p\|q) = HC(\alpha p + (1-\alpha)\,q\|p) \tag{8.37}$$

This results in:

$$FG_{s,\alpha}(p\|q) = \frac{1}{s(s-1)} \sum_i \left[\alpha p_i + (1-\alpha) q_i\right]^s p_i^{1-s} - s\left[\alpha p_i + (1-\alpha) q_i\right] - (1-s) p_i \tag{8.38}$$

It can also be said that this divergence is built, in the sense of Csiszär, on the standard convex function:

$$f_c(x) = \frac{1}{s(s-1)} \left\{\left[\alpha x + (1-\alpha)\right]^s x^{1-s} - s\left[\alpha x + (1-\alpha)\right] - (1-s) x\right\} \tag{8.39}$$

Its gradient with respect to "q" is given by:

$$\frac{\partial FG_{s,\alpha}(p\|q)}{\partial q_j} = \frac{1-\alpha}{s-1} \left\{\left[\frac{\alpha p_j + (1-\alpha) q_j}{p_j}\right]^{s-1} - 1\right\} \tag{8.40}$$

If we are in the case where $\sum_i p_i = \sum_i q_i$, we can use a simplified form of this divergence which is written as follows:

$$FGS_{s,\alpha}(p\|q) = \frac{1}{s(s-1)} \sum_i \left[\alpha p_i + (1-\alpha) q_i\right]^s p_i^{1-s} - p_i \tag{8.41}$$

This is what would have been obtained if, in the mode of construction, a simplified H.C. divergence had been used, but it can also be said that this divergence is constructed in the sense of Csiszär on the simple convex function:

$$f(x) = \frac{1}{s(s-1)} \left\{\left[\alpha x + (1-\alpha)\right]^s x^{1-s} - x\right\} \tag{8.42}$$

8.3.2 Invariant form by change of scale on "q".

Following the usual procedure, we have to solve in K, the equation:

$$\frac{1-\alpha}{s-1} \sum_i q_i \left\{\left[\frac{\alpha p_i + (1-\alpha) K q_i}{p_i}\right]^{s-1} - 1\right\} = 0 \tag{8.43}$$

There is no explicit solution; however, it is possible to use:

$$K^* = \frac{\sum_i p_i}{\sum_i q_i} \tag{8.44}$$

This expression of K^* inserted in the initial divergence (8.38) makes it invariant by scale change and we obtain the simplified invariant divergence:

$$FGI\left(p\|q\right)=\frac{1}{s\left(s-1\right)}\sum_{i}\bar{p}_{i}^{1-s}\left[\alpha\bar{p}_{i}+\left(1-\alpha\right)\bar{q}_{i}\right]^{s} \tag{8.45}$$

In this expression, the normalized variables $\bar{p}$ and $\bar{q}$ are used.
The gradient with respect to "q" is written as :

$$\frac{\partial FGI\left(p\|q\right)}{\partial q_{l}}=\frac{1-\alpha}{\left(s-1\right)\sum_{j}q_{j}}\left\{\left[\frac{\bar{p}_{l}}{\alpha\bar{p}_{l}+\left(1-\alpha\right)\bar{q}_{l}}\right]^{s-1}-\sum_{i}\bar{q}_{i}\left[\frac{\bar{p}_{i}}{\alpha\bar{p}_{i}+\left(1-\alpha\right)\bar{q}_{i}}\right]^{s-1}\right\} \tag{8.46}$$

8.3.3 Dual form.

It is founded on the following constructive mode:

$$FG_{s,\alpha}=HC\left(\alpha q+\left(1-\alpha\right)p\|q\right) \tag{8.47}$$

This leads to:

$$\begin{aligned}FG_{s,\alpha}\left(q\|p\right)=&\frac{1}{s\left(s-1\right)}\sum_{i}\left[\alpha q_{i}+\left(1-\alpha\right)p_{i}\right]^{s}q_{i}^{1-s}\\&-s\left[\alpha q_{i}+\left(1-\alpha\right)p_{i}\right]-\left(1-s\right)q_{i}\end{aligned} \tag{8.48}$$

It can also be said that it is a Csiszär divergence built on the standard convex function:

$$\begin{aligned}\breve{f}_{cs}\left(x\right)=&\frac{1}{s\left(s-1\right)}\left\{\left[\alpha+\left(1-\alpha\right)x\right]^{s}-\left(1-s\right)\right.\\&\left.-s\left[\alpha+\left(1-\alpha\right)x\right]\right\}\end{aligned} \tag{8.49}$$

Its gradient with respect to "q" can be written as follows:

$$\begin{aligned}\frac{\partial FG_{s,\alpha}\left(q\|p\right)}{\partial q_{j}}=&\frac{\alpha}{s-1}\left\{\left[\frac{\alpha q_{j}+\left(1-\alpha\right)p_{j}}{q_{j}}\right]^{s-1}-1\right\}\\&-\frac{1}{s}\left\{\left[\frac{\alpha q_{j}+\left(1-\alpha\right)p_{j}}{q_{j}}\right]^{s}-1\right\}\end{aligned} \tag{8.50}$$

In the case where we have $\sum_{i}p_{i}=\sum_{i}q_{i}$, we obtain the following simplified form:

$$FGS_{s,\alpha}\left(q\|p\right)=\frac{1}{s\left(s-1\right)}\sum_{i}\left[\alpha q_{i}+\left(1-\alpha\right)p_{i}\right]^{s}q_{i}^{1-s}-q_{i} \tag{8.51}$$

It is constructed in the sense of Csiszär, on the simple convex function:

$$\breve{f}_s(x) = \frac{1}{s(s-1)}\left[\{\alpha + (1-\alpha)x\}^s - 1\right] \tag{8.52}$$

A further simplification can be introduced if one is dealing with probability densities, then, $\sum_i p_i = \sum_i q_i = 1$.

8.3.4 Symmetrical form.

The standard convex function for constructing the symmetrical divergence in the Csiszär sense is written (with a factor of 1/2):

$$\begin{aligned}\widehat{f}_{cs}(x) = f_{cs}(x) + \breve{f}_{cs}(x) = \frac{1}{s(s-1)}\Bigg\{x\left[\frac{\alpha x + (1-\alpha)}{x}\right]^s \\ + [\alpha + (1-\alpha)x]^s - (1+x)\Bigg\}\end{aligned} \tag{8.53}$$

The corresponding divergence is written:

$$\begin{aligned}FG_{s,\alpha}(p,q) = &\frac{1}{s(s-1)}\sum_i \left[\alpha p_i + (1-\alpha)q_i\right]^s p_i^{1-s} \\ &+ \left[\alpha q_i + (1-\alpha)p_i\right]^s q_i^{1-s} - (p_i + q_i)\end{aligned} \tag{8.54}$$

The gradient with respect to "q" can be written as follows:

$$\begin{aligned}\frac{\partial FG_{s,\alpha}(p,q)}{\partial q_j} &= \frac{1-\alpha}{s-1}\left[\frac{\alpha p_j + (1-\alpha)q_j}{p_j}\right]^{s-1} + \frac{\alpha}{s-1}\left[\frac{\alpha q_j + (1-\alpha)p_j}{q_j}\right]^{s-1} \\ &- \frac{1}{s}\left[\frac{\alpha q_j + (1-\alpha)p_j}{q_j}\right]^s - \frac{1}{s(s-1)}\end{aligned} \tag{8.55}$$

8.4 J relative divergence of Dragomir.

To broaden what has been proposed by Dragomir et al. [33], we will develop the general weighted case.

8.4.1 Direct form.

Founding on the use of K.L.'s divergences, one can write:

$$JD(p\|q) = KL\left[q\|\alpha p + (1-\alpha)q\right] + KL\left[\alpha p + (1-\alpha)q\|\, q\right] \tag{8.56}$$

Using the notations in the previous sections, one can also write:

$$JD\left(p\|q\right) = F\left(q\|p\right) + G\left(q\|p\right) \tag{8.57}$$

But it can also be said that it is a Csiszär divergence built on the standard convex function represented in the figure (8.9) for $\alpha = 0.5$, written as:

$$f_c\left(x\right) = \left(1-\alpha\right)\left(x-1\right)\log\left[\alpha + \left(1-\alpha\right)x\right] \tag{8.58}$$

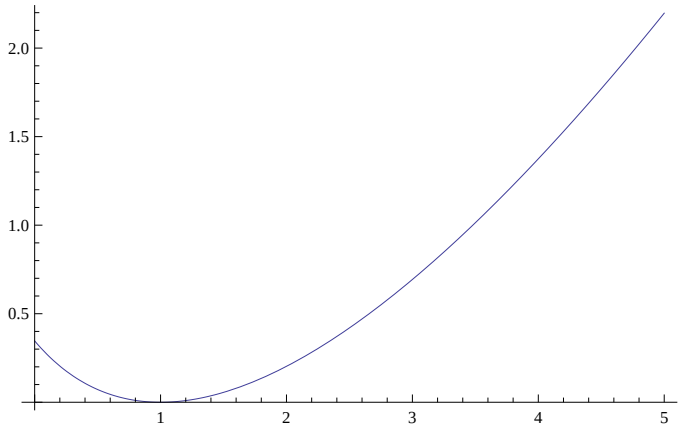

Figure 8.9: Function $f_c\left(x\right) = \left(1-\alpha\right)\left(x-1\right)\log\left[\alpha + \left(1-\alpha\right)x\right]$, $\alpha = 0.5$

Then, we obtain the divergence:

$$JD\left(p\|q\right) = \left(1-\alpha\right)\sum_i \left(p_i - q_i\right)\log\frac{\alpha q_i + \left(1-\alpha\right)p_i}{q_i} \tag{8.59}$$

The gradient with respect to "q" is expressed as:

$$\frac{\partial JD\left(p,q\right)}{\partial q_j} \propto \frac{\alpha\left(p_j - q_j\right)}{\alpha q_j + \left(1-\alpha\right)p_j} - \frac{p_j}{q_j} + \log\frac{q_j}{\alpha q_j + \left(1-\alpha\right)p_j} + 1 \tag{8.60}$$

8.4.2 Dual form.

It is constructed in the sense of Csiszär on the standard convex function, mirror of the function (8.58), represented on the figure (8.10) for $\alpha = 0.5$, that is written:

$$\breve{f}_c\left(x\right) = \left(1-\alpha\right)\left(1-x\right)\log\left[\frac{\alpha x + \left(1-\alpha\right)}{x}\right] \tag{8.61}$$

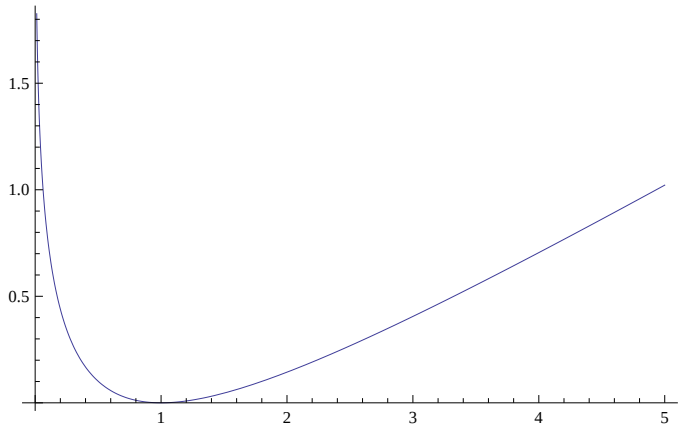

Figure 8.10: Function $\breve{f}_c(x) = (1-\alpha)(1-x)\log\left[\frac{\alpha x+(1-\alpha)}{x}\right]$, $\alpha = 0.5$

This results in:

$$JD(q\|p) = (1-\alpha)\sum_i (q_i - p_i)\log\left[\frac{\alpha p_i + (1-\alpha)q_i}{p_i}\right] \tag{8.62}$$

We can also say that:

$$JD(q\|p) = F(p\|q) + G(p\|q) \tag{8.63}$$

That is:

$$JD(q\|p) = KL[p\|\alpha p + (1-\alpha q)] + KL[\alpha p + (1-\alpha)q\|p] \tag{8.64}$$

The derivation of the gradient with respect to "q" leads to:

$$\frac{\partial JD(q,p)}{\partial q_j} \propto \log\left[\frac{\alpha p_j + (1-\alpha)q_j}{p_j}\right] + (1-\alpha)\frac{q_j - p_j}{\alpha p_j + (1-\alpha)q_j} \tag{8.65}$$

8.4.3 Invariance by change of scale on "q".

For the divergence given by the expression (8.59), the nominal invariance factor cannot be computed explicitly, so we use the expression K^*, which leads to the simplified invariant divergence:

$$JDI(p\|q) = (1-\alpha)\sum_i (\bar{p}_i - \bar{q}_i)\log\frac{\alpha\bar{q}_i + (1-\alpha)\bar{p}_i}{\bar{q}_i} \tag{8.66}$$

The gradient with respect to "q" is given by:

$$\frac{\partial JDI\left(p\|q\right)}{\partial q_l} = \frac{1-\alpha}{\sum_j q_j}\left[\sum_i \bar{q}_i \log\frac{\alpha\bar{q}_i+(1-\alpha)\,\bar{p}_i}{\bar{q}_i} - \log\frac{\alpha\bar{q}_l+(1-\alpha)\,\bar{p}_l}{\bar{q}_l} + (1-\alpha)\,\frac{\bar{p}_l\left(\bar{q}_l-\bar{p}_l\right)}{\bar{q}_l\left[\alpha\bar{q}_l+(1-\alpha)\,\bar{p}_l\right]} - (1-\alpha)\sum_i \frac{\bar{p}_i\left(\bar{q}_i-\bar{p}_i\right)}{\alpha\bar{q}_i+(1-\alpha)\,\bar{p}_i}\right] \tag{8.67}$$

8.4.4 Generalization of Dragomir's J relative divergence.

Starting from the standard convex function (8.58), the proposed generalization consists in replacing the logarithmic function by the "Generalized logarithm" function with the exponent $(1-d)$.
We thus obtain the standard convex function:

$$f_{c,d}\left(x\right) = \left(1-\alpha\right)\left(x-1\right)\left\{\frac{\left[\alpha+(1-\alpha)\,x\right]^{1-d}-1}{1-d}\right\} \tag{8.68}$$

If we use this function to construct a Csiszär divergence, we obtain:

$$JD_d\left(p\|q\right) = \frac{1-\alpha}{1-d}\sum_i \left(p_i-q_i\right)\left\{\left[\frac{\alpha q_i+(1-\alpha)\,p_i}{q_i}\right]^{1-d}-1\right\} \tag{8.69}$$

We can see that when $d \to 2$ this divergence becomes equal to:

$$\left(1-\alpha\right)^2\sum_i \frac{\left(p_i-q_i\right)^2}{\alpha q_i+(1-\alpha)\,p_i} \propto M_{AH} \tag{8.70}$$

The dual form is obtained from the mirror function of the previous one, i.e.:

$$\breve{f}_{c,d}\left(x\right) = \frac{1-\alpha}{1-d}\left(1-x\right)\left\{\left[\frac{\alpha x+(1-\alpha)}{x}\right]^{1-d}-1\right\} \tag{8.71}$$

The corresponding divergence will be written:

$$JD_d\left(q\|p\right) = \frac{1-\alpha}{1-d}\sum_i \left(q_i-p_i\right)\left\{\left[\frac{\alpha p_i+(1-\alpha)\,q_i}{p_i}\right]^{1-d}-1\right\} \tag{8.72}$$

8.5 Generalizations proposed by Taneja.

The generalization technique used by Taneja [92], consists in relying on the divergence based on the entropy of Sharma-Mittal [89] in the simplified form corresponding to normalized variables (4.33).
It is not much more complicated to use the general form corresponding to any variables (4.32) that is written:

$$SM_{\alpha,s}(a\|b) = \frac{1}{\alpha(s-1)}\left\{\left[\sum_i a_i^{\alpha} b_i^{1-\alpha}\right]^{\frac{s-1}{\alpha-1}} - \left[\sum_i \alpha a_i + (1-\alpha) b_i\right]^{\frac{s-1}{\alpha-1}}\right\} \tag{8.73}$$

We can remark that it is in fact the application of the Generalized Logarithm function, on the two terms of $M_{AG}(a\|b)$, with the exponent $\frac{s-1}{\alpha-1}$.
We then define the 1st generalization noted by Taneja ${}^1T_{\alpha}^{s}$:

$${}^1T_{\alpha}^{s}(p\|q) = \frac{1}{2}\left[SM_{\alpha,s}\left(\frac{p+q}{2}\|p\right) + SM_{\alpha,s}\left(\frac{p+q}{2}\|q\right)\right] \tag{8.74}$$

We can observe that for the pleasure of complicating things further, we can imagine replacing the half-sum by a weighted sum, but is it really necessary to do so ?
That being said, the explicit form of ${}^1T_{\alpha}^{s}$ is written:

$$\begin{aligned}{}^1T_{\alpha}^{s}(p\|q) = \frac{1}{2\alpha(s-1)}\Bigg\{&\left[\sum_i \left(\frac{p_i+q_i}{2}\right)^{\alpha} p_i^{1-\alpha}\right]^{\frac{s-1}{\alpha-1}} \\ &+ \left[\sum_i \left(\frac{p_i+q_i}{2}\right)^{\alpha} q_i^{1-\alpha}\right]^{\frac{s-1}{\alpha-1}} \\ &- \left[\sum_i \alpha\left(\frac{p_i+q_i}{2}\right) + (1-\alpha) p_i\right]^{\frac{s-1}{\alpha-1}} \\ &- \left[\sum_i \alpha\left(\frac{p_i+q_i}{2}\right) + (1-\alpha) q_i\right]^{\frac{s-1}{\alpha-1}}\Bigg\}\end{aligned} \tag{8.75}$$

If we consider the simplified configuration corresponding to $\sum_i p_i = \sum_i q_i = 1$, we have:

$$ {}^1TS_\alpha^s(p\|q) = \frac{1}{2\alpha(s-1)} \left\{ \left[\sum_i \left(\frac{p_i+q_i}{2} \right)^\alpha p_i^{1-\alpha} \right]^{\frac{s-1}{\alpha-1}} + \left[\sum_i \left(\frac{p_i+q_i}{2} \right)^\alpha q_i^{1-\alpha} \right]^{\frac{s-1}{\alpha-1}} - 2 \right\} \tag{8.76} $$

Now let's look at the limit expressions of ${}^1T_\alpha^s(p\|q)$ (8.75).
If $s \to \alpha$.
We obtain:

$$ {}^1T_\alpha^\alpha(p\|q) = \frac{1}{\alpha(\alpha-1)} \left\{ \sum_i \left(\frac{p_i+q_i}{2} \right)^\alpha \left(\frac{p_i^{1-\alpha}+q_i^{1-\alpha}}{2} \right) - \sum_i \left(\frac{p_i+q_i}{2} \right) \right\} \tag{8.77} $$

It is a Csizär divergence based on the standard convex function:

$$ f_c(x) = \frac{1}{\alpha(\alpha-1)} \left[\left(\frac{x^{1-\alpha}+1}{2} \right) \left(\frac{x+1}{2} \right)^\alpha - \frac{x+1}{2} \right] \tag{8.78} $$

It is similar to (8.54).
Note that this expression could have been obtained by using the Havrda-Charvat divergence directly in the general definition (8.74) instead of the divergence $SM_{\alpha,s}$.
Furthermore, the special case corresponding to $\sum_i p_i = \sum_i q_i = 1$ is immediately written as follows:

$$ {}^1TS_\alpha^\alpha(p\|q) = \frac{1}{\alpha(\alpha-1)} \left\{ \sum_i \left(\frac{p_i+q_i}{2} \right)^\alpha \left(\frac{p_i^{1-\alpha}+q_i^{1-\alpha}}{2} \right) - 1 \right\} \tag{8.79} $$

Of course, it cannot be constructed as a divergence of Csiszär.

The second limit case is to make $s \to 1$ in ${}^1T_\alpha^s(p\|q)$ (8.75) (we're trying to obtain something that is inspired by Renyi's divergence).

We obtain:

$$
\begin{aligned}
{}^{1}T_{\alpha}^{1}\left(p\|q\right) = \frac{1}{2\alpha\left(\alpha-1\right)} \Bigg\{ & \log \sum_{i} \left(\frac{p_i+q_i}{2}\right)^{\alpha} p_i^{1-\alpha} \\
& + \log \sum_{i} \left(\frac{p_i+q_i}{2}\right)^{\alpha} q_i^{1-\alpha} \\
& - \log \sum_{i} \alpha \left(\frac{p_i+q_i}{2}\right) + \left(1-\alpha\right) p_i \\
& - \log \sum_{i} \alpha \left(\frac{p_i+q_i}{2}\right) + \left(1-\alpha\right) q_i \Bigg\}
\end{aligned}
\tag{8.80}
$$

The divergence corresponding to $\sum_i p_i = \sum_i q_i = 1$ will be written:

$$
\begin{aligned}
{}^{1}TS_{\alpha}^{1}\left(p\|q\right) = \frac{1}{2\alpha\left(\alpha-1\right)} \Bigg\{ & \log \sum_{i} \left(\frac{p_i+q_i}{2}\right)^{\alpha} p_i^{1-\alpha} \\
& + \log \sum_{i} \left(\frac{p_i+q_i}{2}\right)^{\alpha} q_i^{1-\alpha} \Bigg\}
\end{aligned}
\tag{8.81}
$$

The 2nd generalization proposed by Taneja consists in starting from the expression of ${}^{1}T_{\alpha}^{\alpha}\left(p\|q\right)$ (8.77) which is a difference of two positive terms, and apply to each of them the Generalized Logarithm with the power $(s-1/\alpha-1)$ which is an increasing function; thus we obtain:

$$
\begin{aligned}
{}^{2}T_{\alpha}^{s}\left(p\|q\right) = \frac{1}{\alpha\left(s-1\right)} \Bigg\{ & \left[\sum_{i} \left(\frac{p_i+q_i}{2}\right)^{\alpha} \left(\frac{p_i^{1-\alpha}+q_i^{1-\alpha}}{2}\right)\right]^{\frac{s-1}{\alpha-1}} \\
& - \left[\sum_{i} \left(\frac{p_i+q_i}{2}\right)\right]^{\frac{s-1}{\alpha-1}} \Bigg\}
\end{aligned}
\tag{8.82}
$$

It is easy to verify that ${}^{2}T_{\alpha}^{\alpha}\left(p\|q\right) = {}^{1}T_{\alpha}^{\alpha}\left(p\|q\right)$.
Similarly, by making $s \to 1$ in (8.82), we obtain:

$$
\begin{aligned}
{}^{2}T_{\alpha}^{1}\left(p\|q\right) = \frac{1}{\alpha\left(\alpha-1\right)} \Bigg\{ & \log \sum_{i} \left(\frac{p_i+q_i}{2}\right)^{\alpha} \left(\frac{p_i^{1-\alpha}+q_i^{1-\alpha}}{2}\right) \\
& - \log \sum_{i} \left(\frac{p_i+q_i}{2}\right) \Bigg\}
\end{aligned}
\tag{8.83}
$$

In any case, if we consider the simplified version $\sum_i p_i = \sum_i q_i = 1$, we find the forms given by Taneja.

Remark: To fully understand the difference between these 2 generalizations, we can notice that they correspond to the following operations:
we consider the Havrda-Charvat divergences with a parameter "s", noted $HC_s(a\|b)$ with $a = \frac{p+q}{2}$ et $b = p$ on one hand, and $HC_s(a\|b)$ with $a = \frac{p+q}{2}$ et $b = q$ on the other hand.
* - In the case of the 1st generalization, we apply the Generalized Logarithm on the 2 terms of each of these divergences, then we make the half-sum of the divergences obtained, which gives (8.75) and its variants.
* - In the 2nd generalization, we first sum the 2 Havrda-Charvat divergences mentioned above, we obtain a difference of 2 positive terms, then, we apply the Generalized Logarithm on each of these 2 terms.
The results obtained by these two operating modes are different because the two operations carried out successively are not commutative.

To continue this small exercise, we take up an equivalent work already proposed by Taneja [91] and Arndt [6]; it consists in using the relation (8.74) and inverting the order of the arguments in the divergences $SM_{\alpha,s}$, that is:

$$ {}^1I^s_\alpha(p\|q) = \frac{1}{2}\left[SM_{\alpha,s}\left(p\|\frac{p+q}{2}\right) + SM_{\alpha,s}\left(q\|\frac{p+q}{2}\right)\right] \tag{8.84} $$

This can be considered as the 3rd generalization which is similar to the 1st generalization, except for the order of the arguments.
This being said, the explicit form of ${}^1I^s_\alpha$ is written:

$$ \begin{aligned} {}^1I^s_\alpha(p\|q) = \frac{1}{2\alpha(s-1)} \Bigg\{ & \left[\sum_i \left(\frac{p_i+q_i}{2}\right)^{1-\alpha} p_i^\alpha\right]^{\frac{s-1}{\alpha-1}} \\ & + \left[\sum_i \left(\frac{p_i+q_i}{2}\right)^{1-\alpha} q_i^\alpha\right]^{\frac{s-1}{\alpha-1}} \\ & - \left[\sum_i (1-\alpha)\left(\frac{p_i+q_i}{2}\right) + \alpha p_i\right]^{\frac{s-1}{\alpha-1}} \\ & - \left[\sum_i (1-\alpha)\left(\frac{p_i+q_i}{2}\right) + \alpha q_i\right]^{\frac{s-1}{\alpha-1}} \Bigg\} \end{aligned} \tag{8.85} $$

In the simplified case corresponding to $\sum_i p_i = \sum_i q_i = 1$, we will have the simplified forms exhibited in Arndt [6]:

$$
{}^{1}I_{\alpha}^{s}\left(p\|q\right) = \frac{1}{2\alpha\left(s-1\right)}\left\{\left[\sum_{i}\left(\frac{p_i+q_i}{2}\right)^{1-\alpha}p_i^{\alpha}\right]^{\frac{s-1}{\alpha-1}} + \left[\sum_{i}\left(\frac{p_i+q_i}{2}\right)^{1-\alpha}q_i^{\alpha}\right]^{\frac{s-1}{\alpha-1}} - 2\right\} \tag{8.86}
$$

Similarly, if in (8.85) we use $s = \alpha$, we will have:

$$
{}^{1}I_{\alpha}^{\alpha}\left(p\|q\right) = \frac{1}{\alpha\left(\alpha-1\right)}\left\{\sum_{i}\left(\frac{p_i+q_i}{2}\right)^{1-\alpha}\left(\frac{p_i^{\alpha}+q_i^{\alpha}}{2}\right) - \sum_{i}\left(\frac{p_i+q_i}{2}\right)\right\} \tag{8.87}
$$

It is a Csizär divergence based on the standard convex function:

$$
f_c\left(x\right) = \frac{1}{\alpha\left(\alpha-1\right)}\left[\left(\frac{x^{\alpha}+1}{2}\right)\left(\frac{x+1}{2}\right)^{1-\alpha} - \frac{x+1}{2}\right] \tag{8.88}
$$

If now, from (8.85), we change to $s \to 1$, we get:

$$
\begin{aligned}
{}^{1}I_{\alpha}^{1}\left(p\|q\right) = \frac{1}{2\alpha\left(\alpha-1\right)}\Bigg\{&\log\sum_{i}\left(\frac{p_i+q_i}{2}\right)^{1-\alpha}p_i^{\alpha} \\
&+\log\sum_{i}\left(\frac{p_i+q_i}{2}\right)^{1-\alpha}q_i^{\alpha} \\
&-\log\sum_{i}\left(1-\alpha\right)\left(\frac{p_i+q_i}{2}\right)+\alpha p_i \\
&-\log\sum_{i}\left(1-\alpha\right)\left(\frac{p_i+q_i}{2}\right)+\alpha q_i\Bigg\}
\end{aligned} \tag{8.89}
$$

If we take $\sum_i p_i = \sum_i q_i = 1$, we recover a simplified result from Arndt [6]:

$$
\begin{aligned}
{}^{1}IS_{\alpha}^{1}\left(p\|q\right) = \frac{1}{2\alpha\left(\alpha-1\right)}\Bigg\{&\log\sum_{i}\left(\frac{p_i+q_i}{2}\right)^{1-\alpha}p_i^{\alpha} \\
&+\log\sum_{i}\left(\frac{p_i+q_i}{2}\right)^{1-\alpha}q_i^{\alpha}\Bigg\}
\end{aligned} \tag{8.90}
$$

We can then define the 4th generalization which is the analogous of the 2nd one (except for the order of the arguments); to do so, we start from the relation giving ${}^{1}I_{\alpha}^{\alpha}(p\|q)$ (8.87) and an extension is carried out using the Generalized Logarithm, to obtain:

$$ {}^{2}I_{\alpha}^{s}(p\|q) = \frac{1}{\alpha(s-1)} \left\{ \left[\sum_i \left(\frac{p_i + q_i}{2} \right)^{1-\alpha} \left(\frac{p_i^{\alpha} + q_i^{\alpha}}{2} \right) \right]^{\frac{s-1}{\alpha-1}} - \left[\sum_i \left(\frac{p_i + q_i}{2} \right) \right]^{\frac{s-1}{\alpha-1}} \right\} \tag{8.91} $$

This divergence was exhibited by Taneja in [92].
If we perform the simplification $\sum_i p_i = \sum_i q_i = 1$, we retrieve an expression given by Arndt [6]:
On this relationship, moving to the $s \to 1$ limit gives us something close to a Renyi divergence:

$$ {}^{2}I_{\alpha}^{1}(p\|q) = \frac{1}{\alpha(\alpha-1)} \left\{ \log \sum_i \left(\frac{p_i + q_i}{2} \right)^{1-\alpha} \left(\frac{p_i^{\alpha} + q_i^{\alpha}}{2} \right) - \log \sum_i \left(\frac{p_i + q_i}{2} \right) \right\} \tag{8.92} $$

Finally, making the simplification $\sum_i p_i = \sum_i q_i = 1$, we retrieve Arndt's result [6]:

$$ {}^{2}IS_{\alpha}^{1}(p\|q) = \frac{1}{\alpha(\alpha-1)} \left\{ \log \sum_i \left(\frac{p_i + q_i}{2} \right)^{1-\alpha} \left(\frac{p_i^{\alpha} + q_i^{\alpha}}{2} \right) \right\} \tag{8.93} $$

On generalizations 3 and 4, we can make the same remarks as on generalizations 1 and 2 as to the order of the operations which made it possible to obtain them; the non-commutativity of these operations leads to *"Generalization 3 ≠ Generalization 4"*, as we had obtained *"Generalization 1 ≠ Generalization 2"*.
Another generalization (say the 5th generalization) is proposed by Taneja in [92] and may be illustrative of the fact that:

"**Every inequality can give rise to a divergence**".

Indeed, let us consider the inequality between weighted arithmetic mean $(MA)_i = \beta p_i + (1-\beta)\, q_i$ and the weighted geometric mean $(MG)_i = p_i^{\beta}\, q_i^{1-\beta}$;

$$(MG)_i - (MA)_i \leq 0 \quad \Rightarrow \quad (MG)_i^{1-r} - (MA)_i^{1-r} \leq 0 \;\; ; \;\; r \leq 1 \tag{8.94}$$

And then:

$$(MG)_i^{1-r}\,(MA)_i^{r} - (MA)_i \leq 0 \tag{8.95}$$

The extension to the entire field leads to:

$$\sum_i (MG)_i^{1-r}\,(MA)_i^{r} - \sum_i (MA)_i \leq 0 \tag{8.96}$$

And finally:

$$\frac{1}{r-1}\left[\sum_i (MG)_i^{1-r}\,(MA)_i^{r} - \sum_i (MA)_i\right] \geq 0 \tag{8.97}$$

From there, we can apply on each of the terms of the difference, an increasing function without changing the sign of the inequality, why not a generalized logarithmic function with the power $(s-1/r-1)$, to obtain the divergence proposed by Taneja:

$${}^{3}T_r^s\,(p\|q) = \frac{1}{s-1}\left\{\left[\sum_i (MG)_i^{1-r}\,(MA)_i^{r}\right]^{\frac{s-1}{r-1}} - \left[\sum_i (MA)_i\right]^{\frac{s-1}{r-1}}\right\} \tag{8.98}$$

In fact, Taneja [92] proposes something less general than that, because he limits it to the case of variables having a sum equal to 1, in which case the 2nd term of the difference will be equal to 1. The gradient with respect to "q_j" is written as follows:

$$\frac{\partial\left[{}^{3}T_r^s\,(p\|q)\right]}{\partial q_j} = \frac{1-\beta}{r-1}\left\{\left[\sum_i (MG)_i^{1-r}\,(MA)_i^{r}\right]^{\frac{s-r}{r-1}}\left[(1-r)\,\frac{(MA)_i^{r}}{(MG)_i^{r}}\frac{p_j^{\beta}}{q_j^{\beta}} + r\frac{(MA)_i^{r-1}}{(MG)_i^{r-1}}\right] - \left[\sum_i (MA)_i\right]^{\frac{s-r}{r-1}}\right\} \tag{8.99}$$

The 2 classical limit cases can be considered:

* $s = r$ leads to:

$$ {}^{3}T_{r}^{r}\left(p\|q\right) = \frac{1}{r-1}\left\{\sum_{i}\left(MG\right)_{i}^{1-r}\left(MA\right)_{i}^{r} - \sum_{i}\left(MA\right)_{i}\right\} \tag{8.100} $$

* $s \to 1$ leads to the Logarithmic form:

$$ {}^{3}T_{r}^{1}\left(p\|q\right) = \frac{1}{r-1}\left\{\log\left[\sum_{i}\left(MG\right)_{i}^{1-r}\left(MA\right)_{i}^{r}\right] - \log\left[\sum_{i}\left(MA\right)_{i}\right]\right\} \tag{8.101} $$

The corresponding gradients with respect to "q" are obtained immediately as a limit case of (8.99).

This last divergence is referred to in Taneja [92] and Neemuchwala [72] with the name "αGA" which can be noted here as "rGA", and for which they only consider the case of standardized variables.
Of course, this fifth generalization is in relation with the divergences based on inequalities between the generalized means developed in **chapter 7**.

If we want to extend this procedure to other means, we rely on the inequalities between weighted and unweighted means $MH \leq MG \leq MA \leq MQ$.
With:

$$ \left(MH\right)_{i} = \frac{p_{i}q_{i}}{\beta p_{i} - \left(1-\beta\right)q_{i}} \quad ; \quad \left(MQ\right)_{i} = \sqrt{\beta p_{i}^{2} + \left(1-\beta\right)q_{i}^{2}} \tag{8.102} $$

We can write:

$$ {}^{3}THG_{r}^{s}\left(p\|q\right) = \frac{1}{s-1}\left\{\left[\sum_{i}\left(MH\right)_{i}^{1-r}\left(MG\right)_{i}^{r}\right]^{\frac{s-1}{r-1}} - \left[\sum_{i}\left(MG\right)_{i}\right]^{\frac{s-1}{r-1}}\right\} \tag{8.103} $$

$$ {}^{3}THA_{r}^{s}\left(p\|q\right) = \frac{1}{s-1}\left\{\left[\sum_{i}\left(MH\right)_{i}^{1-r}\left(MA\right)_{i}^{r}\right]^{\frac{s-1}{r-1}} - \left[\sum_{i}\left(MA\right)_{i}\right]^{\frac{s-1}{r-1}}\right\} \tag{8.104} $$

$$ {}^3THQ_r^s(p\|q) = \frac{1}{s-1}\left\{\left[\sum_i (MH)_i^{1-r}(MQ)_i^r\right]^{\frac{s-1}{r-1}} - \left[\sum_i (MQ)_i\right]^{\frac{s-1}{r-1}}\right\} \tag{8.105} $$

The divergence:

$$ {}^3TGA_r^s(p\|q) = {}^3T_r^s(p\|q) \tag{8.106} $$

has been previously expressed;

$$ {}^3TGQ_r^s(p\|q) = \frac{1}{s-1}\left\{\left[\sum_i (MG)_i^{1-r}(MQ)_i^r\right]^{\frac{s-1}{r-1}} - \left[\sum_i (MQ)_i\right]^{\frac{s-1}{r-1}}\right\} \tag{8.107} $$

$$ {}^3TAQ_r^s(p\|q) = \frac{1}{s-1}\left\{\left[\sum_i (MA)_i^{1-r}(MQ)_i^r\right]^{\frac{s-1}{r-1}} - \left[\sum_i (MQ)_i\right]^{\frac{s-1}{r-1}}\right\} \tag{8.108} $$

Logarithmic forms can be inferred immediately.
Simplified forms of these divergences are developed in **chapter 7** dealing with divergences between means.

chapter 9 - Smoothness constraint regularization.

We consider here, the application of scale invariant divergences to Tikhonov-type smoothness constraint regularization [93].
This analysis has an obvious interest when one must introduce this type of regularization in a problem for which there is a constraint on the sum of the unknown parameters and, in particular, when it is associated with a non-negativity constraint. Indeed, when regularizing an inverse problem by a smoothness constraint, we are led to minimize in "x", an objective function of the form:

$$J(x) = D_1(y\|Hx) + \gamma D_2(x\|x_d) \tag{9.1}$$

The first term D_1 is a divergence which represents the attachment to the data (data consistency), i.e. a discrepancy between the measurements and the model (here linear), whereas the second term D_2 is a divergence which represents a discrepancy between the solution "x" and a default solution "x_d"; this default solution must obviously satisfy the constraints imposed on the solution, for example, $[x_d]_i \geq 0\ \forall i$, and if we have a sum constraint, $\sum_i [x_d]_i = \sum_i x_i$.
The coefficient "γ" is the regularization factor that adjusts the relative importance of the 2 terms of $J(x)$.
We consider here, the case of divergences which are invariant by change of scale on "q" constructed on the basis of "α" and "β" divergences, and we consider more specifically the logarithmic forms which are not only invariant with respect to "q", but also with respect to "p".
On the other hand, it was emphasized that when a divergence is made invariant with respect to "q" using the invariance factor $K^* = \frac{\sum_j p_j}{\sum_j q_j}$, the resulting divergence is also invariant with respect to "p"; this type of invariant diver-

gence will also be considered.
The invariant divergences with respect to both arguments have the advantage of applying in all cases, whether the default solution is a constant, or a linear function of the solution itself (Laplacian case for example).
In the minimization problem, we will use the expression of the gradient:

$$\frac{\partial J(x)}{\partial x_l} = \frac{\partial D_1(y\|Hx)}{\partial x_l} + \gamma \frac{\partial D_2(x\|x_d)}{\partial x_l} \tag{9.2}$$

If the divergences D_1 and D_2 are invariant divergences, the characteristic property of such divergences will result in:

$$\sum_l x_l \frac{\partial J(x)}{\partial x_l} = 0 \tag{9.3}$$

9.1 Invariant divergences derived from the "α" and "β" divergences.

9.1.1 Constant default solution = C.

In this case, all invariant divergences can be used, whether logarithmic (LAI, LBI) or not (AI, BI); we will just develop the case for logarithmic forms.
In the following discussions, for the two-dimensional case, the tables under consideration are written in lexicographical order.

*** With "LAI".**

We have:

$$LAI(c\|x) = \frac{1}{a}\log\sum_i x_i - \frac{1}{a-1}\log\sum_i c_i + \frac{1}{a(a-1)}\log\sum_i c_i^a x_i^{1-a} \tag{9.4}$$

Taking into account the fact that all c_i are equals whatever "i", we have the gradient:

$$\frac{\partial LAI(c\|x)}{\partial x_j} = \frac{1}{a}\left[\frac{1}{\sum_i x_i} - \frac{x_j^{-a}}{\sum_i x_i^{1-a}}\right] \tag{9.5}$$

*** With "LBI".**

We then have:

$$LBI(c\|x) = \frac{1}{b(b-1)}\log\sum_i c_i^b + \frac{1}{b}\log\sum_i x_i^b - \frac{1}{b-1}\log\sum_i c_i x_i^{b-1} \tag{9.6}$$

Which gradient is:

$$\frac{\partial LBI\left(c\|x\right)}{\partial x_j} = \frac{x_j^{b-1}}{\sum_i x_i^b} - \frac{x_j^{b-2}}{\sum_i x_i^{b-1}} \tag{9.7}$$

*** With the general form "$LABI$".**

The divergence considered is written as follows:

$$\begin{aligned} LABI\left(c\|x\right) = & \frac{1}{a+b-1}\left\{\frac{1}{b-1}\log\sum_i c_i^{a+b-1} - \right. \\ & \frac{a+b-1}{a\left(b-1\right)}\log\sum_i c_i^a x_i^{b-1} + \\ & \left.\frac{1}{a}\log\sum_i x_i^{a+b-1}\right\} \end{aligned} \tag{9.8}$$

After some simple calculations, the gradient is written as follows:

$$\frac{\partial LABI\left(c\|x\right)}{\partial x_j} = \frac{1}{a}\left[\frac{x_j^{a+b-2}}{\sum_i x_i^{a+b-1}} - \frac{x_j^{b-2}}{\sum_i x_i^{b-1}}\right] \tag{9.9}$$

The two previous particular cases can be found without difficulty by respectively making $a+b-1=1$, i.e. $b-1=1-a$ (LAI) and $a=1$ (LBI).

9.1.2 Use of the Laplacian operator.

In this case, the default solution "x_d" is a smoothed version of the solution "x", i.e. in a trivial way, a solution which does not contain high frequencies or anyway they are reduced; in the classical sense of Tikhonov [93], the regularization term is written as follows:

$$D_2\left(x\|x_d\right) = \|Lx\|^2 \tag{9.10}$$

Here we must distinguish between 2 situations:

*** 1 - One-dimensional case.**

In this case, the regularization term "Lx" corresponds to the convolution of "x" by the mask $\left[-\frac{1}{2}\ ;\ 1\ ;\ -\frac{1}{2}\right]$.
We can therefore write the result in the form "$x-Tx$" where the term "Tx"

written in the sense of the matrix vector product, represents the convolution of "x" by the mask $\left[\frac{1}{2}\ ;\ 0\ ;\ \frac{1}{2}\right]$.

*** 2 - Two-dimensional case.**

In this case, the regularization term "Lx" corresponds to the convolution of the table "x" by the mask: $\begin{pmatrix} 0 & -1/4 & 0 \\ -1/4 & 1 & -1/4 \\ 0 & -1/4 & 0 \end{pmatrix}$. We can therefore write the result in the form "$x - Tx$" where the term "Tx" written in the sense of the matrix vector product, represents the convolution of "x" by the mask: $\begin{pmatrix} 0 & 1/4 & 0 \\ 1/4 & 0 & 1/4 \\ 0 & 1/4 & 0 \end{pmatrix}$.

At the end of this operation, the table "Tx" is ordered lexicographically, as well as the table "x".

Note that in these conditions, the matrix T is symmetrical, its columns are normed to 1, that is $\sum_i t_{ij} = 1\ \forall j$, then $\sum_i (Tx)_i = \sum_i x_i$.
Consequently, as a regularization term one must write a divergence invariant by scale change between $p \equiv x$ and $q \equiv Tx$ for example; since "p" and "q" both depend on the "x" solution, it is understandable that we must use the Logarithmic form which is invariant with respect to both arguments (we can verify that the non Logarithmic forms do not allow us to develop algorithms that spontaneously maintain the flow, i.e. which do not spontaneously respect the sum constraint).

*** With "LAI".**

Using the notations $p \equiv x$ and $q \equiv Tx$, the divergence is expressed as:

$$LAI\,(x\|Tx) = \frac{1}{a}\log\sum_i (Tx)_i - \frac{1}{a-1}\log\sum_i x_i + \frac{1}{a\,(a-1)}\log\sum_i x_i^a\,(Tx)_i^{1-a} \tag{9.11}$$

The associated gradient is written:

$$\begin{aligned}\frac{\partial LAI\,(x\|Tx)}{\partial x_j} =& \frac{1}{a}\left[\frac{\sum_i t_{i,j}}{\sum_i (Tx)_i} - \frac{\sum_i t_{i,j} x_i^a\,(Tx)_i^{-a}}{\sum_i x_i^a\,(Tx)_i^{1-a}}\right] + \\ & \frac{1}{a-1}\left[\frac{x_j^{a-1}\,(Tx)_j^{1-a}}{\sum_i x_i^a\,(Tx)_i^{1-a}} - \frac{1}{\sum_i x_i}\right]\end{aligned} \tag{9.12}$$

We may observe that, as always with this type of divergence:
$\sum_j x_j \frac{\partial}{\partial x_j} = 0$.

*** With "LBI".**

With the notations in the previous paragraph, the divergence considered is written as follows:

$$LBI\left(x\|Tx\right) = \frac{1}{b\left(b-1\right)} \log \sum_i x_i^b + \frac{1}{b} \log \sum_i \left(Tx\right)_i^b - \frac{1}{b-1} \log \sum_i x_i \left(Tx\right)_i^{b-1} \tag{9.13}$$

The gradient will be:

$$\begin{aligned} \frac{\partial LBI\left(x\|Tx\right)}{\partial x_j} =& \frac{1}{b-1} \left[\frac{x_j^{b-1}}{\sum_i x_i^b} - \frac{\left(Tx\right)_j^{b-1}}{\sum_i x_i \left(Tx\right)_i^{b-1}} \right] + \\ & \left[\frac{\sum_i t_{i,j} \left(Tx\right)_i^{b-1}}{\sum_i \left(Tx\right)_i^b} - \frac{\sum_i t_{i,j} x_i \left(Tx\right)_i^{b-2}}{\sum_i x_i \left(Tx\right)_i^{b-1}} \right] \end{aligned} \tag{9.14}$$

Again, as always with this type of divergence we have:
$\sum_j x_j \frac{\partial}{\partial x_j} = 0$.

*** With the general form "$LABI$".**

If we use the general form of the divergence, the two previous cases can be recovered without any problem, indeed, the divergence is written as follows:

$$\begin{aligned} LABI\left(x\|Tx\right) =& \frac{1}{a+b-1} \left\{ \frac{1}{b-1} \log \sum_i x_i^{a+b-1} - \right. \\ & \frac{a+b-1}{a\left(b-1\right)} \log \sum_i x_i^a \left(Tx\right)_i^{b-1} + \\ & \left. \frac{1}{a} \log \sum_i \left(Tx\right)_i^{a+b-1} \right\} \end{aligned} \tag{9.15}$$

We can calculate the gradient, which is expressed as follows:

$$\begin{aligned} \frac{\partial LABI\left(x\|Tx\right)}{\partial x_j} =& \frac{1}{a} \left[\frac{\sum_i t_{i,j} \left(Tx\right)_i^{a+b-2}}{\sum_i \left(Tx\right)_i^{a+b-1}} - \frac{\sum_i t_{i,j} x_i^a \left(Tx\right)_i^{b-2}}{\sum_i x_i^a \left(Tx\right)_i^{b-1}} \right] + \\ & \frac{1}{b-1} \left[\frac{x_j^{a+b-2}}{\sum_i x_i^{a+b-1}} - \frac{x_j^{a-1} \left(Tx\right)_j^{b-1}}{\sum_i x_i^a \left(Tx\right)_i^{b-1}} \right] \end{aligned} \tag{9.16}$$

The expressions of the two preceding paragraphs are found immediately by using respectively: $a+b-1=1$, i.e. $b-1=1-a$ (which gives LAI) and $a=1$ (which gives LBI).

9.1.3 Important Note.

If, in the Laplacian regularization, we try to use the non logarithmic forms of the invariant divergences, that is "AI" or "BI", we loose the property of spontaneous maintaining of the flow, in fact these divergences are invariant only with respect to the second argument; to show this, it is enough to evaluate the gradient of $AI(x\|Tx)$ for example, and to note that the relation which allows the maintaining of the flow is not fulfilled.
In fact, if we use the expression of "AI", where $p \equiv x$ and $q \equiv Tx$, it comes:

$$AI(x\|Tx) = \frac{1}{a-1}\left[\left(\sum_i (Tx)_i\right)^{\frac{a-1}{a}}\left(\sum_i x_i^a (Tx)_i^{1-a}\right)^{\frac{1}{a}} - \sum_i x_i\right] \tag{9.17}$$

And according to the expression of the gradient we have:

$$\sum_j x_j \frac{\partial AI(x\|Tx)}{\partial x_j} = \frac{1}{1-a}\left\{\sum_j x_j - \left[\sum_j x_j^a (Tx)_j^{1-a}\right]^{\frac{1}{a}}\left[\sum_j (Tx)_j\right]^{\frac{a-1}{a}}\right\} \tag{9.18}$$

This expression could be equal to zero at convergence ($x \to Tx$), but certainly not at each iteration.

9.2 Effect of the invariance factor $K^* = \frac{\sum_j p_j}{\sum_j q_j}$.

In the study of the invariance factors K, the particular role of $K^* = \frac{\sum_j p_j}{\sum_j q_j}$ has been noted; indeed, the introduction of this factor in any divergence $D(p\|q)$, in order to make it invariant with respect to "q", led to a divergence $DI(p\|q)$ which is invariant with respect to this argument, obtained easily from the divergence $D(p\|q)$ by replacing "p" and "q" by normalized variables $\bar{p}_i = \frac{p_i}{\sum_j p_j}$ and $\bar{q}_i = \frac{q_i}{\sum_j q_j}$ respectively, with a multiplying factor depending only on $\sum_j p_j$ that can be omitted.
Disregarding this multiplicative factor, we obtain a divergence that is not only invariant with respect to "q", but also with respect to "p".

This observation allows us to consider the use of such divergences in the case of smoothness constraint regularization using the Laplacian of the solution, as developed in the previous sections; it is a possible alternative to the use of logarithmic forms "LAI" et"LBI".
Some examples to illustrate this point are given in Appendix 6.

9.2.1 Applications of some divergences to regularization.

We analyze in this paragraph the regularization in the sense of Tikhonov [93] with the Laplacian, as previously considered for the divergences $LAI(p\|q)$ and $LBI(p\|q)$.
We will focus on the cases of a few classical divergences that are rendered invariant by the use of the invariance factor $K^* = \frac{\sum_j p_j}{\sum_j q_j}$; note that for the K.L. divergence, this factor is the nominal invariance factor K_0.
For the application under consideration here, we have:

$$\bar{p}_i = \frac{x_i}{\sum_j x_j}, \qquad \bar{q}_i = \frac{(Tx)_i}{\sum_j (Tx)_j} \tag{9.19}$$

With $\sum_j t_{jl} = 1 \ \forall l$ (the columns of T are normalized to 1), it comes:

$$\frac{\partial \bar{p}_i}{\partial x_l} = \frac{1}{\sum_j x_j}(\delta_{i,l} - \bar{p}_i), \qquad \frac{\partial \bar{q}_i}{\partial x_l} = \frac{1}{\sum_j (Tx)_j}(t_{il} - \bar{q}_i) \tag{9.20}$$

1 - Mean square deviation.

The invariant divergence we obtain is written as follows (see (83) in Appendix 6):

$$EQMI(p\|q) = \sum_i (\bar{p}_i - \bar{q}_i)^2 \tag{9.21}$$

Then:

$$\frac{\partial EQMI(p\|q)}{\partial x_l} = \sum_i (\bar{p}_i - \bar{q}_i)\left(\frac{\partial \bar{p}_i}{\partial x_l} - \frac{\partial \bar{q}_i}{\partial x_l}\right) \tag{9.22}$$

Taking into account the fact that $\sum_j x_j = \sum_j (Tx)_j$, we obtain after a few simple calculations:

$$\frac{\partial EQMI(p\|q)}{\partial x_l} = \frac{1}{\sum_j x_j}\left[(\bar{p}_l - \bar{q}_l) - \sum_i (\bar{p}_i - \bar{q}_i)^2 - \sum_i t_{il}(\bar{p}_i - \bar{q}_i)\right] \tag{9.23}$$

or also:

$$\frac{\partial EQMI\left(p\|q\right)}{\partial x_l} = \frac{1}{\sum_j x_j}\left\{\left[\left(I - T^T\right)\left(\bar{p} - \bar{q}\right)\right]_l - \sum_i \left(\bar{p}_i - \bar{q}_i\right)^2\right\} \tag{9.24}$$

Coming back to the true variables (9.19), it comes:

$$\frac{\partial EQMI\left(p\|q\right)}{\partial x_l} = \frac{1}{\left(\sum_j x_j\right)^2}\left\{\left[\left(I - T^T\right)\left(x - Tx\right)\right]_l - \frac{1}{\sum_j x_j}\sum_i \left[x_i - \left(Tx\right)_i\right]^2\right\} \tag{9.25}$$

We can verify that, as always:

$$\sum_l x_l \frac{\partial EQMI\left(p\|q\right)}{\partial x_l} = 0 \tag{9.26}$$

2 - Kullback-Leibler divergence.

Taking into account the invariance factor that is used, and the possible simplifications, the following is obtained (see (87) in Appendix 6):

$$KLI\left(p\|q\right) = \sum_i \bar{p}_i \log\frac{\bar{p}_i}{\bar{q}_i} \tag{9.27}$$

Then:

$$\frac{\partial KLI\left(p\|q\right)}{\partial x_l} = \sum_i \left(\log\frac{\bar{p}_i}{\bar{q}_i} + 1\right)\frac{\partial \bar{p}_i}{\partial x_l} - \sum_i \frac{\bar{p}_i}{\bar{q}_i}\frac{\partial \bar{q}_i}{\partial x_l} \tag{9.28}$$

With the partial derivative expressions previously mentioned (9.20), and taking into account that $\sum_j x_j = \sum_j \left(Tx\right)_j$, it comes all calculations made:

$$\frac{\partial KLI\left(p\|q\right)}{\partial x_l} = \frac{1}{\sum_j x_j}\left[\log\frac{\bar{p}_l}{\bar{q}_l} + 1 - \sum_i t_{il}\frac{\bar{p}_i}{\bar{q}_i} - \sum_i \bar{p}_i \log\frac{\bar{p}_i}{\bar{q}_i}\right] \tag{9.29}$$

With $\sum_i t_{il} = 1$, we can write:

$$\frac{\partial KLI\left(p\|q\right)}{\partial x_l} = \frac{1}{\sum_j x_j}\left\{\log\frac{\bar{p}_l}{\bar{q}_l} - \sum_i \bar{p}_i \log\frac{\bar{p}_i}{\bar{q}_i} + \left[T^T\left(1 - \frac{\bar{p}}{\bar{q}}\right)\right]_l\right\} \tag{9.30}$$

Coming back to the true variables (9.19), it comes:

$$\frac{\partial KLI\left(p\|q\right)}{\partial x_l} = \frac{1}{\sum_j x_j}\left\{\log\frac{x_l}{\left(Tx\right)_l} - \sum_i \frac{x_i}{\sum_j x_j}\log\frac{x_i}{\left(Tx\right)_i} + \left[T^T\left(1 - \frac{x}{\left(Tx\right)}\right)\right]_l\right\} \tag{9.31}$$

We can verify that, as always:

$$\sum_l x_l \frac{\partial KLI\left(p\|q\right)}{\partial x_l} = 0 \tag{9.32}$$

3 - Neyman's Chi2 divergence.

The invariant divergence is written (see (91) in Appendix 6):

$$\chi^2_N I\left(p\|q\right) = \sum_i \frac{\left(\bar{p}_i - \bar{q}_i\right)^2}{\bar{q}_i} \tag{9.33}$$

Then:

$$\frac{\partial \chi^2_N I\left(p\|q\right)}{\partial x_l} = 2\sum_i \left(\frac{\bar{p}_i}{\bar{q}_i} - 1\right)\frac{\partial \bar{p}_i}{\partial x_l} + \sum_i \left[1 - \left(\frac{\bar{p}_i}{\bar{q}_i}\right)^2\right]\frac{\partial \bar{q}_i}{\partial x_l} \tag{9.34}$$

Considering the expressions of the partial derivatives of $\bar{p}_i$ and $\bar{q}_i$ with respect to "x_l" (9.20), it comes all calculations made, and with the simplifications already indicated:

$$\frac{\partial \chi^2_N I\left(p\|q\right)}{\partial x_l} = \frac{1}{\sum_j x_j}\left[2\frac{\bar{p}_l}{\bar{q}_l} - \sum_i t_{il}\left(\frac{\bar{p}_i}{\bar{q}_i}\right)^2 - \sum_i \frac{\bar{p}_i^2}{\bar{q}_i}\right] \tag{9.35}$$

Turning back to the true variables, we have:

$$\frac{\partial \chi^2_N I\left(p\|q\right)}{\partial x_l} = \frac{1}{\sum_j x_j}\left\{2\frac{x_l}{(Tx)_l} - \left[T^T \frac{x^2}{(Tx)^2}\right]_l - \frac{1}{\sum_j x_j}\sum_i \frac{x_i^2}{(Tx)_i}\right\} \tag{9.36}$$

And, then:

$$\sum_l x_l \frac{\partial \chi^2_N I\left(p\|q\right)}{\partial x_l} = 0 \tag{9.37}$$

4 - Pearson's Chi2 divergence.

The invariant divergence is written (see (95) in Appendix 6):

$$\chi^2_P I\left(p\|q\right) = \sum_i \frac{\left(\bar{p}_i - \bar{q}_i\right)^2}{\bar{p}_i} \tag{9.38}$$

Then:

$$\frac{\partial \chi^2_P I\left(p\|q\right)}{\partial x_l} = \sum_i \left[1 - \left(\frac{\bar{q}_i}{\bar{p}_i}\right)^2\right]\frac{\partial \bar{p}_i}{\partial x_l} + 2\sum_i \left[\left(\frac{\bar{q}_i}{\bar{p}_i}\right) - 1\right]\frac{\partial \bar{q}_i}{\partial x_l} \tag{9.39}$$

Considering the expressions of the partial derivatives of $\bar{p}_i$ and $\bar{q}_i$ with respect to "x_l" (9.20), it comes with all the calculations made, with the simplifications already indicated:

$$\frac{\partial \chi^2_P I\left(p\|q\right)}{\partial x_l} = \frac{1}{\sum_j x_j}\left[-\left(\frac{\bar{q}_l}{\bar{p}_l}\right)^2 + 2\sum_i t_{il}\frac{\bar{q}_i}{\bar{p}_i} - \sum_i \frac{\bar{q}_i^2}{\bar{p}_i}\right] \tag{9.40}$$

And, turning back to the true variables:

$$\frac{\partial \chi^2_P I\left(p\|q\right)}{\partial x_l} = \frac{1}{\sum_j x_j}\left\{-\left(\frac{(Tx)_l}{x_l}\right)^2 + 2\left[T^T\frac{(Tx)}{x}\right]_l - \frac{1}{\sum_j (Tx)_j}\sum_i \frac{(Tx)_i^2}{x_i}\right\} \tag{9.41}$$

Then, we have:

$$\sum_l x_l \frac{\partial \chi^2_P I\left(p\|q\right)}{\partial x_l} = 0 \tag{9.42}$$

chapitre 10 - Algorithmic point of view.

10.1 Some preliminary remarks.

In this chapter, different algorithmic methods for minimizing divergences, under non-negativity constraint and under sum constraint, are described. The divergences considered are assumed to be convex and differentiable.
We will first consider the minimization problem under the non-negativity constraint.
We will distinguish three cases corresponding to specific problems:
* Simplified divergences that can be used for fields of explicitly equal sums (possibly equal to 1).
* General divergences which contain no simplification related to the sum of the data fields.
* Divergences invariant by change of scale.
In a second step, one will introduce in addition, the sum constraint on the unknown parameters.

We're essentially interested in the separable divergences that will be noted:

$$D\left(p\|q\right) = \sum_i D\left(p_i\|q_i\right) = \sum_i d_i \tag{10.1}$$

The non-separable case is easily deduced from this.

Moreover, we will only consider the case of a linear model:

$$q = Hx \quad \Rightarrow \quad q_i = \sum_j h_{ij} x_j \quad \Rightarrow \quad \frac{\partial q_i}{\partial x_l} = h_{il} \tag{10.2}$$

And we will have:

$$\frac{\partial D\left(p\|q\right)}{\partial x_l} = \sum_i \frac{\partial d_i}{\partial q_i}\frac{\partial q_i}{\partial x_l} \tag{10.3}$$

Consequently, with the notation:

$$\frac{\partial d_i}{\partial q_i} = \left[\frac{\partial D}{\partial q}\right]_i \tag{10.4}$$

We will have:

$$\frac{\partial D\left(p\|q\right)}{\partial x_l} = \sum_i h_{il}\frac{\partial d_i}{\partial q_i} = \left[H^T\frac{\partial D}{\partial q}\right]_l \tag{10.5}$$

This justifies "*a posteriori*", the fact that for the divergences considered in this work, the expression of the gradient with respect to the variable "q" has been given almost systematically.

10.2 Split Gradient Method (S.G.M.).

In this section, we present the method of minimization under non-negativity constraint, widely developed elsewhere [62], [61], which allows us to understand the origin of multiplicative algorithms and we can easily see why these algorithms can suffer from convergence problems.
This method applies to divergences $D\left(p\|q\right)$ having a minimum for $p_i = q_i \ \ \forall i$; it is quite obvious that this excludes simplified divergences which do not have this property. We will always consider that the "p_i" are the "y_i" measures and that the model "q_i" is linearly related to the true unknowns "x_i" by the relation $q_i = (Hx)_i$.
In a later section dedicated to the introduction of the sum constraint, it will be shown that, subject to a change of variables, the S.G.M. method can be applied on simplified divergences which have the property of being zero for $p_i = q_i \ \ \forall i$.
In the following sections we will note:

$$D\left(p\|q\right) = D\left(y\|Hx\right) \equiv D\left(x\right) \tag{10.6}$$

10.2.1 Non-negativity constraint.

For a strictly convex divergence $D\left(x\right)$, we consider the problem:
Minimize with respect to "x": $D\left(x\right)$
Subject to the constraint $x \geq 0 \quad \equiv \quad x_i \geq 0 \ \ \forall i$.

Let $L(x,\lambda)$ the Lagrangian associated with this problem, which is expressed in the form of:

$$L(x,\lambda) = D(x) - \langle \lambda, g(x) \rangle \tag{10.7}$$

"λ" is the vector of Lagrange multipliers whose components are positive or zero ($\lambda_i \geq 0 \quad \forall i$); $\langle \lambda, g(x) \rangle$ is for the scalar product; $g(x) \equiv g(x_i) \equiv g_i(x) \quad \forall i$ is a function for expressing constraints; this function must be positive when the constraints are inactive ($x_i > 0$) and zero when the constraints are active ($x_i = 0$), moreover, if "x^*" denotes the optimum, the zeros of $g_i(x^*) / [\nabla g(x^*)]_i$ must be the same as those of $g_i(x^*)$.

We propose to build a constrained minimization algorithm based on Karush, Kuhn, Tucker (KKT) conditions [19], [17],[28] which must be satisfied at the optimum of the problem, i.e. at the solution (x^*, λ^*).
These conditions are written:

$$\nabla_x L(x^*,\lambda^*) = 0 \;\Rightarrow\; \lambda_i^* [\nabla g(x^*)]_i = [\nabla D(x^*)]_i \;\Rightarrow\; \lambda_i^* = \frac{[\nabla D(x^*)]_i}{[\nabla g(x^*)]_i} \tag{10.8}$$

$$g(x^*) \geq 0 \quad \Rightarrow \quad g_i(x^*) \geq 0 \quad \forall i \tag{10.9}$$

$$\lambda^* \geq 0 \quad \Rightarrow \quad \lambda_i^* \geq 0 \quad \forall i \tag{10.10}$$

$$\langle \lambda^* g(x^*) \rangle = 0 \tag{10.11}$$

Note that this last condition, which is a scalar product, results in fact, taking into account (10.9) and (10.10), in a set of complementary conditions:

$$\lambda_i^* g_i(x^*) = 0 \quad \Rightarrow \quad \frac{[\nabla D(x^*)]_i}{[\nabla g(x^*)]_i} g_i(x^*) = 0 \quad \forall i \tag{10.12}$$

Taking into account the properties of $g(x)$, this last condition can be written as follows:

$$[\nabla D(x^*)]_i \, g_i(x^*) = 0 \quad \forall i \tag{10.13}$$

To express it in more explicit way, we will have:

1. Either:

$$g_i(x^*) > 0 \quad \Rightarrow \quad \textit{The constraint is inactive} \tag{10.14}$$

then (10.13) is fulfilled if:

$$\lambda_i^* = 0 \quad \Leftrightarrow \quad [\nabla D(x^*)]_i = 0 \tag{10.15}$$

then "x^*" is the unconstrained solution of the problem.

2. Or:

$$\lambda_i^* > 0 \quad \Leftrightarrow \quad [\nabla D(x^*)]_i > 0 \tag{10.16}$$

then (10.13) is fulfilled if:

$$g_i(x^*) = 0 \quad \Rightarrow \quad \textit{The constraint is active} \tag{10.17}$$

3. or, in an extreme case, we have simultaneously:

$$\lambda_i^* = 0 \quad \Rightarrow \quad [\nabla D(x^*)]_i = 0 \quad et \quad g_i(x^*) = 0 \tag{10.18}$$

Then, (10.13) is "a fortiori" fulfilled, this is the case where the minimum corresponding to $\nabla D(x) = 0$ is located on the constraint.

These situations are illustrated (in the one-dimensional case) in the figure (10.1) for a constraint $x_i \geq m \quad \forall i$, by expressing the constraint using the function $g(x) = x - m$.

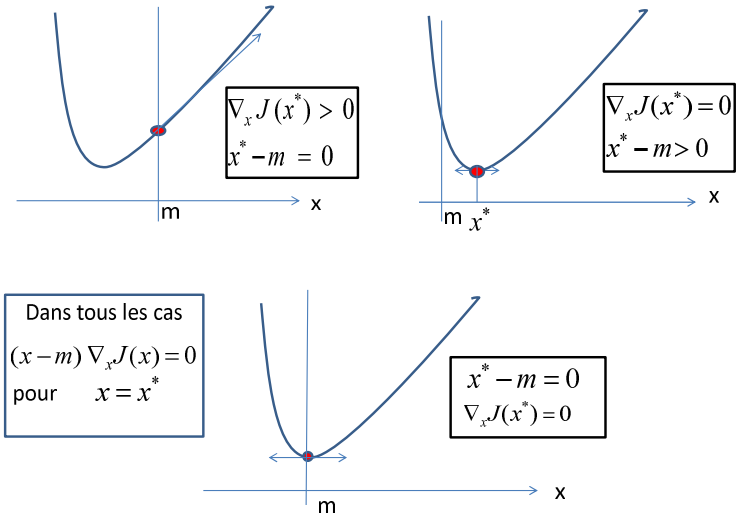

Figure 10.1: Karush, Kuhn, Tucker conditions for a constraint of the form $x \geq m$

The form of the function $g(x)$ used to express the constraints is of fundamental importance since it directly influences the final form and the behavior

of the algorithm.
For a non-negativity constraint, in the simplest case, one chooses $g(x) = x$ that is $g_i(x) = x_i \quad \forall i$.

Algorithm.

To solve (10.13), we rely on the fact that $-\nabla D(x)$ is a descent direction for the unconstrained problem, and that if M is a positive definite matrix, then any vector of the form $M[-\nabla D(x)]$ is also a descent direction for this problem; this is especially true if M is a diagonal matrix with positive terms.
We can also similarly rely on the successive substitution method [45] to write an iterative algorithm in the form:

$$x_i^{k+1} = x_i^k + \alpha_i^k x_i^k f_i\left(x^k\right)\left[-\nabla D\left(x^k\right)\right]_i \tag{10.19}$$

In this expression $\alpha_i^k > 0$ is the descent step size which will ensure both the fulfillment of the constraint and the convergence of the algorithm (in the sense of contracting applications, we can say that the proper choice of the step α_i^k ensures that the algorithm (10.19) of the type $x^{k+1} = F\left(x^k\right)$ is a contraction).
The function $f_i(x)$ is a function having positive values when we are in the constraint domain; it depends on the forms of $D(x)$ and $g(x)$ as well as on the wished properties for the algorithm, in particular if we want to obtain a purely multiplicative form.
We now analyze the properties of the algorithm, and in particular, whether or not the solution fulfills the KKT conditions.
The non-negativity constraint imposes restrictions on the descent step size. Indeed, to remain within a general framework, we consider that even if $D(x)$ is not defined everywhere, its definition domain is entirely contained within the constraint domain.

Therefore, the condition $x_i \geq 0 \ \forall i$ must be imposed first.

Thus, we define an interior points method (all successive estimates will fulfill the constraints).
So, as a first step, we need to ensure that at each iteration we have:

$$x_i^k \geq 0 \quad \Rightarrow \quad x_i^{k+1} \geq 0 \quad \forall i \tag{10.20}$$

Which leads, from (10.19), to the condition:

$$1 + \alpha_i^k f_i\left(x^k\right)\left[-\nabla D\left(x^k\right)\right]_i \geq 0 \quad \forall i \tag{10.21}$$

For $\left[-\nabla D\left(x^k\right)\right]_i > 0$, the condition (10.21) is always satisfied, in this case, for the relevant "i" components, there are no restrictions on the descent stepsize as far as the fulfillment of the constraints is concerned.
On the other hand, for $\left[-\nabla D\left(x^k\right)\right]_i < 0$ the condition (10.21) leads for the corresponding "i" components to:

$$\alpha_i^k \leq \frac{1}{f_i\left(x^k\right)\left[\nabla D\left(x^k\right)\right]_i} \tag{10.22}$$

This results in a set of descent step size values that guarantee the non-negativity of each component separately.
At iteration "k", the maximum step size to ensure the non-negativity of all the components of the solution will therefore be :

$$\alpha_M^k = \min_i \frac{1}{f_i\left(x^k\right)\left[\nabla D\left(x^k\right)\right]_i} \quad \forall i/ \quad \left[\nabla D\left(x^k\right)\right]_i > 0, \quad x_i > 0 \tag{10.23}$$

This maximum step size value being obtained, the value of the descent step size independent of the components, ensuring the convergence of the algorithm, must be calculated by a one-dimensional minimisation method (possibly a simplified method of the Armijo type [5]), within the interval $\left[0, \alpha_M^k\right]$ in the direction:

$$d_k = diag\left[f_i\left(x^k\right)\right] \; diag\left[x_i^k\right] \left[-\nabla D\left(x^k\right)\right] \tag{10.24}$$

This direction, which is no longer that of the opposite of the gradient, remains nevertheless a descent direction for $D\left(x\right)$.
The general algorithm is therefore written:

$$x^{k+1} = x^k + \alpha^k \; d^k \tag{10.25}$$

We can now check that the conditions of KKT (10.15) and (10.16) are fulfilled at optimum, indeed:

- if $x_i^* > 0$, according to (10.19), we have necessarily $\left[\nabla D\left(x^*\right)\right]_i = 0$, that is $\lambda_i^* = 0$; the optimum is the unconstrained minimum, the constraint is inactive.

- if $x_i \approx 0$ ($x_i = \epsilon > 0$), according to (10.19), to attain the constraint it will be necessary to have $\left[\nabla D\left(x\right)\right]_i > 0$ that is $\lambda_i^* > 0$, then, the optimum is located on the constraint which is active.

- of course, one can have simultaneously $x_i^* = 0$ and $\left[\nabla D\left(x^*\right)\right]_i = 0$; then, the optimum is located on the constraint which is active and it also corresponds to the unconstrained minimum.

10.2.2 Standard Algorithms - Non multiplicative

The algorithms of this form that we will be led to use, strictly correspond to the previous method, in which we take:

$$f_i(x) = 1 \quad \forall i \tag{10.26}$$

10.2.3 Multiplicative Algorithms.

In the algorithm defined by (10.24) and (10.25), a particular choice of the function $f_i(x)$ will allow to obtain purely multiplicative algorithms.
To characterize it for this specific purpose, we observe that if at the convergence of the algorithm we have $\nabla D(x^*) = 0$, then, necessarily the gradient can be written as the difference of 2 terms of the same sign (we can even say positive without taking anything away from the generality).
We will therefore write:

$$-\nabla D\left(x^k\right) = U\left(x^k\right) - V\left(x^k\right) \quad ; \quad U\left(x^k\right) > 0 \qquad V\left(x^k\right) > 0 \tag{10.27}$$

If we define now:

$$f_i\left(x^k\right) = \frac{1}{\left[V\left(x^k\right)\right]_i} \tag{10.28}$$

The algorithm (10.19), (10.24), (10.25) can be written:

$$x_i^{k+1} = x_i^k + \alpha_i^k x_i^k \frac{1}{\left[V\left(x^k\right)\right]_i}\left[U\left(x^k\right) - V\left(x^k\right)\right]_i \tag{10.29}$$

Or also:

$$x_i^{k+1} = x_i^k + \alpha_i^k x_i^k \left\{\frac{\left[U\left(x^k\right)\right]_i}{\left[V\left(x^k\right)\right]_i} - 1\right\} \tag{10.30}$$

We can say, even if it is somewhat trivial, that we have substituted the convergence condition $\left[U\left(x^k\right)\right]_i - \left[V\left(x^k\right)\right]_i = 0$ by the condition $\frac{\left[U\left(x^k\right)\right]_i}{\left[V\left(x^k\right)\right]_i} = 1$.
According to (10.22), we will have:

$$\alpha_i^k \leq \frac{1}{1 - \frac{\left[U\left(x^k\right)\right]_i}{\left[V\left(x^k\right)\right]_i}} \quad ; if \quad \left[-\nabla D\left(x^k\right)\right]_i < 0 \quad \Leftrightarrow \quad \left[U\left(x^k\right)\right]_i < \left[V\left(x^k\right)\right]_i \tag{10.31}$$

The values for α_i^k will all necessarily be greater than 1, so does α_M^k, and we have to remember that these are limit values ensuring that the solution is

positive or zero.
If we choose a descent step size $\alpha^k = 1\ \forall k$, the algorithm takes the form "Purely Multiplicative (P.M.)" which is written:

$$x_i^{k+1} = x_i^k \left\{ \frac{\left[U\left(x^k\right)\right]_i}{\left[V\left(x^k\right)\right]_i} \right\} \tag{10.32}$$

On the contrary, the algorithm (10.30) can be interpreted as a relaxed version of (10.32).
It is quite obvious that in (10.32) the successive estimates will be positive if the initial estimate is positive, and that moreover, if at a given iteration, a component of the solution reaches the value zero (i.e. is on the constraint), this component will remain at zero during the following iterations.
However, it should be noted that in general, nothing guarantees the convergence of a purely multiplicative algorithm of type (10.32), only a specific analysis of the divergence $D(x)$ considered will make it possible to ensure it.
Such an analysis has been performed for 2 particular divergences, the root mean square (RMS) divergence that led to the ISRA algorithm [29], [30] and the Kullback-Leibler divergence that led to the E.M. algorithms [32],[58] or Richardson-Lucy [83], [69].
Moreover, we must note that even if the algorithm (10.32) converges, its convergence speed is fixed, because there is no longer any adjustment parameter available. Indeed, for a fixed descent direction, this role is played by the descent step size.
The low convergence speed is one of the drawbacks (frequently mentioned in the literature), of purely multiplicative algorithms of the type (10.32).

10.2.4 Effect of the function expressing the constraints.

We will show that by modifying the function expressing the constraints $g_i(x)$, we can achieve a change in the speed of convergence of the algorithms exhibited in the preceding sections (as long as they converge).
For this aim we choose for example $g_i(x) = x_i^{1/n}$, then, the relation (10.13) is written:

$$\left[x_i^k\right]^{1/n} \left\{ \left[U\left(x^k\right)\right]_i - \left[V\left(x^k\right)\right]_i \right\} = 0 \tag{10.33}$$

or also:

$$x_i^k \left\{ \left[U\left(x^k\right)\right]_i^n - \left[V\left(x^k\right)\right]_i^n \right\} = 0 \tag{10.34}$$

Multiplicative Algorithms.

The equation (10.34) can be written as follows:

$$\frac{x_i^k}{\left[V\left(x^k\right)\right]_i^n}\left\{\left[U\left(x^k\right)\right]_i^n-\left[V\left(x^k\right)\right]_i^n\right\}=0 \quad \Rightarrow \quad x_i^k\left\{\frac{\left[U\left(x^k\right)\right]_i^n}{\left[V\left(x^k\right)\right]_i^n}-1\right\}=0 \tag{10.35}$$

Then, the algorithm (10.19)(10.25) can be written as:

$$x_i^{k+1}=x_i^k+\alpha^k x_i^k\left\{\underbrace{\frac{\left[U\left(x^k\right)\right]_i^n}{\left[V\left(x^k\right)\right]_i^n}-1}_{A}\right\} \tag{10.36}$$

To illustrate the change in the behavior of the algorithm (10.36) with respect to the algorithm (10.30), in connection with the form of $g_i\left(x^k\right)$, we are looking at the corrective term "**A**" which can be developed in the form:

$$\frac{1}{\left[V\left(x^k\right)\right]_i^n}\left\{\sum_{p=0}^{n-1}\left[U\left(x^k\right)\right]_i^{n-p-1}\left[V\left(x^k\right)\right]_i^p\right\}\left\{\left[U\left(x^k\right)\right]_i-\left[V\left(x^k\right)\right]_i\right\}=0 \tag{10.37}$$

Which means that the $f_i\left(x^k\right)$ factor here takes the form:

$$f_i\left(x^k\right)=\frac{1}{\left[V\left(x^k\right)\right]_i^n}\left\{\sum_{p=0}^{n-1}\left[U\left(x^k\right)\right]_i^{n-p-1}\left[V\left(x^k\right)\right]_i^p\right\} \tag{10.38}$$

The expression of $f_i\left(x^k\right)=\frac{1}{\left[V\left(x^k\right)\right]_i}$ (10.28) here is multiplied by:

$$B=\frac{1}{\left[V\left(x^k\right)\right]_i^{n-1}}\left\{\sum_{p=0}^{n-1}\left[U\left(x^k\right)\right]_i^{n-p-1}\left[V\left(x^k\right)\right]_i^p\right\} \tag{10.39}$$

Apart from the obvious change in the direction of descent, we observe that when we are close to convergence, i.e. when $\left[U\left(x^k\right)\right]_i \to \left[V\left(x^k\right)\right]_i$ this change results in $B \to n$; so the descent step size is progressively multiplied by a factor that tends towards "n"; so we can expect a gain of a factor of about "n" on the convergence speed.
In a relaxed algorithm, this effect is counterbalanced by the computation of the descent step size in the modified direction, so there's no risk of divergence

for the algorithm, but then the gain in speed is partially suppressed.
On the other hand, if we consider the algorithm written in the form (10.36) and we choose a descent step equal to 1, we will obtain a purely multiplicative algorithm which is written as follows:

$$x_i^{k+1} = x_i^k \frac{\left[U\left(x^k\right)\right]_i^n}{\left[V\left(x^k\right)\right]_i^n} \tag{10.40}$$

This is the form frequently proposed in the litterature [90], [97], to achieve a convergence speed gain of a "n" factor, but again, as is often the case in fixed step size algorithms, nothing guarantees convergence, and, to put it roughly: if you want to go too fast, you often go to the wall....

Standard Algorithm - Non multiplicative.

In this case, the equation (10.34) can be written:

$$x_i^k \left\{ \sum_{p=0}^{n-1} \left[U\left(x^k\right)\right]_i^{n-p-1} \left[V\left(x^k\right)\right]_i^p \right\} \left\{ \left[U\left(x^k\right)\right]_i - \left[V\left(x^k\right)\right]_i \right\} = 0 \tag{10.41}$$

The fonction $f_i\left(x^k\right) = 1$ is here replaced by the expression:

$$f_i\left(x^k\right) = \sum_{p=0}^{n-1} \left[U\left(x^k\right)\right]_i^{n-p-1} \left[V\left(x^k\right)\right]_i^p \tag{10.42}$$

It should be noted, however, that with such algorithms, for convergence problems, the calculation of the descent step size is indispensable, which somewhat cancels out the effect of acceleration, if any.

10.2.5 Specific case of simplified divergences.

The application of the SGM method and, in general, of the methods described in the previous section, implies that the divergences considered have a minimum (possibly zero) when $p_i = q_i \ \forall i$.
This being the case, in the expressions of certain divergences, simplifications of the type $\sum_i p_i = \sum_i q_i$ have been explicitly introduced, or in the case of comparisons between probability densities, $\sum_i p_i = \sum_i q_i = 1$.
When this type of simplification has been introduced in a divergence, although the divergence considered is zero for $p_i = q_i \ \forall i$, the gradient with respect to "q" may either never cancel out or cancel out for $p_i \neq q_i$, as has

been shown in many examples.
These divergences are thus "a priori" not usable in the context which is ours; however one can nevertheless try to exploit them in spite of the simplifications carried out, by introducing variables which are explicitly of the same sum, for example normalized variables.
We will notice that this operation, whose objective is to enable the use of the simplified divergences, allows to take into account the sum constraint at the same time as the non-negativity constraint.
This is fully understandable because when one tries to make a divergence invariant with respect to "q" by introducing an invariance factor $K^{*}(p,q) = \frac{\sum_j p_j}{\sum_j q_j}$, one obtains a divergence which has the same expression as the initial divergence, but with normalized variables $\bar{p}_i = \frac{p_i}{\sum_j p_j}$ and $\bar{q}_i = \frac{q_i}{\sum_j q_j}$, thus, the normalized variables are used "de facto", so it is not surprising that the use of such variables allows to take into account the sum constraint; this is a kind of justification for the introduction of normalized variables that may seem somewhat artificial.
In order to explicitly account for these normalizations in the simplified $D(p\|q)$ divergences, we will write them down as $D(\bar{p}\|\bar{q})$.
If we denote the measurements by "y_i" and the linear model by $(Hx)_i$, the variables used in the expressions of the simplified divergences will be:

$$\bar{p}_i = \frac{p_i}{\sum_j p_j} C = \frac{y_i}{\sum_j y_j} C \quad ; \quad \bar{q}_i = \frac{q_i}{\sum_j q_j} C = \frac{(Hx)_i}{\sum_j (Hx)_j} C \quad ; \quad C > 0 \tag{10.43}$$

Two cases occur in a classical way (not restrictive): either the data fields are typically probability densities, then $C = 1$, or we deal with deconvolution problems such as those encountered in image deconvolution, then $C = \sum_i y_i$.
Of course, the problem of convexity of the divergence thus modified arises; indeed, taking into account the normalization introduced, if the divergence considered $D(\bar{p}\|\bar{q})$ is convex with respect to " $\bar{q}$ ", is it still convex with respect to "q", therefore with respect to "x"?
One fact is certain: if $\sum_j (q)_j = \sum_j (Hx)_j$ is maintained constant during the minimization process, $D(\bar{p}\|\bar{q})$ will be convex with respect to "x".
We will use the notation:

$$D(\bar{p}\|\bar{q}) = \sum_i d(\bar{p}_i\|\bar{q}_i) = \sum_i d_i \tag{10.44}$$

Then, we have:

$$\frac{\partial D(\bar{p}\|\bar{q})}{\partial q_m} = \sum_i \frac{\partial d(\bar{p}_i\|\bar{q}_i)}{\partial \bar{q}_i} \frac{\partial \bar{q}_i}{\partial q_m} \tag{10.45}$$

With:

$$\frac{\partial \bar{q_i}}{\partial q_m} = \frac{C}{\sum_j q_j}\left(\delta_{im} - \frac{\bar{q_i}}{C}\right) \tag{10.46}$$

So it comes:

$$\frac{\partial D\left(\bar{p}\|\bar{q}\right)}{\partial q_m} = \frac{C}{\sum_j q_j}\left[\frac{\partial d\left(\bar{p_m}\|\bar{q_m}\right)}{\partial \bar{q_m}} - \sum_i \frac{\bar{q_i}}{C}\frac{d\left(\bar{p_i}\|\bar{q_i}\right)}{\partial \bar{q_i}}\right] \tag{10.47}$$

If we now take into account the fact that we are using a linear model, we have:

$$\frac{\partial D\left(\bar{p}\|\bar{q}\right)}{\partial x_l} = \sum_j \frac{\partial D\left(\bar{p}\|\bar{q}\right)}{\partial q_j}\frac{\partial q_j}{\partial x_l} = \sum_j h_{jl}\frac{\partial D\left(\bar{p}\|\bar{q}\right)}{\partial q_j} \tag{10.48}$$

so:

$$-\frac{\partial D\left(\bar{p}\|\bar{q}\right)}{\partial x_l} = \frac{C}{\sum_j q_j}\left\{\sum_i h_{i,l}\left[-\frac{\partial d_i\left(\bar{p}\|\bar{q}\right)}{\partial \bar{q_i}}\right] - \left(\sum_j h_{j,l}\right)\sum_i \frac{\bar{q_i}}{C}\left[-\frac{\partial d_i\left(\bar{p}\|\bar{q}\right)}{\partial \bar{q_i}}\right]\right\} \tag{10.49}$$

Or also:

$$-\frac{\partial D\left(\bar{p}\|\bar{q}\right)}{\partial x_l} = \frac{C}{\sum_j \left(Hx\right)_j}\left\{\left[H^T - \left(\sum_j h_{j,l}\right)\frac{\bar{Q}}{C}\right]\left[-\frac{\partial D\left(\bar{p}\|\bar{q}\right)}{\partial \bar{q}}\right]\right\}_l \tag{10.50}$$

In this expression, $\overline{Q}$ is a matrix in which all the lines are identical and equal to $[\bar{q_1}\ \bar{q_2}\ .....\bar{q_N}]$ with $\sum_i \bar{q_i} = C$ and $\left[\frac{\partial D}{\partial \bar{q}}\right]$ is the vector of partial derivatives $\left[\frac{\partial d_i}{\partial \bar{q_i}}\right]$.
Another writting may be more appealing, considering Richardson's [83] and Lucy's [69] deconvolution algorithms (for example):

$$-\frac{\partial D\left(\bar{p}\|\bar{q}\right)}{\partial x_l} = C\frac{\sum_j h_{j,l}}{\sum_j q_j}\left\{\left[\overline{H}^T - \frac{\overline{Q}}{C}\right]\left[-\frac{\partial D\left(\bar{p}\|\bar{q}\right)}{\partial \bar{q}}\right]\right\}_l \tag{10.51}$$

Here, $\overline{H}$ is a matrix whose columns are of sum 1.

Algorithm.

From the above considerations, we can write an SGM-type algorithm in the form:

$$x_l^{k+1} = x_l^k + \delta^k x_l^k\left[-\frac{\partial D\left(\bar{p}\|\bar{q}\right)}{\partial x_l}\right] \tag{10.52}$$

that is:

$$x_l^{k+1} = x_l^k + \delta^k x_l^k C \frac{\sum_j h_{j,l}}{\sum_j q_j^k} \left\{ \left[\overline{H}^T - \frac{Q^k}{C} \right] \left[-\frac{\partial D\left(\bar{p}\|\bar{q}\right)}{\partial \bar{q}^k} \right] \right\}_l \tag{10.53}$$

or also:

$$x_l^{k+1} = x_l^k + \delta^k \left\{ x_l^k C \frac{\sum_j h_{j,l}}{\sum_j q_j^k} \left[\overline{H}^T \left[-\frac{\partial D\left(\bar{p}\|\bar{q}\right)}{\partial \bar{q}^k} \right] \right]_l - x_l^k \frac{\sum_j h_{j,l}}{\sum_j q_j^k} \left[\overline{Q}^k \left[-\frac{\partial D\left(\bar{p}\|\bar{q}\right)}{\partial \bar{q}^k} \right] \right]_l \right\} \tag{10.54}$$

With this algorithm, we have:

$$\sum_l x_l^{k+1} = \sum_l x_l^k \tag{10.55}$$

To show this, we calculate: $\sum_l x_l^{k+1}$ and we look at what are becoming the terms between the braces of (10.54).
For the first of these terms, we have:

$$\sum_l x_l^k C \frac{\sum_j h_{j,l}}{\sum_j q_j^k} \left[\overline{H}^T \left[-\frac{\partial D}{\partial \bar{q}^k} \right] \right]_l = \frac{C}{\sum_j q_j^k} \sum_l q_l^k \left[-\frac{\partial D}{\partial \bar{q}^k} \right]_l = \sum_l \bar{q}_l^k \left[-\frac{\partial D}{\partial \bar{q}^k} \right]_l \tag{10.56}$$

For the second term of the brace, given the properties of the matrix $\overline{Q}$ in which all the lines are identical, the term:

$$\left[\overline{Q}^k \left[-\frac{\partial D}{\partial \bar{q}^k} \right] \right]_l = \left[\sum_i \bar{q}_i^k \left[-\frac{\partial D}{\partial \bar{q}^k} \right]_i \right]_l \tag{10.57}$$

is independent of "l"; therefore, we can write:

$$\sum_l x_l^k \frac{\sum_j h_{j,l}}{\sum_j q_j^k} \left[\overline{Q}^k \left[-\frac{\partial D}{\partial \bar{q}^k} \right] \right]_l = \frac{\sum_j \sum_l h_{j,l} x_l^k}{\sum_j q_j^k} \sum_i \bar{q}_i^k \left[-\frac{\partial D}{\partial \bar{q}^k} \right]_i = \sum_i \bar{q}_i^k \left[-\frac{\partial D}{\partial \bar{q}^k} \right]_i \tag{10.58}$$

To summarize, if we are returning to the equation (10.54), the term between braces will give in the sum, a null contribution, and we will have:

$$\sum_l x_l^{k+1} = \sum_l x_l^k \tag{10.59}$$

In conclusion, the algorithm (10.53),(10.54) ensures that the sum constraint is satisfied. This is a consequence of the introduction of the variables in the form (10.43).

However, it should be noted that the relation (10.59) is absolutely not equivalent to:

$$\sum_l \left(Hx^k\right)_l = \sum_l \left(Hx^{k+1}\right)_l \tag{10.60}$$

so that the convexity of the initial divergence from the "x" variable is still not assured unless the H matrix is column-normalized.
If this is the case, the relations (10.59) and (10.60) become equivalent; this property will be reused in the following section.

Purely multiplicative form of the algorithm.

Given the properties of $\overline{H}$ and of the matrix $\frac{\overline{Q}}{C}$ (for each line, the sum of the terms is equal to 1), in the relation (10.53), we can make $-\frac{\partial D}{\partial \bar{q}} \geq 0$ without changing the expression of the gradient, by adding a constant $(-\min_i \left[-\frac{\partial d_i}{\partial \bar{q}_i}\right])$.
We denote $\left[-\frac{\partial D}{\partial \bar{q}}\right]_d$ the shifted vector.

$$\left[-\frac{\partial D}{\partial \bar{q}}\right]_d = -\frac{\partial D}{\partial \bar{q}} - \min_i \left[-\frac{\partial d_i}{\partial \bar{q}_i}\right] + \epsilon \tag{10.61}$$

We can therefore envisage a decomposition of the gradient into a difference of 2 positive terms.
From this remark, in the case of the simplified divergences considered here, one can write an algorithm in the form:

$$x_l^{k+1} = x_l^k + \delta^k x_l^k \left\{ \left[\overline{H}^T - \frac{\overline{Q}^k}{C}\right] \left[-\frac{\partial D}{\partial \bar{q}^k}\right]_d \right\}_l \tag{10.62}$$

Or also:

$$x_l^{k+1} = x_l^k + \delta^k x_l^k \left\{ \left[\overline{H}^T \left(-\frac{\partial D}{\partial \bar{q}^k}\right)_s\right]_l - \sum_i \frac{\bar{q}_i^k}{C} \left(-\frac{\partial d_i}{\partial \bar{q}_i^k}\right)_d \right\} \tag{10.63}$$

It has already been noted that the second term of the expression between braces is a constant independent of the component "l".
As long as $\sum_i \bar{q}_i^k \left(-\frac{\partial d_i}{\partial \bar{q}_i^k}\right)_d \neq 0$, which is generally the case, and with a descent step size $\delta_l = 1$, we can exhibit a purely multiplicative form:

$$x_l^{k+1} = x_l^k \frac{\left[\overline{H}^T \left[-\frac{\partial D}{\partial \bar{q}^k}\right]_d\right]_l}{\sum_i \frac{\bar{q}_i^k}{C} \left(-\frac{\partial d_i}{\partial \bar{q}_i^k}\right)_d} = x_l^k \frac{C \left[H^T \left[-\frac{\partial D}{\partial \bar{q}^k}\right]_d\right]_l}{\sum_j h_{j,l} \quad \sum_i \bar{q}_i^k \left(-\frac{\partial d_i}{\partial \bar{q}_i^k}\right)_d} \tag{10.64}$$

At this point, if we calculate $\sum_l x_l^{k+1}$, we obtain nothing but a weighted sum that is difficult to interpret, indeed, $\sum_j h_{j,l}$ depends on the "l" component, on the other hand, if we make the assumption of a matrix H normalized to 1 in columns, i.e. $\sum_j h_{j,l} = 1 \; \forall l$, (which is the case in classical deconvolution problems), we arrive at something simpler which is written as follows:

$$\sum_l x_l^{k+1} = \sum_l \left(Hx^k\right)_l = \sum_l x_l^k \tag{10.65}$$

Consequently, by introducing the variables in the form (10.43), we also obtained a purely multiplicative algorithm making it possible to fulfill the sum constraint, at least in the case where the H matrix is normalized in columns, i.e. in the case of the usual convolution with a positive kernel with integral equal to 1.

10.3 Introduction of the sum constraint.

In this section, we will rely on the developments of the previous section and introduce an additional constraint which is the sum constraint on the unknowns "x_i". Several methods are possible to introduce this constraint; their use depends essentially on the properties of the divergences considered.

10.3.1 Simplified divergences.

The case of these divergences has been addressed in the previous paragraph. Indeed, the introduction of an invariance factor equal to

$$K^*(p,q) = \frac{\sum_j p_j}{\sum_j q_j} \tag{10.66}$$

allows to show simplified divergences in which the variables are explicitly of the same sum. Moreover, we observe that (with the exception of a multiplicative factor which depends only on $\sum_j p_j$ and which can be omitted), these divergences are not only invariant with respect to the variable "q", but also with respect to the variable "p".
This property will make it possible to use this type of divergence for regularization problems in which the both arguments of the divergence depends on the true variable "x".

10.3.2 Non-simplified divergences.

This paragraph concerns the Csiszär divergences built on standard convex functions as well as the Bregman divergences and the Jensen divergences.
With these divergences, we will have $p_i = y_i$ and $q_i = (Hx)_i$.
The algorithms for minimizing these divergences under non-negativity constraint are derived from KKT conditions and can result in multiplicative algorithms.
The sum constraint $\sum_i x_i = C, \;\; C > 0$ can be taken into account by introducing the variable change $x_i = \frac{u_i}{\sum_j u_j} C$.
The problem that arises then is that of the convexity of the divergence to be minimized with respect to the new variable "u".
An answer to this problem can be the following: the objective function (i.e. the considered divergence) being supposed to be convex with respect to "q", therefore with respect to "x" since $q = Hx$, it will be convex with respect to "u" if $\sum_j u_j$ is constant during the minimization process, i.e. in the course of the iterations.
We must therefore write an algorithm of minimization with respect to "u" of the divergence $D(p\|q) = D(y\|Hx)$ with: $x_i = \frac{u_i}{\sum_j u_j} C$; this algorithm must be such that during the iterations $\sum_j u_j^k = Cst$.
Using the notation:

$$D(p\|q) = \sum_i d(p_i\|q_i) \tag{10.67}$$

we will have:

$$\frac{\partial D}{\partial x_m} = \sum_i \frac{\partial d(p_i\|q_i)}{\partial q_i} \frac{\partial q_i}{\partial x_m} = \left[H^T \frac{\partial D(p\|q)}{\partial q} \right]_m \tag{10.68}$$

Then the gradient of D with respect to "u" is written as follows:

$$\left[\frac{\partial D}{\partial u} \right]_l = \sum_m \frac{\partial D}{\partial x_m} \frac{\partial x_m}{\partial u_l} \tag{10.69}$$

With:

$$\frac{\partial x_m}{\partial u_l} = \frac{1}{\sum_j u_j} (C \; \delta_{l\,m} - x_l) \tag{10.70}$$

we have:

$$\left[\frac{\partial D}{\partial u} \right]_l = \frac{1}{\sum_j u_j} \left[C \; \frac{\partial D}{\partial x_l} - \sum_m x_m \frac{\partial D}{\partial x_m} \right] \tag{10.71}$$

and we have, immediately:

$$\sum_l u_l \left[-\frac{\partial D}{\partial u}\right]_l = 0 \tag{10.72}$$

This relationship is the key point of the affair.

Indeed, the iterative minimization algorithm with respect to "u", founded on the KKT conditions, ensuring non-negativity is written in the form (10.19):

$$u_l^{k+1} = u_l^k + \delta^k u_l^k \left[-\frac{\partial D}{\partial u^k}\right]_l \tag{10.73}$$

Then, with (10.72) we will have:

$$\sum_l u_l^{k+1} = \sum_l u_l^k \tag{10.74}$$

In the course of the iterations, the trajectory of the solutions will thus be contained in the convexity domain of the objective function; one can thus return to the initial variable "x" by dividing the 2 members of (10.73) by $\sum_l u_l^{k+1} = \sum_l u_l^k$, and we will have:

$$x_l^{k+1} = x_l^k + \delta^k \frac{1}{\sum_j u_j^k} x_l^k \left[C \left(-\frac{\partial D}{\partial x^k}\right)_l - \sum_m x_m^k \left(-\frac{\partial D}{\partial x^k}\right)_m\right] \tag{10.75}$$

Or also:

$$x_l^{k+1} = x_l^k + \zeta^k x_l^k \left[C \left(-\frac{\partial D}{\partial x^k}\right)_l - \sum_m x_m^k \left(-\frac{\partial D}{\partial x^k}\right)_m\right] \tag{10.76}$$

Note that in this algorithm the second term in square brackets is a constant that is independent of the component considered; moreover, the quantity:

$$\zeta^k = \delta^k \frac{1}{\sum_j u_j^k} \tag{10.77}$$

represents globally a descent step size which is calculated at each iteration by a one-dimensional minimization procedure as indicated for the S.G.M. method.

Taking into account the relation $\sum_i x_i = C$, we will thus have at each iteration, independently of the descent step size:

$$\sum_l x_l^{k+1} = \sum_l x_l^k \tag{10.78}$$

To fulfill the sum constraint at each iteration, it is thus only necessary to set an initial estimate "x^0" such that $\sum_i x_i^0 = C$.

*** Multiplicative form.**

From the algorithm, a multiplicative form can be derived.
To do this, we shift the components of the opposite of the gradient to make them all positive:

$$\left[-\frac{\partial D}{\partial x_j^k}\right]_d = \left[-\frac{\partial D}{\partial x_j^k}\right] - \min_l \left[-\frac{\partial D}{\partial x_l^k}\right] + \epsilon \tag{10.79}$$

The introduction of this offset doesn't change the algorithm.
We can then propose an algorithm that will be the basis of the purely multiplicative form, which is written:

$$x_l^{k+1} = x_l^k + \zeta^k x_l^k \left[\frac{C\ \left(-\frac{\partial D}{\partial x_l^k}\right)_d}{\sum_m x_m^k \left(-\frac{\partial D}{\partial x_m^k}\right)_d} - 1\right] \tag{10.80}$$

With this algorithm, the sum-holding property is preserved over iterations, regardless of the descent step size.
The purely multiplicative form will, as always, be obtained by using a descent step size $\zeta^k = 1\ \forall k$, but then convergence is no longer guaranteed.

10.3.3 Divergences invariant by change of scale.

For these divergences, we rely on the method developed in the previous sections to build a minimization algorithm that takes into account the non-negativity constraint.
As far as the summation constraint is concerned, we have two methods at our disposition:
* Either we normalize at each iteration which does not lead to any variation of the objective function taking into account the invariance property.
* Either we rely on a specific property of invariant divergences we previously established.

*** Property:** Considering a divergence $D(p\|q)$, which we render invariant by scale change using any invariance factor, whether it corresponds to it ("*nominal*" invariance factor) or not, we obtain a divergence $DI(p\|q)$, for which we always have:

$$\sum_l q_l \frac{\partial DI(p\|q)}{\partial q_l} = 0 \tag{10.81}$$

This property is analogous to(10.72).

Example: Mean square deviation.

The basic divergence is:

$$MC = \sum_i (p_i - q_i)^2 \tag{10.82}$$

By introducing the invariance factor K, we have:

$$MCI = \sum_i (p_i - Kq_i)^2 \tag{10.83}$$

The gradient with respect to "q_j" is written, all simplifications made:

$$\frac{\partial MCI}{\partial q_j} = -2K(p_j - Kq_j) - 2\frac{\partial K}{\partial q_j}\sum_i q_i(p_i - Kq_i) \tag{10.84}$$

Then, we have:

$$\sum_j q_j \frac{\partial MCI}{\partial q_j} = -2\left[\sum_j q_j(p_j - Kq_j)\right]\left(K + \sum_j q_j \frac{\partial K}{\partial q_j}\right) \tag{10.85}$$

Then, if we use the nominal invariance factor, i.e.:

$$K_0(p,q) = \frac{\sum_l p_l q_l}{\sum_l q_l^2} \tag{10.86}$$

which is a solution to the differential equation:

$$K + \sum_j q_j \frac{\partial K}{\partial q_j} = 0 \tag{10.87}$$

the expression of the gradient (10.84) becomes simpler because the second term is zero.
We still have the property (10.81).
Similarly, if we use another expression of K, for example:

$$K^*(p,q) = \frac{\sum_l p_l}{\sum_l q_l} \tag{10.88}$$

which is another solution of (10.87), we also obtain an invariant divergence, the expression of the gradient (10.84) does not simplify, but we still have the property (10.81).

Some comments on the use of $K^*(p,q)$.

We recall that when a divergence $D(p\|q)$ is made invariant by using the invariance factor $K^*(p,q)$, the divergence obtained is similar to the initial divergence in which the variables "p" and "q" are replaced by normalized variables $\bar{p}_i = \frac{p_i}{\sum_j p_j}$ and $\bar{q}_i = \frac{q_i}{\sum_j q_j}$, possibly with a multiplicative scalar factor which depends only on the measures "p_i" and which can be omitted; simplifications can then appear.
Therefore, the invariant divergence $DI(p\|q)$ obtained with this particular invariance factor is denoted as $D(\bar{p}\|\bar{q})$ with:

$$D(\bar{p}\|\bar{q}) = \sum_i d(\bar{p}_i\|\bar{q}_i) \tag{10.89}$$

and the gradient with respect to "q" is written as follows:

$$\frac{\partial DI(p\|q)}{\partial q_l} = \frac{\partial D(\bar{p}\|\bar{q})}{\partial q_l} \tag{10.90}$$

It is calculated as follows:

$$\frac{\partial D(\bar{p}\|\bar{q})}{\partial q_l} = \sum_i \frac{\partial d(\bar{p}_i\|\bar{q}_i)}{\partial q_l} = \sum_i \frac{\partial d(\bar{p}_i\|\bar{q}_i)}{\partial \bar{q}_i}\frac{\partial \bar{q}_i}{\partial q_l} \tag{10.91}$$

With:

$$\frac{\partial \bar{q}_i}{\partial q_l} = \frac{\delta_{il}}{\sum_j q_j} - \frac{\bar{q}_i}{\sum_j q_j} \tag{10.92}$$

So:

$$\frac{\partial D(\bar{p}\|\bar{q})}{\partial q_l} = \frac{1}{\sum_j q_j}\left[\frac{\partial d(\bar{p}_l\|\bar{q}_l)}{\partial \bar{q}_l} - \sum_i \bar{q}_i \frac{\partial d(\bar{p}_i\|\bar{q}_i)}{\partial \bar{q}_i}\right] \tag{10.93}$$

We can observe that the classical relationship for invariant divergences:

$$\sum_l q_l \frac{\partial DI(p\|q)}{\partial q_l} = \sum_l q_l \frac{\partial D(\bar{p}\|\bar{q})}{\partial q_l} = 0 \tag{10.94}$$

is fulfilled.

Algorithm:
In the context of the use of a linear model $q = Hx$, the calculation of the gradient with respect to "x", true unknowns of the problem, follows by the relation:

$$\frac{\partial DI(p\|q)}{\partial x_l} = \left[H^T \frac{\partial DI(p\|q)}{\partial q}\right]_l \tag{10.95}$$

Given this relationship, the algorithms for minimizing scale invariant divergences under non-negativity and sum-of-unknowns constraints can be written directly according to the method indicated above, in the form of:

$$x_l^{k+1} = x_l^k + \alpha^k x_l^k \left[-H^T \frac{\partial DI(p\|q^k)}{\partial q^k} \right]_l \tag{10.96}$$

Non-negativity and convergence are ensured by a computation of the descent step size, but whatever its value, we will always have:

$$\sum_l x_l^{k+1} = \sum_l x_l^k \tag{10.97}$$

Indeed:

$$\sum_l x_l^{k+1} = \sum_l x_l^k + \alpha^k \sum_l \left\{ x_l^k \left[H^T \left(-\frac{\partial DI\,(p\|q)}{\partial q^k} \right) \right]_l \right\} \tag{10.98}$$

Which can be written:

$$\sum_l x_l^{k+1} = \sum_l x_l^k + \alpha^k \sum_l \left\{ \left(Hx^k \right)_l \left(-\frac{\partial DI\,(p\|q)}{\partial q^k} \right)_l \right\} \tag{10.99}$$

And then:

$$\sum_l x_l^{k+1} = \sum_l x_l^k + \alpha^k \sum_l \left\{ q_l^k \left(-\frac{\partial DI\,(p\|q)}{\partial q^k} \right)_l \right\} \tag{10.100}$$

Taking into account the property (10.81) we have:

$$\sum_l x_l^{k+1} = \sum_l x_l^k \tag{10.101}$$

It is therefore enough to fix the sum of the components of the initial estimate to maintain this sum during the iterative process.
Moreover, as for all the algorithms proposed in the preceding sections, **the maximum step size α_M^k ensuring non-negativity is calculated first**, then the descent step size ensuring the convergence of the algorithm is calculated at each iteration by a one-dimensional search method in the interval $[0, \alpha_M^k]$.
On the other hand, it should be noted that a purely multiplicative algorithm deduced from the SGM method would not spontaneously possess this property of maintaining the sum.

However, nothing prevents us from proposing such a purely multiplicative form. Indeed, for all the divergences considered, we can observe (although there is no formal demonstration of this property) that the opposite of the gradient is always written as the difference of 2 positive terms, i.e.:

$$\left[-H^T \frac{\partial DI(p\|q^k)}{\partial q^k}\right]_l = U_l^k - V_l^k \quad ; U_l^k > 0 \quad ; V_l^k > 0 \tag{10.102}$$

Then, following the S.G.M. method, we can propose a multiplicative algorithm obtained using a descent step size $\alpha^k = 1 \ \ \forall k$ written as:

$$x_l^{k+1} = x_l^k \left[\frac{U^k}{V^k}\right]_l \tag{10.103}$$

Of course, there is no evidence of convergence of such algorithms for all the divergences, since the descent step size is fixed, however, the non-negativity constraint is satisfied for a positive initial estimate.
On the other hand, the sum constraint, which is not automatic as in algorithms of the form (10.96), can be ensured here by a normalization at each iteration; indeed, taking into account the invariance property of the divergences considered, this operation does not modify the value of the objective function i.e. of the divergence.

chapter 11 - Applications to the Non Negative Matrix Factorization.

11.1 Introduction.

In this chapter, there is no attempt to repeat the work dealing with factorization in non-negative matrices (N.M.F.); a very extensive bibliography on this subject can be found in [25] and in [15].
Instead, we will develop algorithms that take into account the sum constraints involved in NMF.
In addition, we will demonstrate the interest of using scale invariant divergences; indeed, some properties of these divergences allow to take into account easily sum constraints. These properties will be established, and their influence on the corresponding algorithms will be shown.
Regularized algorithms are also considered, whether the divergences are invariant or not.
Some problems related to purely multiplicative regularized algorithms as proposed in the literature have been reported in **Appendix 9** and it will be explained why such problems do not appear in the methods proposed here.

11.1.1 Linear unmixing.

The problem is presented in the context of hyperspectral data.
At the simplest level, which is the **linear unmixing problem**, we have the measurement of a spectrum (1 only). The measured data are therefore the intensity values at different wavelengths, arranged as a vector "y"; this can

be seen as the decomposition over wavelengths of the intensity in 1 pixel of an image. The sum of these intensities thus represents the total intensity "a" in the pixel considered $\sum_i y_i = a$.
On the other hand, we have a certain number of spectra of simple elements (endmembers); each of these spectra is arranged in a column (vector), the juxtaposition of these columns forming a table that can be considered as a matrix H.
The problem is to find the weights "x" (positive or zero, of course) such that the observed spectrum is described as a weighted sum of the basic spectra, that is, of endmembers.
So we have to solve with respect to "x" a problem that is written in a matrix form:

$$y = Hx \tag{11.1}$$

Furthermore, we would want that these unknown coefficients sum to 1, in order to obtain percentages of each of the elementary spectra in the measured spectrum.
Without further details on the elementary spectra available (for example the integrals of these spectra), the problem seems to me to be unsolvable; for a solution to be envisaged, all the spectra considered, both the measured and the reference spectra, would have to be of the same integral (i.e. the corresponding vectors would have to be of the same sum).
One possibility is to use reference spectra normalised to 1, and to impose as a constraint on the sum of the weighting coefficients $\sum_i x_i = \sum_i y_i$, the percentages being then easily obtained.
Another solution is to use reference spectra normalized to 1, and to normalize the measured spectrum to 1, then the constraint on the sum of the weighting coefficients at 1 is immediate.
In both cases, the common point is the normalization of the reference spectra, i.e. the sum to 1 of the columns of the H matrix.
However, if we perform a simulation, that is, if we generate the measured spectrum as a weighted sum of elementary spectra with weighting coefficients of sum 1, we can hope to find a solution to the inverse problem, whatever the integrals of the reference spectra, but then we are very far from the real problem.
The real case that comes closest to this situation is the one where the "endmembers" with the "right" properties are extracted from the data prior to solving the inverse problem.
Finally, in a real case, an additional problem may arise, because if we have a very large set of reference spectra, even if we assume that all integrals prob-

lems are solved, we may think that all reference spectra will not necessarily contribute to the measured spectrum, then the corresponding weighting coefficients will be zero, so we have to consider that the solution of the problem i.e. the vector of weighting coefficients has a "**sparse**" structure, which introduces an additional constraint.

11.1.2 Multispectral case.

Here, the problem becomes a little more complicated.
Suppose we have an image, the result of an observation; the intensity in each of the pixels of the image is decomposed according to wavelengths as in the previous section, so we will have as many measured spectra as there are pixels in the image; each of these spectra (vector) constitutes a column of a table that represents the measurements. For the moment, it is only a table (not yet a matrix).
Note that the sum of the components in a column represents the total intensity in the corresponding pixel and there is no reason why all these values should be equal.
Knowing as in the previous section a set of reference spectra (column vectors) juxtaposed to form a matrix H, the problem is thus to find for each measured spectrum, the vector (column) of the weighting coefficients.
This is just a succession of problems similar to the one described in the previous section.
If, now, to see the problem as a whole, the (non-negative) weights are arranged in column vectors, the juxtaposition of these columns forms a table X which can be called a matrix and treated as such.
The problem can now be written in matrix form:

$$Y = H\ X \tag{11.2}$$

Where Y is of dimension $(L * C)$; "L" is the number of wavelengths, "C" is the number of pixels in the observed image,
H is of dimension $(L * M)$; "M" is the number of reference spectra,
X is $(M * C)$.
Obviously, "L" and "C" are fixed by the experimental conditions from which the measured spectra are obtained, while "M" depends on the number of reference spectra that are available.
If the columns of the matrix H are normalized to 1, then we have $\sum_i Y_{ij} = \sum_i X_{ij},\ \forall j$; this last inequality will constitute a constraint of the problem.

11.1.3 Case of the NMF.

Here, things get even more complicated, indeed, in this case, we only have the Y measurements and we're looking for both the reference spectra matrix H and the weighting coefficients matrix X.
The first difficulty which appears in an obvious way is the multiplicity of the solutions, indeed, the relation $Y = HX$ can be written $Y = HDD^{-1}X$, with D invertible with non-negative terms, and any decomposition of the form $\widetilde{H} = HD$, $\widetilde{X} = D^{-1}X$ also provides a solution.
From this point of view, the introduction of constraints such as $\sum_i H_{ij} = 1 \ \forall j$ and $\sum_i X_{ij} = \sum_i Y_{ij} \ \forall j$, helps to reduce the ambiguity.
However, an appropriate choice of "M" is generally critical, this choice is of course problem dependent; nevertheless, "M" is often chosen such that $M << \min(L, C)$, which is singularly imprecise. In Lee and Seung's article [67] another rule is proposed, it is written $M < \frac{LC}{L+C}$, which corresponds to the rule *"number of data > number of unknowns"* which can be understood in the context of systems of linear equations.
Finally, it's clear that the product HX is only an approximation of Y to the rank "M".

11.2 Generalities.

All the problems previously mentioned imply looking for the solution of a minimization problem with respect to the unknown parameters, a discrepancy between the measures $Y \equiv P$ and the linear model $HX \equiv Q$. By noting $D(P\|Q)$ the divergence between the tables (matrices) P and Q:

$$D(P\|Q) = \sum_{ij} D(P_{ij}\|Q_{ij}) = \sum_{ij} D_{ij} \tag{11.3}$$

With for our applications:

$$P_{ij} = Y_{ij}; \quad Q_{ij} = [HX]_{ij} = \sum_l H_{il} X_{lj} \tag{11.4}$$

The minimization techniques proposed in the literature and used in this book require the calculation of the gradient of this divergence with respect to the elements of the matrix H as well as with respect to the elements of the matrix X.

*** Gradient with respect to the elements of the matrix X.**

By introducing the matrix $[A]$ such that:

$$[A]_{ij} = \left[\frac{\partial D}{\partial Q}_{ij}\right] \tag{11.5}$$

We can write in matrix form:

$$\left[\frac{\partial D\left(P\|Q\right)}{\partial X}\right]_{nm} = \frac{\partial D\left(P\|Q\right)}{\partial X_{nm}} = \left[H^T A\right]_{nm} = \sum_l H^T_{nl}\left[\frac{\partial D}{\partial Q}_{lm}\right] \tag{11.6}$$

- Details of the calculation.

The gradient is derived as follows:

$$\frac{\partial D\left(P\|Q\right)}{\partial X_{nm}} = \sum_{ij} \frac{D\left(Y_{ij}\|\left[HX\right]_{ij}\right)}{\partial\left[HX\right]_{ij}} \frac{\partial\left[HX\right]_{ij}}{\partial X_{nm}} \tag{11.7}$$

With:

$$\frac{\partial\left[HX\right]_{ij}}{\partial X_{nm}} = \sum_l H_{il}\frac{\partial X_{lj}}{\partial X_{nm}} = \sum_l H_{il}\delta_{ln}\delta_{jm} \tag{11.8}$$

We have:

$$\frac{\partial\left[HX\right]_{ij}}{\partial X_{nm}} = H_{in}\delta_{jm} \tag{11.9}$$

Then:

$$\frac{\partial D\left(P\|Q\right)}{\partial X_{nm}} = \sum_{ij} \frac{\partial D\left(Y_{ij}\|\left[HX\right]_{ij}\right)}{\partial\left[HX\right]_{ij}} H_{in}\delta_{jm} \tag{11.10}$$

And finally:

$$\frac{\partial D\left(P\|Q\right)}{\partial X_{nm}} = \sum_i \frac{\partial D\left(Y_{im}\|\left[HX\right]_{im}\right)}{\partial\left[HX\right]_{im}} H_{in} \tag{11.11}$$

Or also:

$$\frac{\partial D\left(P\|Q\right)}{\partial X_{nm}} = \sum_i H^T_{ni}\frac{\partial D\left(Y_{im}\|\left[HX\right]_{im}\right)}{\partial\left[HX\right]_{im}} = \sum_i H^T_{ni}\left[\frac{\partial D}{\partial Q}_{im}\right] \tag{11.12}$$

*** Gradient with respect to the elements of the matrix H.**

Using the matrix $[A]$ (11.5), we will have in matrix form:

$$\left[\frac{\partial D\left(P\|Q\right)}{\partial H}\right]_{nm} = \frac{\partial D\left(P\|Q\right)}{\partial H_{nm}} = \left[AX^T\right]_{nm} = \sum_j \left[\frac{\partial D}{\partial Q_{nj}}\right] X^T_{jm} \tag{11.13}$$

- Details of the calculation.

The calculation is as follows:

$$\frac{\partial D\left(P\|Q\right)}{\partial H_{nm}}=\sum_{ij}\frac{\partial D\left(Y_{ij}\|\left[HX\right]_{ij}\right)}{\partial\left[HX\right]_{ij}}\frac{\partial\left[HX\right]_{ij}}{\partial H_{nm}} \tag{11.14}$$

We have:

$$\frac{\partial\left[HX\right]_{ij}}{\partial H_{nm}}=\sum_{l}\frac{\partial H_{il}}{\partial H_{nm}}X_{lj}=\sum_{l}\delta_{in}\delta_{lm}X_{lj} \tag{11.15}$$

Then:

$$\frac{\partial\left[HX\right]_{ij}}{\partial H_{nm}}=\delta_{in}X_{mj} \tag{11.16}$$

So:

$$\frac{\partial D\left(P\|Q\right)}{\partial H_{nm}}=\sum_{ij}\frac{\partial D\left(Y_{ij}\|\left[HX\right]_{ij}\right)}{\partial\left[HX\right]_{ij}}\delta_{in}X_{mj} \tag{11.17}$$

And finally:

$$\frac{\partial D\left(P\|Q\right)}{\partial H_{nm}}=\sum_{j}\frac{\partial D\left(Y_{nj}\|\left[HX\right]_{nj}\right)}{\partial\left[HX\right]_{nj}}X_{mj} \tag{11.18}$$

Or also:

$$\frac{\partial D\left(P\|Q\right)}{\partial H_{nm}}=\sum_{j}\frac{\partial D\left(Y_{nj}\|\left[HX\right]_{nj}\right)}{\partial\left[HX\right]_{nj}}X_{jm}^{T}=\sum_{j}\left[\frac{\partial D}{\partial Q_{nj}}\right]X_{jm}^{T} \tag{11.19}$$

11.3 Algorithmic.

11.3.1 Principle of the method.

The unknowns of the problem are the matrices H and X of the product $HX\equiv Q$, the data are the matrix $Y\equiv P$. We look for H and X by minimizing a divergence between the matrices Y and HX which is written as follows:

$$X;\ H;=\underbrace{Arg\ min}_{X,H}\ D\left(Y\|\left[HX\right]\right)=\underbrace{Arg\ min}_{X,H}\sum_{j}\sum_{i}D\left(Y_{ij}\|\left[HX\right]_{ij}\right) \tag{11.20}$$

We can notice that we so write a divergence between the columns of the two matrices, and that we then sum over all the columns.

The method of minimization generally proposed is an iterative method in which one operates alternately on the two unknowns which can be summarized for example according to the diagram:
1 - Iteration on H: X^k , $H^k \rightarrow X^k$, H^{k+1}
2 - Iteration on X: X^k , $H^{k+1} \rightarrow X^{k+1}$, H^{k+1}
It is this flowchart that will be considered in the rest of the computations, but the order of update is not mandatory.
In this diagram, one point remains to be clarified: is the updating done element by element, or column by column?
From my point of view, given the constraints imposed on the columns of H and X, it is obvious to operate column by column.
Indeed, the constraints imposed on the unknowns are as follows:
* $H_{ij} \geq 0 \quad \forall i, j.$
* $X_{ij} \geq 0 \quad \forall i, j.$
* $\sum_i H_{ij} = 1 \quad \forall j.$
* $\sum_i X_{ij} = \sum_i Y_{ij} \quad \forall j.$
If the columns of the matrix Y are normalized to 1, the sum constraint on X becomes:
* $\sum_i X_{ij} = 1 \quad \forall j.$
The main difficulty encountered in this problem is that even if the divergence considered is separately convex in H and X, it is not jointly convex with respect to the two unknowns; consequently, there is no guarantee that the absolute minimum is reached by the proposed algorithmic method, one can only say that a local minimum of the objective function is reached, therefore a solution that is not necessarily optimal.

* An alternative approach.

In the problems of searching for saddle points in constrained optimization based on Lagrangian methods, we are led to operate on primary variables in a minimization step and on dual variables in a maximization step.
Two algorithmic methods are proposed [28] (p.131-132, fig.3.10-3.11):
1 - The Arrow-Hurwitz method, in which an iteration step (minimization) is performed on the primal variables followed by an iteration step (maximization) on the dual variables.
The procedure described in the previous section has a clear analogy with this method, it is of course understood that in the case of NMF, we must successively minimize with respect to the two unknowns.
2 - The Uzawa method in which one performs a mininization (several iterative steps) on the primal variables until a stop criterion is met, followed by

a maximization (several iterative steps) on the dual variables until a stop criterion is met, and so on.
The Uzawa method allows us to consider a procedure for MMF that would consist of first minimizing H (for example), doing several iteration steps (until a stop criterion is fulfilled), then repeating the same procedure on X, and so on.
Such a procedure has never been considered to my knowledge for NMF, however, the example shown in [28] seems to indicate better results with the Uzawa method in terms of convergence speed.
Of course, this is just an example applied to saddle point research, i.e. in a very different context, and therefore nothing is well established....
This alternative method is not used here; it has been used for blind deconvolution in [59], [60].

11.3.2 Iterations with respect to H.

Whatever the type of divergence used, whether it is scale invariant with respect to Q or non-invariant, taking into account the sum constraint on the columns of H implies to proceed by means of a variable change method.

- Important remark.
Multiplicative algorithms can always be obtained by introducing a shift in the terms constituting the gradient of the divergence considered; this point will be developed in the following section.

****In this section, we note $Q = H^k X^k$.**

1 - Non-multiplicative algorithm.

In order to take into account the constraint $\sum_i H_{ij} = 1 \ \forall j$, we proceed to the change of variables:

$$H_{ij} = \frac{z_{ij}}{\sum_l z_{lj}} \tag{11.21}$$

First, we build an iterative algorithm on "z", then we go back to an algorithm on H.
To do that, we have:

$$\frac{\partial D}{\partial z_{nm}} = \sum_i \frac{\partial D}{\partial H_{im}} \frac{\partial H_{im}}{\partial z_{nm}} \tag{11.22}$$

With:

$$\frac{\partial H_{im}}{\partial z_{nm}} = \frac{1}{\sum_l z_{lm}} (\delta_{in} - H_{im}) \tag{11.23}$$

We have:

$$\frac{\partial D}{\partial z_{nm}} = \sum_i \frac{\partial D}{\partial H_{im}} \frac{1}{\sum_l z_{lm}} (\delta_{in} - H_{im}) \tag{11.24}$$

And finally:

$$\frac{\partial D}{\partial z_{nm}} = \frac{1}{\sum_l z_{lm}} \left[\frac{\partial D}{\partial H_{nm}} - \sum_i H_{im} \frac{\partial D}{\partial H_{im}} \right] \tag{11.25}$$

One can immediately verify that we have:

$$\sum_n z_{nm} \frac{\partial D}{\partial z_{nm}} = 0 \tag{11.26}$$

We are reasoning on the column "m" and we write in a first step, an algorithm on the "n" component:

$$z_{nm}^{k+1} = z_{nm}^k + \alpha_{nm}^k z_{nm}^k \left[-\frac{\partial D}{\partial z_{nm}^k} \right] \tag{11.27}$$

Taking into account the considerations on the determination of the descent step size outlined in **Chapter 10**, an algorithm that ensures the non-negativity and convergence of the algorithm for all "n" components can be written in the form:

$$z_{nm}^{k+1} = z_{nm}^k + \alpha_m^k z_{nm}^k \left[-\frac{\partial D}{\partial z_{nm}^k} \right] \tag{11.28}$$

and finally, taking into account (11.25):

$$z_{nm}^{k+1} = z_{nm}^k + \alpha_m^k \frac{z_{nm}^k}{\sum_l z_{lm}^k} \left[\left(-\frac{\partial D}{\partial H_{nm}^k} \right) - \sum_i H_{im}^k \left(-\frac{\partial D}{\partial H_{im}^k} \right) \right] \tag{11.29}$$

From this expression, according to (11.21) and (11.26) we can easily verify that we have, $\forall m$:

$$\sum_n z_{nm}^{k+1} = \sum_n z_{nm}^k \tag{11.30}$$

With such an algorithm, we move in a solution space such that $\sum_n z_{nm}^k = Cte \ \forall k$, so if the divergence considered is convex with respect to H, the convexity is maintained during the change of variables.

Therefore, going back to the algorithm (11.29), and dividing the whole by $\sum_n z_{nm}^k$, it comes:

$$H_{nm}^{k+1} = H_{nm}^k + \alpha_m^k \frac{H_{nm}^k}{\sum_l z_{lm}^k} \left[\left(-\frac{\partial D}{\partial H_{nm}^k} \right) - \sum_i H_{im}^k \left(-\frac{\partial D}{\partial H_{im}^k} \right) \right] \quad (11.31)$$

or also, with $\delta_m^k = \frac{\alpha_m^k}{\sum_l z_{lm}^k}$:

$$H_{nm}^{k+1} = H_{nm}^k + \delta_m^k H_{nm}^k \left[\left(-\frac{\partial D}{\partial H_{nm}^k} \right) - \sum_i H_{im}^k \left(-\frac{\partial D}{\partial H_{im}^k} \right) \right] \quad (11.32)$$

Which can be written in condensed form:

$$H_{nm}^{k+1} = H_{nm}^k + \delta_m^k H_{nm}^k \left\{ \left(-\frac{\partial D}{\partial H_{nm}^k} \right) - \left[\left(H^k \right)^T \left(-\frac{\partial D}{\partial H^k} \right) \right]_{mm} \right\} \quad (11.33)$$

Note that the second of the terms between braces is a constant for the column "m" considered.
It can then be observed that regardless of the descent step size δ_m^k, we have $\forall m$:

$$\sum_n H_{nm}^{k+1} = \sum_n H_{nm}^k = 1 \quad (11.34)$$

The algorithm (11.32) (11.33) represents the algorithm corresponding to the iterations on H for the NMF.
It is initialized with:
*** H^0 such that $\sum_n H_{nm}^0 = 1 \ \forall m$.**
*** X^0 such that $\sum_n X_{nm}^0 = \sum_n Y_{nm} \ \forall m$.**
Note that in this algorithm, one can update the H matrix, column by column.
Using the expression of the gradient relative to H (11.13), we can also write the algorithm (11.32) in the form:

$$H_{nm}^{k+1} = H_{nm}^k + \delta_m^k H_{nm}^k \left[\left(\sum_l \left[-\frac{\partial D}{\partial Q_{nl}} \right] X_{lm}^T \right) - \sum_i H_{im}^k \left(\sum_l \left[-\frac{\partial D}{\partial Q_{il}} \right] X_{lm}^T \right) \right] \quad (11.35)$$

Or else in condensed form:

$$H_{nm}^{k+1} = H_{nm}^k + \delta_m^k H_{nm}^k \left\{ \left[\left(-\frac{\partial D}{\partial Q} \right) X^T \right]_{nm} - \left[\left(H^k \right)^T \left(-\frac{\partial D}{\partial Q} \right) X^T \right]_{mm} \right\} \quad (11.36)$$

With, of course, in that expression, $X \equiv X^k$ and $Q = H^k X^k$.

2 - Multiplicative algorithm.

If we want to obtain a multiplicative algorithm while preserving the non-negativity and sum constraints, we must first introduce an shift of the components of the opposite of the gradient $\left[-\frac{\partial D}{\partial H_{ij}^k}\right]$ in such a way as to make them all non-negative; to do this, one writes:

$$\left[-\frac{\partial D}{\partial H_{ij}^k}\right]_d = \left[-\frac{\partial D}{\partial H_{ij}^k}\right] - \min_i \left[-\frac{\partial D}{\partial H_{ij}^k}\right] + \epsilon \tag{11.37}$$

Taking into account the definition of H_{ij} (11.21), the algorithm (11.32) can be rewritten:

$$H_{nm}^{k+1} = H_{nm}^k + \delta_m^k H_{nm}^k \left[\left(-\frac{\partial D}{\partial H_{nm}^k}\right)_d - \sum_i H_{im}^k \left(-\frac{\partial D}{\partial H_{im}^k}\right)_d\right] \tag{11.38}$$

With this algorithm, the property (11.34) is maintained and using the SGM method presented in **Chapter 10**, we can write an algorithm which is the basis of the multiplicative forms, as follows:

$$H_{nm}^{k+1} = H_{nm}^k + \delta_m^k H_{nm}^k \left[\frac{\left(-\frac{\partial D}{\partial H_{nm}^k}\right)_d}{\sum_i H_{im}^k \left(-\frac{\partial D}{\partial H_{im}^k}\right)_d} - 1\right] \tag{11.39}$$

As is always the case, a purely multiplicative algorithm can be obtained if one chooses for all iterations a descent step equal to 1.
The algorithm is then written:

$$H_{nm}^{k+1} = H_{nm}^k \left[\frac{\left(-\frac{\partial D}{\partial H_{nm}^k}\right)_d}{\sum_i H_{im}^k \left(-\frac{\partial D}{\partial H_{im}^k}\right)_d}\right] \tag{11.40}$$

Using the expression of (11.13), we can also write, with $X \equiv X^k$ et $Q = H^k X^k$:

$$H_{nm}^{k+1} = H_{nm}^k \left[\frac{\left(\sum_l \left[-\frac{\partial D}{\partial Q_{nl}}\right] X_{lm}^T\right)_d}{\sum_i H_{im}^k \left(\sum_l \left[-\frac{\partial D}{\partial Q_{il}}\right] X_{lm}^T\right)_d}\right] \tag{11.41}$$

But of course, if the non-negativity and sum constraints are fulfilled, there is no guarantee that such an algorithm will be convergent.

11.3.3 Iterations on X.

For these unknowns, in order to impose the constraint $\sum_i X_{ij} = \sum_i Y_{ij}\ \forall j$ (possibly =1), two cases must be distinguished depending on whether the divergence to be minimized is scale invariant or not.
If a non-invariant divergence is used, a change in variables similar to that used for H must be introduced.
If an invariant divergence is used, the implementation of the general property of these divergences deduced from (3.7), allows to ensure that the sum constraint is satisfied, without it being necessary to introduce a change of variables; this aspect will be developed in the section devoted to the scale invariant divergences.

In this section, we develop the case of non-invariant divergences; we therefore proceed by change of variables.

The use of non-invariant divergences associated with the variable change method will lead to multiplicative algorithms for X in a similar way to what has been done for H.

We have obtained H^{k+1}, we know X^k, we are looking for X^{k+1}, so here, $Q = H^{k+1}X^k$.

1 - Non-multiplicative algorithm.

Taking into account the sum constraints imposed on the columns of the matrix X, the change of variables corresponding to X is written as follows:

$$X_{ij} = \frac{t_{ij}}{\sum_l t_{lj}} \sum_l Y_{lj} \quad \Leftrightarrow \quad \sum_i X_{ij} = \sum_l Y_{lj} \tag{11.42}$$

For a divergence $D\left(P\|Q\right)$, the gradient with respect to "t" is written as:

$$\frac{\partial D\left(P\|Q\right)}{\partial t_{nm}} = \sum_i \frac{\partial D\left(P\|Q\right)}{\partial X_{im}} \frac{\partial X_{im}}{\partial t_{nm}} \tag{11.43}$$

With:

$$\frac{\partial X_{im}}{\partial t_{nm}} = \frac{\sum_l Y_{lm}}{\sum_l t_{lm}} \left[\delta_{in}\right] - \frac{X_{im}}{\sum_l t_{lm}} = \frac{\sum_l Y_{lm}}{\sum_l t_{lm}} \left[\delta_{in}\right] - \frac{X_{im}}{\sum_l t_{lm}} \tag{11.44}$$

We have:

$$\frac{\partial D\left(P\|Q\right)}{\partial t_{nm}} = \frac{\sum_l Y_{lm}}{\sum_l t_{lm}} \frac{\partial D\left(P\|Q\right)}{\partial X_{nm}} - \frac{1}{\sum_l t_{lm}} \sum_i X_{im} \frac{\partial D\left(P\|Q\right)}{\partial X_{im}} \tag{11.45}$$

From this we deduce the iterative algorithm corresponding to the component "n" of the column "m":

$$t_{nm}^{k+1} = t_{nm}^{k} + \alpha_{nm}^{k} t_{nm}^{k} \left[\frac{\sum_l Y_{lm}}{\sum_l t_{lm}^{k}} \left(-\frac{\partial D\left(P\|Q\right)}{\partial X_{nm}^{k}} \right) - \frac{1}{\sum_l t_{lm}^{k}} \sum_i X_{im}^{k} \left(-\frac{\partial D\left(P\|Q\right)}{\partial X_{im}^{k}} \right) \right] \tag{11.46}$$

Taking into account the considerations on the determination of the descent step size as specified in **Chapter 10**, an algorithm that will guarantee the non-negativity and convergence for all the components "n" can be written in the form:

$$t_{nm}^{k+1} = t_{nm}^{k} + \alpha_{m}^{k} t_{nm}^{k} \left[\frac{\sum_l Y_{lm}}{\sum_l t_{lm}^{k}} \left(-\frac{\partial D\left(P\|Q\right)}{\partial X_{nm}^{k}} \right) - \frac{1}{\sum_l t_{lm}^{k}} \sum_i X_{im}^{k} \left(-\frac{\partial D\left(P\|Q\right)}{\partial X_{im}^{k}} \right) \right] \tag{11.47}$$

We can then easily verify that we have:

$$\sum_n t_{nm}^{k+1} = \sum_n t_{nm}^{k} \tag{11.48}$$

Consequently, if the divergence considered is convex with respect to X, the change of variables will not modify this property.
By introducing the descent step size $\delta_m^k = \frac{\alpha_m^k}{\sum_l t_{lm}^k}$, we can therefore go back to the initial variables by dividing the two members of (11.47) by $\sum_n t_{nm}^k$, which leads to:

$$X_{nm}^{k+1} = X_{nm}^{k} + \delta_{m}^{k} X_{nm}^{k} \left[\left(\sum_l Y_{lm} \right) \left(-\frac{\partial D\left(P\|Q\right)}{\partial X_{nm}^{k}} \right) - \sum_i X_{im}^{k} \left(-\frac{\partial D\left(P\|Q\right)}{\partial X_{im}^{k}} \right) \right] \tag{11.49}$$

Which can also be written in condensed form:

$$X^{k+1} = X^{k} + \delta_{m}^{k} X^{k} \left\{ \left(\sum_l Y_{lm} \right) \left(-\frac{\partial D\left(P\|Q\right)}{\partial X^{k}} \right)_{nm} - \left[\left(X^{k} \right)^{T} \left(-\frac{\partial D\left(P\|Q\right)}{\partial X^{k}} \right) \right]_{mm} \right\} \tag{11.50}$$

Note that the second term in the braces is a constant for the whole "m" column.
With an initialization such as $\sum_n X^0_{nm} = \sum_n Y_{nm}\ \forall m$, taking into account the change of variables (11.42), we can check without difficulty that the sum constraint on the columns of X is satisfied and that we have:

$$\sum_n X^{k+1}_{nm} = \sum_n X^k_{nm} = \sum_n Y_{nm} \quad \forall m \tag{11.51}$$

In this algorithm, we will use the expression of the gradient with respect to X:

$$\left[\frac{\partial D\left(P\|Q\right)}{\partial X_{nm}}\right] = \sum_l H^T_{nl}\left[\frac{\partial D}{\partial Q}\right]_{lm} \tag{11.52}$$

With $H \equiv H^{k+1}$ and $Q = H^{k+1}X^k$.

2 - Multiplicative algorithm.

If we wish to obtain a multiplicative algorithm while preserving the non-negativity and sum constraints, we must first introduce a shift of the components of the opposite of the gradient $\left[-\frac{\partial D}{\partial X^k_{ij}}\right]$ in such a way as to make them all non-negative; to do this, one writes:

$$\left[-\frac{\partial D}{\partial X^k_{ij}}\right]_d = \left[-\frac{\partial D}{\partial X^k_{ij}}\right] - \min_i\left[-\frac{\partial D}{\partial X^k_{ij}}\right] + \epsilon \tag{11.53}$$

We can then rewrite the algorithm (11.49) in the form:

$$X^{k+1}_{nm} = X^k_{nm} + \delta^k_m X^k_{nm}\left[\left(\sum_l Y_{lm}\right)\left(-\frac{\partial D\left(P\|Q\right)}{\partial X^k_{nm}}\right)_d - \sum_i X^k_{im}\left(-\frac{\partial D\left(P\|Q\right)}{\partial X^k_{im}}\right)_d\right] \tag{11.54}$$

According to the SGM method presented in **Chapter 10**, we can deduce the algorithm:

$$X^{k+1}_{nm} = X^k_{nm} + \delta^k_m X^k_{nm}\left[\frac{\left(\sum_l Y_{lm}\right)\left(-\frac{\partial D(P\|Q)}{\partial X^k_{nm}}\right)_d}{\sum_i X^k_{im}\left(-\frac{\partial D(P\|Q)}{\partial X^k_{im}}\right)_d} - 1\right] \tag{11.55}$$

Then, using a descent step size equal to 1, we obtain the purely multiplicative algorithm:

$$X_{nm}^{k+1} = X_{nm}^{k} \left[\frac{(\sum_l Y_{lm}) \left(-\frac{\partial D(P\|Q)}{\partial X_{nm}^k} \right)_d}{\sum_i X_{im}^k \left(-\frac{\partial D(P\|Q)}{\partial X_{im}^k} \right)_d} \right] \tag{11.56}$$

We can easily verify that we have: $\sum_n X_{nm}^{k+1} = \sum_n Y_{nm}$.
Nevertheless, for this algorithm, as for all algorithms with a fixed descent step size, there is no guarantee of convergence, even if all the constraints are fulfilled.

*** Important remark :** We will see in the following sections that, in the particular case of **"Alpha divergences"**, given the specific expression of the gradient, the iterative algorithm on H (11.32) and the iterative algorithm on X (11.49) have a simplified expression that makes the shift unnecessary. Thus, a multiplicative algorithm can be written directly using the SGM method presented in **Chapter 10**.

11.3.4 Non-invariant divergences - Overview of the algorithms.

By performing alternately the iterations, first on H, then on X, and with the initialisations:
* H^0 such that $\sum_n H_{nm}^0 = 1 \ \forall m$
* X^0 such that $\sum_n X_{nm}^0 = \sum_n Y_{nm} \ \forall m$;

we first have:
** $X \equiv X^k$, $H \equiv H^k$ et $Q = H^k X^k$

The algorithm obtained using the change of variables on H is written as follows (11.32):

$$H_{nm}^{k+1} = H_{nm}^{k} + \delta_m^k H_{nm}^k \left[\left(-\frac{\partial D}{\partial H_{nm}^k} \right) - \sum_i H_{im}^k \left(-\frac{\partial D}{\partial H_{im}^k} \right) \right] \tag{11.57}$$

or also (11.35):

$$H_{nm}^{k+1} = H_{nm}^{k} + \delta_m^k H_{nm}^k \left[\left(\sum_l \left[-\frac{\partial D}{\partial Q_{nl}} \right] X_{lm}^T \right) - \sum_i H_{im}^k \left(\sum_l \left[-\frac{\partial D}{\partial Q_{il}} \right] X_{lm}^T \right) \right] \tag{11.58}$$

The algorithm can equally be obtained in a purely multiplicative form (11.41):

$$H_{nm}^{k+1} = H_{nm}^{k} \left[\frac{\left(\sum_l \left[\frac{-\partial D}{\partial Q_{nl}} \right] X_{lm}^T \right)_d}{\sum_i H_{im}^k \left(\sum_l \left[-\frac{\partial D}{\partial Q_{il}} \right] X_{lm}^T \right)_d} \right] \tag{11.59}$$

At this point, we have:

** $H \equiv H^{k+1}$, $X \equiv X^k$ and $Q = H^{k+1} X^k$.

With $\delta_m^k = \frac{\alpha_m^k}{\sum_l t_{lm}^k}$, the algorithm obtained using the change of variables on X (11.49) is written:

$$X_{nm}^{k+1} = X_{nm}^k + \delta_m^k X_{nm}^k \left[\left(\sum_l Y_{lm} \right) \left(-\frac{\partial D\left(P \| Q\right)}{\partial X_{nm}^k} \right) - \sum_i X_{im}^k \left(-\frac{\partial D\left(P \| Q\right)}{\partial X_{im}^k} \right) \right] \tag{11.60}$$

In this algorithm, we will use the expression of the gradient with respect to X:

$$\left[\frac{\partial D\left(P \| Q\right)}{\partial X_{nm}} \right] = \sum_l H_{nl}^T \left[\frac{\partial D}{\partial Q} \right]_{lm} \tag{11.61}$$

The corresponding algorithm in purely multiplicative form (11.56) is written as follows:

$$X_{nm}^{k+1} = X_{nm}^k \left[\frac{\left(\sum_l Y_{lm} \right) \left(-\frac{\partial D(P \| Q)}{\partial X_{nm}^k} \right)_d}{\sum_i X_{im}^k \left(-\frac{\partial D(P \| Q)}{\partial X_{im}^k} \right)_d} \right] \tag{11.62}$$

At this point, a cycle has been run on the 2 variables; we have therefore obtained H^{k+1} and X^{k+1}.

11.4 Applications to some typical divergences.

In this section, the algorithms corresponding to the "Alpha" and "Beta" divergences are discussed.

11.4.1 "Alpha" divergences.

The basic divergence (5.15) when applied to tables is written as follows:

$$A\left(P\|Q\right)=\frac{1}{\lambda\left(\lambda-1\right)}\sum_{j}\sum_{i}\left[P_{ij}^{\lambda}Q_{ij}^{1-\lambda}-\lambda P_{ij}-\left(1-\lambda\right)Q_{ij}\right] \tag{11.63}$$

The expression of the gradient with respect to Q is written as follows:

$$\frac{\partial A\left(P\|Q\right)}{\partial Q_{nm}}=\frac{1}{\lambda}\left[1-\left(\frac{P_{nm}}{Q_{nm}}\right)^{\lambda}\right] \tag{11.64}$$

We proceed alternately to iterations, first on H, then on X.

*** Itérations on H.**

** First, we have: $X\equiv X^{k}$, $H\equiv H^{k}$, $P=Y$ and $Q=H^{k}X^{k}$

From (11.64), we have:

$$\frac{\partial A\left(P\|Q\right)}{\partial H_{nm}}=\sum_{j}\frac{\partial A\left(P\|Q\right)}{\partial Q_{nj}}X_{jm}^{T}=\frac{1}{\lambda}\sum_{j}\left[1-\left(\frac{P_{nj}}{Q_{nj}}\right)^{\lambda}\right]X_{jm}^{T} \tag{11.65}$$

The iterative algorithm on H is then written from (11.35):

$$\begin{aligned}H_{nm}^{k+1}=H_{nm}^{k}+\frac{\delta_{m}^{k}}{\lambda}H_{nm}^{k}\Bigg[&\left(\sum_{l}\left[\left(\frac{P_{nl}}{Q_{nl}}\right)^{\lambda}-1\right]X_{lm}^{T}\right)\\&-\sum_{i}H_{im}^{k}\left(\sum_{l}\left[\left(\frac{P_{il}}{Q_{il}}\right)^{\lambda}-1\right]X_{lm}^{T}\right)\Bigg]\end{aligned} \tag{11.66}$$

This algorithm can be used as it is, or a multiplicative algorithm can be obtained by performing an offset as defined in the previous section.
However, in the particular case of the Alpha divergences, a simplification occurs and the algorithm (11.66) is rewritten:

$$\begin{aligned}H_{nm}^{k+1}=H_{nm}^{k}+\frac{\delta_{m}^{k}}{\lambda}H_{nm}^{k}\Bigg[&\left(\sum_{l}\left(\frac{P_{nl}}{Q_{nl}}\right)^{\lambda}X_{lm}^{T}\right)\\&-\sum_{i}H_{im}^{k}\left(\sum_{l}\left(\frac{P_{il}}{Q_{il}}\right)^{\lambda}X_{lm}^{T}\right)\Bigg]\end{aligned} \tag{11.67}$$

Consequently, the multiplicative algorithm will be obtained without the need for an offset, by taking a descent step size equal to $\frac{\delta_m^k}{\lambda} = 1$, from the expression:

$$H_{nm}^{k+1} = H_{nm}^k + \frac{\delta_m^k}{\lambda} H_{nm}^k \left[\frac{\sum_l \left(\frac{P_{nl}}{Q_{nl}}\right)^\lambda X_{lm}^T}{\sum_i H_{im}^k \left(\sum_l \left(\frac{P_{il}}{Q_{il}}\right)^\lambda X_{lm}^T\right)} - 1 \right] \tag{11.68}$$

*** Itérations on X.**

** First, we have: $X \equiv X^k$, $H \equiv H^{k+1}$, $P = Y$ and $Q = H^{k+1}X^k$

From the expression (11.64), we have:

$$\frac{\partial A\left(P\|Q\right)}{\partial X_{nm}} = \sum_j H_{nj}^T \frac{\partial A\left(P\|Q\right)}{\partial Q_{jm}} \tag{11.69}$$

Or also:

$$\frac{\partial A\left(P\|Q\right)}{\partial X_{nm}} = \frac{1}{\lambda}\left[\sum_j H_{nj}^T - \sum_j H_{nj}^T \left(\frac{P_{jm}}{Q_{jm}}\right)^\lambda\right] \tag{11.70}$$

Considering the properties of H, it comes:

$$\frac{\partial A\left(P\|Q\right)}{\partial X_{nm}} = \sum_j H_{nj}^T \frac{\partial A\left(P\|Q\right)}{\partial Q_{jm}} = \frac{1}{\lambda}\left[1 - \sum_j H_{nj}^T \left(\frac{P_{jm}}{Q_{jm}}\right)^\lambda\right] \tag{11.71}$$

From (11.49) the iterative algorithm on X is then written:

$$\begin{aligned} X_{nm}^{k+1} = X_{nm}^k + \frac{\delta_m^k}{\lambda} X_{nm}^k \Bigg\{ & \sum_l Y_{lm} \left[\sum_j H_{nj}^T \left(\frac{P_{jm}}{Q_{jm}}\right)^\lambda - 1\right] \\ & - \sum_i X_{im}^k \left[\sum_j H_{ij}^T \left(\frac{P_{jm}}{Q_{jm}}\right)^\lambda - 1\right] \Bigg\} \end{aligned} \tag{11.72}$$

Which can be simplified by:

$$\begin{aligned} X_{nm}^{k+1} = X_{nm}^k + \frac{\delta_m^k}{\lambda} X_{nm}^k \Bigg\{ & \sum_l Y_{lm} \left[\sum_j H_{nj}^T \left(\frac{P_{jm}}{Q_{jm}}\right)^\lambda\right] \\ & - \sum_i X_{im}^k \left[\sum_j H_{ij}^T \left(\frac{P_{jm}}{Q_{jm}}\right)^\lambda\right] \Bigg\} \end{aligned} \tag{11.73}$$

The multiplicative algorithm can be obtained from (11.72) by introducing an offset to render the bracketed terms positive.
However, due to the simplified form of the Alpha divergences (11.73), a multiplicative algorithm can also be obtained without the need for an offset. Indeed, such an algorithm is obtained by taking a descent step $\frac{\delta_m^k}{\lambda}$ equal to $1\ \forall k$ from the expression:

$$X_{nm}^{k+1} = X_{nm}^k + \frac{\delta_m^k}{\lambda} X_{nm}^k \left\{ \frac{\sum_l Y_{lm} \left[\sum_j H_{nj}^T \left(\frac{P_{jm}}{Q_{jm}} \right)^\lambda \right]}{\sum_i X_{im}^k \left[\sum_j H_{ij}^T \left(\frac{P_{jm}}{Q_{jm}} \right)^\lambda \right]} - 1 \right\} \tag{11.74}$$

11.4.2 "Beta" divergences.

The basis divergence (5.20) applied to tables is written as follows:

$$B\left(P \| Q\right) = \frac{1}{\lambda\left(\lambda - 1\right)} \sum_j \sum_i \left[P_{ij}^\lambda - \lambda P_{ij} Q_{ij}^{\lambda-1} - \left(1 - \lambda\right) Q_{ij}^\lambda \right] \tag{11.75}$$

The expression of the gradient with respect to Q is written as follows:

$$\frac{\partial B\left(P \| Q\right)}{\partial Q_{nm}} = Q_{nm}^{\lambda-1} - Q_{nm}^{\lambda-2} P_{nm} \tag{11.76}$$

We proceed alternately to iterations, first on H, then on X.

*** Itérations on H.**

** First, we have: $X \equiv X^k$, $H \equiv H^k$, $P = Y$ and $Q = H^k X^k$

From the expression (11.76), we have:

$$\frac{\partial B\left(P \| Q\right)}{\partial H_{nm}} = \sum_j \frac{\partial B\left(P \| Q\right)}{\partial Q_{nj}} X_{jm}^T = \sum_j \left[Q_{nj}^{\lambda-1} - Q_{nj}^{\lambda-2} P_{nj} \right] X_{jm}^T \tag{11.77}$$

The iterative algorithm on H is then written from (11.35):

$$H_{nm}^{k+1} = H_{nm}^k + \delta_m^k H_{nm}^k \left[\sum_j \left(Q_{nj}^{\lambda-2} P_{nj} - Q_{nj}^{\lambda-1} \right) X_{jm}^T - \sum_i H_{im}^k \sum_j \left(Q_{ij}^{\lambda-2} P_{ij} - Q_{ij}^{\lambda-1} \right) X_{jm}^T \right] \tag{11.78}$$

To obtain a multiplicative algorithm that maintains the flux, there is no other solution than to shift the terms in "$\sum_j$" of the previous expression to render them positive.

*** Itérations on X.**

** First, we have: $X \equiv X^k$, $H \equiv H^{k+1}$, $P = Y$ and $Q = H^{k+1}X^k$

From the expression (11.76), we have:

$$\frac{\partial B\left(P\|Q\right)}{\partial X_{nm}} = \sum_j H_{nj}^T \frac{\partial B\left(P\|Q\right)}{\partial Q_{jm}} = \sum_j H_{nj}^T \left(Q_{jm}^{\lambda-1} - Q_{jm}^{\lambda-2} P_{jm}\right) \tag{11.79}$$

From (11.49) the iterative algorithm on X is written as follows:

$$X_{nm}^{k+1} = X_{nm}^k + \delta_m^k X_{nm}^k \left[\left(\sum_l Y_{lm} \right) \sum_j H_{nj}^T \left(Q_{jm}^{\lambda-2} P_{jm} - Q_{jm}^{\lambda-1}\right) - \sum_i X_{im}^k \sum_j H_{ij}^T \left(Q_{jm}^{\lambda-2} P_{jm} - Q_{jm}^{\lambda-1}\right) \right] \tag{11.80}$$

Considering the properties of the columns in the X matrix, the flux will be maintained throughout the iterations.
The multiplicative algorithm can be obtained from (11.80) by introducing an offset in order to make positive the terms of the form "$\sum_j$" of the previous expression.

Note: Unlike the case of Alpha divergences, a grouping of terms with the same sign in (11.78) and (11.80) can lead to a multiplicative algorithm; however, the algorithm thus obtained will not have the properties of preserving the sum constraints during iterations.

11.5 Fundamental property of scale invariant divergences.

Two cases are considered in this paragraph:
In the first case, the divergence we will construct will be scale invariant when all the elements of the Q matrix are multiplied by the same positive

constant. The divergences obtained will have interesting properties in the case of blind deconvolution for example.
In the second case the scale invariance property relates to each column of the matrix Q, i.e. the divergence obtained will be invariant when each column of Q is multiplied by a positive scalar which is specific to it. The divergences obtained will have properties that will be specifically adapted to the case of NMF.

11.5.1 General case - All components of Q are concerned simultaneously.

Here, starting with a divergence $D(P\|Q)$, we first construct a divergence $DI(P\|Q)$ that remains scale invariant when the elements of the Q matrix are multiplied by the same positive constant.
In this case, the nominal invariance factor, valid for the Q matrix as a whole, is calculated as follows:

$$K_0(P,Q) = Arg\ min_K \sum_{ij} \frac{\partial D(P_{ij}\|KQ_{ij})}{\partial K} \tag{11.81}$$

*** Fundamental property.**

The general property of the scale invariant divergences with respect to the Q matrix as a whole is written as follows:

$$\sum_{nm} Q_{nm} \frac{\partial D(P\|KQ)}{\partial Q_{nm}} = \sum_{nm} Q_{nm} \frac{\partial DI(P\|Q)}{\partial Q_{nm}} = 0 \tag{11.82}$$

Obviously, this is strictly equivalent to the calculus developed in Chapter 3, equation (3.7), as long as the matrices Q and $\frac{\partial DI(P\|Q)}{\partial Q}$ are written lexicographically.

Note that with the matrix relationship $Q \equiv HX$, the invariance on Q translates into an invariance on X, but not on H.

- Detail of the calculation.

The general property of these divergences can be established as follows.
We have:

$$\frac{\partial D(P\|KQ)}{\partial Q_{nm}} = \sum_{ij} \frac{\partial D(P_{ij}\|KQ_{ij})}{\partial (KQ_{ij})} \frac{\partial (KQ_{ij})}{\partial Q_{nm}} \tag{11.83}$$

but:

$$\frac{\partial\left(KQ_{ij}\right)}{\partial Q_{nm}}=\frac{\partial\left(KQ_{ij}\right)}{\partial K}\frac{\partial K}{\partial Q_{nm}}+\frac{\partial\left(KQ_{ij}\right)}{\partial Q_{ij}}\frac{\partial Q_{ij}}{\partial Q_{nm}} \tag{11.84}$$

That is:

$$\frac{\partial\left(KQ_{ij}\right)}{\partial Q_{nm}}=Q_{ij}\frac{\partial K}{\partial Q_{nm}}+K\delta_{in}\delta_{jm} \tag{11.85}$$

which results in:

$$\frac{\partial D\left(P\|KQ\right)}{\partial Q_{nm}}=\sum_{ij}Q_{ij}\frac{\partial D\left(P_{ij}\|KQ_{ij}\right)}{\partial\left(KQ_{ij}\right)}\frac{\partial K}{\partial Q_{nm}}+K\frac{\partial D\left(P_{nm}\|KQ_{nm}\right)}{\partial\left(KQ_{nm}\right)} \tag{11.86}$$

then, with $Q_{ij}=\frac{\partial(KQ_{ij})}{\partial K}$, we obtain:

$$\sum_{nm}Q_{nm}\frac{\partial D\left(P\|KQ\right)}{\partial Q_{nm}}=\left[\sum_{ij}\frac{\partial D\left(P_{ij}\|KQ_{ij}\right)}{\partial K}\right]\left[K+\sum_{nm}Q_{nm}\frac{\partial K}{\partial Q_{nm}}\right] \tag{11.87}$$

The first term of the second member is zero by definition of the nominal invariance factor, otherwise the second term of the second member is zero if the invariance factor satisfies the differential equation, whether or not the invariance factor is the nominal invariance factor.
So we have:

$$\sum_{nm}Q_{nm}\frac{\partial DI\left(P\|Q\right)}{\partial Q_{nm}}=0 \tag{11.88}$$

Therefore, this relation is equivalent to the relation (3.7) which is established in the **Chapter 3** dealing with scale invariant divergences.

11.5.2 The invariance is derived for each column of Q.

This situation is particularly adapted to NMF; it corresponds to the case where a sum constraint on the component occurs on each column of the X matrix (in the product $[HX]\equiv Q$).
It is quite clear that the multiplication of the column $X_{.j}$ by a constant, results in the multiplication of the homologous column $[HX]_{.j}\equiv Q_{.j}$ by the same constant.
A divergence must therefore be constructed with a scale invariance property when multiplying the columns of Q by various coefficients; consequently, an invariance factor must be calculated for each column of Q.

By denoting the column index as "j" and the row index as "i", the basic divergence is written as follows:

$$D\left(P\|Q\right)=\sum_{j}\sum_{i}D\left(P_{ij}\|Q_{ij}\right)=\sum_{j}\sum_{i}D_{ij} \tag{11.89}$$

The nominal invariance factor corresponding to the column "m" of the matrix Q is defined by:

$$K_{0m}=Arg\ min_{K}\sum_{i}D\left(P_{im}\|KQ_{im}\right) \tag{11.90}$$

When the derivation is possible, the different nominal invariance factors are therefore obtained by calculating for every column "m", K_{0m} such that:

$$\sum_{i}\frac{\partial D\left(P_{im}\|KQ_{im}\right)}{\partial K}=0 \tag{11.91}$$

Obviously, the scalar factor K_{0m} will only contain terms corresponding to the columns labeled "m" in the P and Q matrices.
The resulting scale invariant divergence will be written:

$$DI\left(P\|Q\right)=\sum_{j}\sum_{i}D\left(P_{ij}\|K_{0j}Q_{ij}\right) \tag{11.92}$$

*** Property.**

The general property of the scale invariant divergences, when the invariance factor is calculated separately for each column "m" of Q is written:

$$\sum_{n}Q_{nm}\frac{\partial DI\left(P\|Q\right)}{\partial Q_{nm}}=0 \tag{11.93}$$

- Détail of the calculation.

The derivation of the gradient with respect to Q is as follows:

$$\frac{\partial DI\left(P\|Q\right)}{\partial Q_{nm}}=\sum_{j}\sum_{i}\frac{\partial D\left(P_{ij}\|K_{0j}Q_{ij}\right)}{\partial\left(K_{0j}Q_{ij}\right)}\frac{\partial\left(K_{0j}Q_{ij}\right)}{\partial Q_{nm}} \tag{11.94}$$

In this expression, the term $\frac{\partial(K_{0j}Q_{ij})}{\partial Q_{nm}}$ is zero, unless $j=m$, so it leaves:

$$\frac{\partial DI\left(P\|Q\right)}{\partial Q_{nm}}=\sum_{i}\frac{\partial D\left(P_{im}\|K_{0m}Q_{im}\right)}{\partial\left(K_{0m}Q_{im}\right)}\frac{\partial\left(K_{0m}Q_{im}\right)}{\partial Q_{nm}} \tag{11.95}$$

The second term of the second member is written:

$$\frac{\partial\left(K_{0m}Q_{im}\right)}{\partial Q_{nm}}=\frac{\partial\left(K_{0m}Q_{im}\right)}{\partial K_{0m}}\frac{\partial K_{0m}}{\partial Q_{nm}}+\frac{\partial\left(K_{0m}Q_{im}\right)}{\partial Q_{im}}\frac{\partial Q_{im}}{\partial Q_{nm}} \tag{11.96}$$

That is:

$$\frac{\partial\left(K_{0m}Q_{im}\right)}{\partial Q_{nm}}=Q_{im}\frac{\partial K_{0m}}{\partial Q_{nm}}+K_{0m}\delta_{in} \tag{11.97}$$

Coming back then to the expression of the gradient (11.95), we obtain:

$$\frac{\partial DI\left(P\|Q\right)}{\partial Q_{nm}}=\sum_{i}Q_{im}\frac{\partial D\left(P_{im}\|K_{0m}Q_{im}\right)}{\partial\left(K_{0m}Q_{im}\right)}\frac{\partial K_{0m}}{\partial Q_{nm}}+K_{0m}\frac{\partial D\left(P_{nm}\|K_{0m}Q_{nm}\right)}{\partial\left(K_{0m}Q_{nm}\right)} \tag{11.98}$$

And finally:

$$\sum_{n}Q_{nm}\frac{\partial DI\left(P\|Q\right)}{\partial Q_{nm}}=\left[\sum_{n}\frac{\partial D\left(P_{nm}\|K_{0m}Q_{nm}\right)}{\partial\left(K_{0m}Q_{nm}\right)}\right]\left[K_{0m}+\sum_{n}Q_{nm}\frac{\partial K_{0m}}{\partial Q_{nm}}\right] \tag{11.99}$$

The first term of the second member is zero if the invariance factor is the nominal invariance factor for the column under consideration ("m"), while the second term of the second member is zero for all invariance factors satisfying the differential equation:

$$K_{0m}+\sum_{n}Q_{nm}\frac{\partial K_{0m}}{\partial Q_{nm}}=0 \tag{11.100}$$

Consequently:

$$\sum_{n}Q_{nm}\frac{\partial DI\left(P\|Q\right)}{\partial Q_{nm}}=0\ \forall m \tag{11.101}$$

This relation is similar to the property (3.7) applied on each column of the matrix Q.

11.6 Iterative algorithm for invariant divergences.

For these divergences denoted DI, by operating column by column, we rely on the property (11.93) (11.101).

11.6.1 Iterations on H.

Since the invariance on Q translates into an invariance on X, but not on H, there is no particular thing to expect regarding iterations on H.

Consequently, the iterative algorithm on H will be deduced from the variable change method; it is given in (11.32) (11.35) and is written for an invariant divergence DI, with $H = H^k,\ X = X^k,\ Q = H^k X^k$:

$$H_{nm}^{k+1} = H_{nm}^k + \delta_m^k H_{nm}^k \left[\left(\sum_l \left[-\frac{\partial DI}{\partial Q_{nl}} \right] X_{lm}^T \right) - \sum_i H_{im}^k \left(\sum_l \left[-\frac{\partial DI}{\partial Q_{il}} \right] X_{lm}^T \right) \right] \tag{11.102}$$

A multiplicative algorithm can be obtained by proceeding in an similar way to what has been previously developed for non-invariant divergences.

At this point, we have obtained H^{k+1}.

11.6.2 Iterations on X.

1 - Non-multiplicative algorithm.

The iterative algorithm on X allowing to exploit the properties of invariant divergences, can be written, with the method developed in **Chapter 10**, as follows:

$$X_{nm}^{k+1} = X_{nm}^k + \alpha_m^k X_{nm}^k \left[-\frac{\partial DI}{\partial X_{nm}^k} \right] \tag{11.103}$$

With this algorithm, the constraint $\sum_n X_{nm}^k = Cte\ \forall k$ is ensured as shown in the following calculation.
In the previous sections, we have shown (11.5) that with:

$$[A]_{ij} = \left[\frac{\partial D}{\partial Q} \right]_{ij} \tag{11.104}$$

One could write (11.6) in matrix form:

$$\left[\frac{\partial DI\,(P\|Q)}{\partial X} \right]_{nm} = \frac{\partial DI\,(P\|Q)}{\partial X_{nm}} = \left[H^T A \right]_{nm} = \sum_l H_{nl}^T \frac{\partial DI\,(P\|Q)}{\partial Q_{lm}} \tag{11.105}$$

Taking into account the fact that $H \equiv H^{k+1}$ and $Q = H^{k+1} X^k$, the algorithm (11.103) can therefore be rewritten:

$$X_{nm}^{k+1} = X_{nm}^k + \alpha_m^k X_{nm}^k \left[\sum_i H_{ni}^T \left(-\frac{\partial DI\,(P\|Q)}{\partial Q_{im}} \right) \right] \tag{11.106}$$

Then:

$$\sum_n X_{nm}^{k+1} = \sum_n X_{nm}^k + \alpha_m^k \sum_n \left\{ X_{nm}^k \left[\sum_i H_{ni}^T \left(-\frac{\partial DI\,(P\|Q)}{\partial Q_{im}} \right) \right] \right\} \tag{11.107}$$

By reversing the order of the summations, it comes:

$$\sum_n X_{nm}^{k+1} = \sum_n X_{nm}^k + \alpha_m^k \sum_i \left(-\frac{\partial DI\,(P\|Q)}{\partial Q_{im}} \right) \sum_n H_{ni}^T X_{nm}^k \tag{11.108}$$

or also:

$$\sum_n X_{nm}^{k+1} = \sum_n X_{nm}^k + \alpha_m^k \sum_i \left(HX^k \right)_{im} \left(-\frac{\partial DI\,(P\|Q)}{\partial Q_{im}} \right) \tag{11.109}$$

So, with (11.93), it comes:

$$\sum_n X_{nm}^{k+1} = \sum_n X_{nm}^k \tag{11.110}$$

It is therefore sufficient to choose an initial estimate with the correct sum (i.e. $\sum_n X_{nm}^0 = \sum_n Y_{nm}$), to ensure that this sum is maintained during the iterations.

The algorithm (11.103) (11.106) represents the iterative algorithm on X when minimizing an invariant divergence.

2 - Multiplicative algorithm.

A multiplicative algorithm can be obtained from (11.103) (11.106) by operating as follows:
The SGM method is strictly applied; the opposite of the gradient is first split into a difference of 2 positive terms:

$$-\frac{\partial DI}{\partial X_{nm}^k} = U_{nm}^k - V_{nm}^k \;;\quad U_{nm}^k > 0 \;;\quad V_{nm}^k > 0 \tag{11.111}$$

This allows in a first step (with the usual restrictions on the descent step size), to write a multiplicative algorithm in the form:

$$\widetilde{X}_{nm}^{k+1} = X_{nm}^k \left[\frac{U_{nm}^k}{V_{nm}^k} \right] \tag{11.112}$$

But then the property $\sum_n \widetilde{X}_{nm}^{k+1} = \sum_n X_{nm}^k$ is lost, unlike what we had with algorithms (11.103) (11.106).
However, we can recover this property by introducing an extra step of normalization by writing:

$$X_{nm}^{k+1} = \frac{\widetilde{X}_{nm}^{k+1}}{\sum_n \widetilde{X}_{nm}^{k+1}} \sum_n Y_{nm} \tag{11.113}$$

Due to the property of invariance with respect to X, this normalization does not lead to any change in the invariant divergence considered.
The association of (11.112) and (11.113) is thus a multiplicative algorithm having the summation property on the columns of X. This is subject of course to the convergence of (11.112) in relation to the choice of the descent step size (equal to 1).

11.7 Applications to some particular invariant divergences.

In the previous sections, it was shown that the proposed algorithms used, for a given divergence, the gradient expression:

$$\frac{\partial D\left(P\|Q\right)}{\partial Q_{nm}} \tag{11.114}$$

Consequently, we give for some invariant divergences, the expression of this gradient and the iterative algorithm on X that we can obtain.

11.7.1 Invariant "Alpha" divergences.

From the basic divergence (11.63), the invariance on each column of the table (matrix) Q is obtained by deriving the nominal invariance factor (which is possible); this gives, for the column "j":

$$K_{0j} = \left(\frac{\sum_i P_{ij}^{\lambda} Q_{ij}^{1-\lambda}}{\sum_i Q_{ij}} \right)^{\frac{1}{\lambda}} \tag{11.115}$$

The divergence we obtain, invariant with respect to the columns of Q, is written as follows:

$$AI_Q\left(P\|Q\right) = \frac{1}{\lambda - 1} \sum_j \left[\left(\frac{\sum_l P_{lj}^{\lambda} Q_{lj}^{1-\lambda}}{\sum_l Q_{lj}} \right)^{\frac{1}{\lambda}} \sum_i Q_{ij} - \sum_i P_{ij} \right] \tag{11.116}$$

That is:

$$AI_Q\left(P\|Q\right) = \frac{1}{\lambda - 1} \sum_j \left[K_{0j} \sum_i Q_{ij} - \sum_i P_{ij} \right] \tag{11.117}$$

The construction of the algorithms implies the calculation of the gradient with respect to the elements of the column "m" of the matrix Q which is written after some calculations:

$$\frac{\partial AI_Q\left(P\|Q\right)}{\partial Q_{nm}} = \frac{1}{\lambda} \left[\left(\frac{\sum_l P_{lm}^{\lambda} Q_{lm}^{1-\lambda}}{\sum_l Q_{lm}} \right)^{\frac{1}{\lambda}} - \left(\frac{\sum_l P_{lm}^{\lambda} Q_{lm}^{1-\lambda}}{\sum_i Q_{lm}} \right)^{\frac{1-\lambda}{\lambda}} \frac{P_{nm}^{\lambda}}{Q_{nm}^{\lambda}} \right] \tag{11.118}$$

That is:

$$\frac{\partial AI_Q\left(P\|Q\right)}{\partial Q_{nm}} = \frac{1}{\lambda} \left[K_{0m} - K_{0m}^{1-\lambda} \frac{P_{nm}^{\lambda}}{Q_{nm}^{\lambda}} \right] \tag{11.119}$$

We can verify that the fundamental relation (11.93) for scale invariant divergences with respect to the columns of Q is satisfied and that we have:

$$\sum_n Q_{nm} \frac{\partial AI_Q\left(P\|Q\right)}{\partial Q_{nm}} = 0 \tag{11.120}$$

In the particular case ($\lambda = 1$) which corresponds to the Kullback-Leibler divergence, we obtain from the expression of the gradient (11.119), the gradient of the K.L. divergence , invariant with respect to the columns of Q; such divergence is denoted KLI_Q:

$$\frac{\partial KLI_Q\left(P\|Q\right)}{\partial Q_{nm}} = \frac{\sum_l P_{lm}}{\sum_l Q_{lm}} - \frac{P_{nm}}{Q_{nm}} \tag{11.121}$$

*** Itérations on H.**

With (11.119), it comes:

$$\frac{\partial AI_Q\left(P\|Q\right)}{\partial H_{nm}} = \sum_l \frac{\partial AI_Q\left(P\|Q\right)}{\partial Q_{nl}} X_{lm}^T = \frac{1}{\lambda} \sum_l \left[K_{0l} - K_{0l}^{1-\lambda} \frac{P_{nl}^{\lambda}}{Q_{nl}^{\lambda}} \right] X_{lm}^T \tag{11.122}$$

With $X = X^k$, $H = H^k$, $Q = H^k X^k$ and taking into account the fact that the invariance factor K_{0l} given by (11.115) varies with the iteration because

it depends on Q, the iterative algorithm on H deduced from (11.102) is then written after simplification:

$$H_{nm}^{k+1} = H_{nm}^{k} + \frac{\delta_m^k}{\lambda} H_{nm}^{k} \left[\sum_l K_{0l}^{1-\lambda} \left(\frac{P_{nl}^{\lambda}}{Q_{nl}^{\lambda}} - \sum_i H_{im}^{k} \frac{P_{il}^{\lambda}}{Q_{il}^{\lambda}} \right) X_{lm}^{T} \right] \quad (11.123)$$

*** Itérations on X.**

At this point in the calculation, we have $H \equiv H^{k+1}$, $X \equiv X^k$, $Q \equiv H^{k+1}X^k$.
It has been shown that:

$$\frac{\partial AI_Q(P\|Q)}{\partial X_{nm}} = \sum_j H_{nj}^{T} \frac{\partial AI_Q(P\|Q)}{\partial Q_{jm}} \quad (11.124)$$

That is, with (11.119):

$$\frac{\partial AI_Q(P\|Q)}{\partial X_{nm}} = \frac{1}{\lambda} \sum_j H_{nj}^{T} \left[K_{0m} - K_{0m}^{1-\lambda} \frac{P_{jm}^{\lambda}}{Q_{jm}^{\lambda}} \right] \quad (11.125)$$

Considering that $\sum_j H_{nj}^{T} = 1$, the iterative algorithm on X given by (11.103) is written:

$$X_{nm}^{k+1} = X_{nm}^{k} + \frac{\alpha_m^k}{\lambda} X_{nm}^{k} \left[\left(K_{0m}^{k} \right)^{1-\lambda} \sum_j H_{nj}^{T} \left(\frac{P_{jm}^{\lambda}}{Q_{jm}^{\lambda}} \right) - K_{0m}^{k} \right] \quad (11.126)$$

In this expression, we have taken into account the fact that the invariance factor depends on iteration, indeed, we recall that:

$$K_{0m} = \left(\frac{\sum_i P_{im}^{\lambda} Q_{im}^{1-\lambda}}{\sum_i Q_{im}} \right)^{\frac{1}{\lambda}} \quad (11.127)$$

with $Q \equiv H^{k+1}X^k$; to avoid any ambiguity, we have denoted his expression: K_{0m}^{k}.
Since $K_{0m}^{k} > 0$ and considering the very particular expression of the gradient, this also leads to the expression which is at the basis of the multiplicative algorithm:

$$X_{nm}^{k+1} = X_{nm}^{k} + \frac{K_{0m}^{k} \alpha_m^k}{\lambda} X_{nm}^{k} \left[\left(K_{0m}^{k} \right)^{-\lambda} \sum_j H_{nj}^{T} \left(\frac{P_{jm}^{\lambda}}{Q_{jm}^{\lambda}} \right) - 1 \right] \quad (11.128)$$

By taking a descent step size $\frac{K_{0m}^k \alpha_m^k}{\lambda}$ equal to 1, we obtain the purely multiplicative form:

$$X_{nm}^{k+1} = X_{nm}^k \left(K_{0m}^k\right)^{-\lambda} \sum_j H_{nj}^T \left(\frac{P_{jm}^\lambda}{Q_{jm}^\lambda}\right) \tag{11.129}$$

Using the expression of K_{0m}^k, it can be shown after some simple calculations that $\sum_n X_{nm}^{k+1} = \sum_n X_{nm}^k$.

Logarithmic form.

As proposed for Alpha divergences, we can give a logarithmic form for the invariant divergence AI_Q, it is written as follows:

$$LAI_Q\left(P\|Q\right) = \frac{1}{\lambda - 1}\left\{\log\left[\sum_j \sum_i \left(\frac{\sum_l P_{lj}^\lambda Q_{lj}^{1-\lambda}}{\sum_l Q_{lj}}\right)^{\frac{1}{\lambda}} Q_{ij}\right] - \log\left[\sum_j \sum_i P_{ij}\right]\right\} \tag{11.130}$$

It can be observed that, as with all logarithmic forms, this divergence is invariant not only with respect to the columns of Q, but also with respect to the columns of P.
The gradient with respect to Q is written as:

$$\frac{\partial LAI_Q\left(P\|Q\right)}{\partial Q_{nm}} = \frac{A}{\lambda}\left[\left(\frac{\sum_l P_{lm}^\lambda Q_{lm}^{1-\lambda}}{\sum_l Q_{lm}}\right)^{\frac{1}{\lambda}} - \left(\frac{\sum_l P_{lm}^\lambda Q_{lm}^{1-\lambda}}{\sum_i Q_{lm}}\right)^{\frac{1-\lambda}{\lambda}} \frac{P_{nm}^\lambda}{Q_{nm}^\lambda}\right] \tag{11.131}$$

That is:

$$\frac{\partial LAI_Q\left(P\|Q\right)}{\partial Q_{nm}} = A\frac{\partial AI_Q\left(P\|Q\right)}{\partial Q_{nm}} \tag{11.132}$$

with:

$$A = \frac{1}{\sum_i \sum_j \left(\frac{\sum_l P_{lj}^\lambda Q_{lj}^{1-\lambda}}{\sum_l Q_{lj}}\right)^{\frac{1}{\lambda}} Q_{ij}} = \frac{1}{\sum_i \sum_j K_{0j} Q_{ij}} \tag{11.133}$$

The gradient with respect to Q can then be written in a simpler way:

$$\frac{\partial LAI_Q\left(P\|Q\right)}{\partial Q_{nm}} = A\frac{K_{0m}}{\lambda}\left(1 - K_{0m}^{-\lambda}\frac{P_{nm}^{\lambda}}{Q_{nm}^{\lambda}}\right) \tag{11.134}$$

We then deduce the expression of the gradient with respect to X:

$$\frac{\partial LAI_Q\left(P\|Q\right)}{\partial X_{nm}} = \sum_j H_{nj}^T \frac{\partial LAI_Q\left(P\|Q\right)}{\partial Q_{jm}} \tag{11.135}$$

That is:

$$\frac{\partial LAI_Q\left(P\|Q\right)}{\partial X_{nm}} = A\frac{K_{0m}}{\lambda}\left(1 - K_{0m}^{-\lambda}\sum_j H_{nj}^T\frac{P_{jm}^{\lambda}}{Q_{jm}^{\lambda}}\right) \tag{11.136}$$

Hence the iterative algorithm on X:

$$X_{nm}^{k+1} = X_{nm}^k + A\frac{K_{0m}^k\alpha_m^k}{\lambda}X_{nm}^k\left[\left(K_{0m}^k\right)^{-\lambda}\sum_j H_{nj}^T\left(\frac{P_{jm}^{\lambda}}{Q_{jm}^{\lambda}}\right) - 1\right] \tag{11.137}$$

As with the $AI(P\|Q)$ divergence, we will take into account the fact that $H \equiv H^{k+1}$, $Q = H^{k+1}X^k$, and therefore $K_{0m} \equiv K_{0m}^k$ depends on the iteration.
Note that this algorithm is analogous to (11.128), except for the multiplicative factor "A" that occurs in the correction term and can be included in the descent step size; therefore, the sum constraint (flux-holding) property highlighted in (11.128) is maintained.
Due to the particular expression of the Gradient, a multiplicative flux-holding algorithm is obtained as for the $AI(P\|Q)$ divergence.

11.7.2 Invariant "Beta" divergences .

The basis divergence (5.20) applied to tables is written as follows:

$$B\left(p\|q\right) = \frac{1}{\lambda\left(\lambda-1\right)}\sum_j\sum_i\left[P_{ij}^{\lambda} - \lambda P_{ij}Q_{ij}^{\lambda-1} - \left(1-\lambda\right)Q_{ij}^{\lambda}\right] \tag{11.138}$$

The invariance on each column of the table (matrix) Q is obtained by calculating the nominal invariance factor (which is possible); thus, for the column "j" we obtain:

$$K_{0j} = \frac{\sum_i P_{ij}Q_{ij}^{\lambda-1}}{\sum_i Q_{ij}^{\lambda}} \tag{11.139}$$

The divergence generated, invariant with respect to the columns of Q, is written as follows:

$$BI_Q(P\|Q) = \frac{1}{\lambda(\lambda-1)} \sum_j \left[\sum_i P_{ij}^\lambda - \left(\frac{\sum_l P_{lj} Q_{lj}^{\lambda-1}}{\sum_l Q_{lj}^\lambda} \right)^\lambda \sum_i Q_{ij}^\lambda \right] \quad (11.140)$$

Or also:

$$BI_Q(P\|Q) = \frac{1}{\lambda(\lambda-1)} \sum_j \left[\sum_i P_{ij}^\lambda - K_{0j}^\lambda \sum_i Q_{ij}^\lambda \right] \quad (11.141)$$

The calculation of the gradient of this divergence with respect to Q leads to:

$$\frac{\partial BI_Q(P\|Q)}{\partial Q_{nm}} = \left[\left(\frac{\sum_l P_{lm} Q_{lm}^{\lambda-1}}{\sum_l Q_{lm}^\lambda} \right)^\lambda Q_{nm}^{\lambda-1} - \left(\frac{\sum_l P_{lm} Q_{lm}^{\lambda-1}}{\sum_l Q_{lm}^\lambda} \right)^{\lambda-1} P_{nm} Q_{nm}^{\lambda-2} \right] \quad (11.142)$$

That is to say more simply:

$$\frac{\partial BI_Q(P\|Q)}{\partial Q_{nm}} = \left[K_{0m}^\lambda Q_{nm}^{\lambda-1} - K_{0m}^{\lambda-1} P_{nm} Q_{nm}^{\lambda-2} \right] \quad (11.143)$$

One can verify that the fundamental relationship for divergences invariant with respect to the columns of Q is satisfied.
The case ($\lambda = 1$) corresponding to the Kullback-Leibler divergence is obtained without difficulty; its gradient has been given by (11.121), whereas in the particular case ($\lambda = 2$) corresponding to the mean square deviation MCI_Q, the gradient obtained from (11.142) is written as follows:

$$\frac{\partial MCI_Q(P\|Q)}{\partial Q_{nm}} = \left(\frac{\sum_l P_{lm} Q_{lm}}{\sum_l Q_{lm}^2} \right)^2 Q_{nm} - \left(\frac{\sum_l P_{lm} Q_{lm}}{\sum_l Q_{lm}^2} \right) P_{nm} \quad (11.144)$$

*** Itérations on H.**

With (11.143), we have:

$$\frac{\partial BI_Q(P\|Q)}{\partial H_{nm}} = \sum_l \frac{\partial BI_Q(P\|Q)}{\partial Q_{nl}} X_{lm}^T = \sum_l \left[K_{0l}^\lambda Q_{nl}^{\lambda-1} - K_{0l}^{\lambda-1} P_{nl} Q_{nl}^{\lambda-2} \right] X_{lm}^T \quad (11.145)$$

With $X = X^k$, $H = H^k$, $Q = H^k X^k$ and taking into account that the invariance factor K_{0l} given by (11.139) varies with the iteration via Q, the iterative algorithm on H deduced from (11.102) is written as follows:

$$H_{nm}^{k+1} = H_{nm}^k + \delta_m^k H_{nm}^k \left[\sum_l K_{0l}^{\lambda-1} \left(P_{nl} Q_{nl}^{\lambda-2} - \sum_i H_{im}^k P_{il} Q_{il}^{\lambda-2} \right) X_{lm}^T - \sum_l K_{0l}^{\lambda} \left(\sum_i H_{im}^k Q_{il}^{\lambda-1} - Q_{nl}^{\lambda-1} \right) X_{lm}^T \right] \quad (11.146)$$

*** Itérations on X.**

At this point of the calculus, we have $H \equiv H^{k+1}$, $X \equiv X^k$, $Q \equiv H^{k+1} X^k$. Taking into account that:

$$\frac{\partial BI_Q(P\|Q)}{\partial X_{nm}} = \sum_j H_{nj}^T \frac{\partial BI_Q(P\|Q)}{\partial Q_{jm}} \quad (11.147)$$

It comes from the expression (11.143):

$$\frac{\partial BI_Q(P\|Q)}{\partial X_{nm}} = \sum_j H_{nj}^T \left[K_{0m}^{\lambda} Q_{jm}^{\lambda-1} - K_{0m}^{\lambda-1} P_{jm} Q_{jm}^{\lambda-2} \right] \quad (11.148)$$

To take into account the fact that the invariance factor varies with iteration, we will denote K_{0m}^k; it is given by the expression (11.139), where $Q \equiv H^{k+1} X^k$.
The iterative algorithm on X given by (11.103) is then written as follows:

$$X_{nm}^{k+1} = X_{nm}^k + \alpha_m^k \left(K_{0m}^k \right)^{\lambda-1} X_{nm}^k \left[\sum_j H_{nj}^T \left(P_{jm} Q_{jm}^{\lambda-2} - K_{0m}^k Q_{jm}^{\lambda-1} \right) \right] \quad (11.149)$$

One can check that with this algorithm, the property $\sum_n X_{nm}^{k+1} = \sum_n X_{nm}^k$ is ensured, however, a multiplicative form deduced from this expression will not have this property any more.
To obtain a multiplicative form having the property of maintaining the flow, it will be necessary to proceed in 2 steps as indicated in **(11.6.2)**.

Logarithmic form.

As proposed for the Alpha divergences, we can give a logarithmic form for the BI_Q divergence; it is written as follows:

$$LBI_Q(P\|Q) = \frac{1}{\lambda(\lambda-1)}\left\{\log\left[\sum_j\sum_i P_{ij}^{\lambda}\right] - \log\left[\sum_j\sum_i\left(\frac{\sum_l P_{lj}Q_{lj}^{\lambda-1}}{\sum_l Q_{lj}^{\lambda}}\right)^{\lambda} Q_{ij}^{\lambda}\right]\right\} \quad (11.150)$$

It can be observed that, as with all logarithmic forms, this divergence is invariant not only with respect to columns of Q, but also with respect to columns of P.
The gradient with respect to Q is written as follows:

$$\frac{\partial LBI_Q(P\|Q)}{\partial Q_{nm}} = B\left[\left(\frac{\sum_l P_{lm}Q_{lm}^{\lambda-1}}{\sum_l Q_{lm}^{\lambda}}\right)^{\lambda} Q_{nm}^{\lambda-1} - \left(\frac{\sum_l P_{lm}Q_{lm}^{\lambda-1}}{\sum_l Q_{lm}^{\lambda}}\right)^{\lambda-1} P_{nm}Q_{nm}^{\lambda-2}\right] \quad (11.151)$$

That is:

$$\frac{\partial LBI_Q(P\|Q)}{\partial Q_{nm}} = B\,\frac{\partial BI_Q(P\|Q)}{\partial Q_{nm}} \quad (11.152)$$

with:

$$B = \frac{1}{\sum_j\sum_i Q_{ij}^{\lambda}\left(\frac{\sum_l P_{lj}Q_{lj}^{\lambda-1}}{\sum_l Q_{lj}^{\lambda}}\right)^{\lambda}} = \frac{1}{\sum_i\sum_j K_{0j}^{\lambda}Q_{ij}^{\lambda}} \quad (11.153)$$

The gradient with respect to Q is written in a simpler way:

$$\frac{\partial LBI_Q(P\|Q)}{\partial Q_{nm}} = BK_{0m}^{\lambda-1}\left(K_{0m}Q_{nm}^{\lambda-1} - P_{nm}Q_{nm}^{\lambda-2}\right) \quad (11.154)$$

We then deduce the expression of the gradient with respect to X:

$$\frac{\partial LBI_Q(P\|Q)}{\partial X_{nm}} = \sum_j H_{nj}^T\frac{\partial LBI_Q(P\|Q)}{\partial Q_{jm}} \quad (11.155)$$

That is:

$$\frac{\partial LBI_Q\left(P\|Q\right)}{\partial X_{nm}} = BK_{0m}^{\lambda-1}\left[\sum_j H_{nj}^T\left(K_{0m}Q_{jm}^{\lambda-1} - P_{jm}Q_{jm}^{\lambda-2}\right)\right] \tag{11.156}$$

Hence the iterative algorithm on X:

$$X_{nm}^{k+1} = X_{nm}^k + \alpha_m^k BK_{0m}^{\lambda-1}X_{nm}^k\left[\sum_j H_{nj}^T\left(P_{jm}Q_{jm}^{\lambda-2} - K_{0m}^kQ_{jm}^{\lambda-1}\right)\right] \tag{11.157}$$

As with the $BI(P\|Q)$ divergence, we will take into account the fact that $H \equiv H^{k+1}$, $Q = H^{k+1}X^k$ and that, consequently, $K_{0m} \equiv K_{0m}^k$ is dependent on iteration.

Note that this algorithm is analogous to (11.149), except for the multiplicative factor "B" which is involved in the corrective term and can be included in the descent step size.

We can verify that with this algorithm, the property $\sum_n X_{nm}^{k+1} = \sum_n X_{nm}^k$ is ensured, however, a multiplicative form deduced from this expression will no longer possess this property .

To obtain a multiplicative form having the property of flow holding, it will be necessary to proceed in 2 steps as indicated in **(11.6.2)**.

11.8 Régularisation. Non-invariant divergences

In the context of NMF, a smoothness constraint regularization can only be performed on the columns of the H matrix, in fact, each column represents the spectrum of an elementary component.

With respect to the X matrix, whose columns contain weighting factors, such a smoothness constraint regularization has no meaning.

On the other hand, a much more logical form of regularization is to introduce a certain degree of "*sparsity*" on the elements in the columns of X.

In any case, solving the regularized problem requires minimizing with respect to H and X, a functional of the form:

$$J\left(H,X\right) = D\left(Y\|HX\right) + \gamma DH\left(H\right) + \mu DX\left(X\right) \tag{11.158}$$

The minimization is carried out under the non-negativity and sum constraints already specified.

The "γ" and "μ" coefficients are the positive regularization factors.

Since the iterative process of minimization is performed alternately on H then on X (for example), one works first concerning H on a composite divergence:

$$DCH = D\left(Y \| HX\right) + \gamma DH\left(H\right) \tag{11.159}$$

then, concerning X, on a composite divergence:

$$DCX = D\left(Y \| HX\right) + \mu DX\left(X\right) \tag{11.160}$$

11.8.1 Regularization of the columns of the matrix H.

The regularization is performed separately on each of the columns of H. The basic algorithm is given by (11.32) applied to the composite divergence DCH; it is written:

$$H_{nm}^{k+1} = H_{nm}^{k} + \delta_{m}^{k} H_{nm}^{k} \left\{ -\frac{\partial\left(DCH\right)}{\partial H_{nm}^{k}} + \sum_{i} H_{im}^{k} \frac{\partial\left(DCH\right)}{\partial H_{im}^{k}} \right\} \tag{11.161}$$

We are now considering 2 particular cases of regularization by smoothness constraint:

1 - Euclidean norm of the solution.

$$DH\left(H\right) = DH\left(H \| C\right) = \frac{1}{2} \sum_{j} \sum_{i} \left(H_{ij} - C\right)^{2} \tag{11.162}$$

For columns of H with "N" components, taking into account the constraint $\sum_{i} H_{ij} = 1 \ \forall j$, we will have $C = \frac{1}{N}$.

$$\frac{\partial DH}{\partial H_{nm}} = \left(H_{nm} - C\right) \tag{11.163}$$

2 - Euclidean norm of the Laplacian of the solution.

One can as well use a regularization term expressed by the Euclidean norm of the Laplacian, which will be expressed by:

$$DH\left(H\right) = DH\left(H \| TH\right) = \frac{1}{2} \sum_{j} \sum_{i} \left[H_{ij} - \left(TH_{.j}\right)_{i}\right]^{2} \tag{11.164}$$

Where the matrix operation $\left(TH_{.j}\right)$ corresponds to the convolution of the column "j" of H by the mask [0.5 ; 0 ; 0.5].

The regularization is performed independently for each column of H.
In this case, we have:

$$\frac{\partial DH}{\partial H_{nm}} = H_{nm} - \left[\left(T + T^T\right) H_{.m}\right]_n + \left[\left(T^T T\right) H_{.m}\right]_n \tag{11.165}$$

Of course, taking into account the symmetry of the convolution mask, $T^T = T$.

11.8.2 Regularization of the columns of the matrix X.

The regularized algorithm is developed on the basis of the algorithm (11.49), taking into account the fact that the divergence considered is the composite divergence DCX;

$$X_{nm}^{k+1} = X_{nm}^k + \delta_m^k X_{nm}^k \left[\left(\sum_l Y_{lm}\right)\left(-\frac{\partial DCX}{\partial X_{nm}^k}\right) - \sum_i X_{im}^k \left(-\frac{\partial DCX}{\partial X_{im}^k}\right)\right] \tag{11.166}$$

*** Expression of the penalty term.**

Here again, the regularization applies to each column of the X matrix, but, contrary to the H matrix, the regularization does not relate to the smoothness of the solution, but to its "*sparsity*".
The measure of "*sparsity*" used here was proposed by Hoyer [47]; by denoting X_j the column "j" of X, if "N" is the number of elements of X_j, the sparsity factor for this column is expressed by:

$$s_j = \frac{\sqrt{N} - \frac{\|X_j\|_1}{\|X_j\|_2}}{\sqrt{N} - 1} \tag{11.167}$$

This relationship induces a relationship between $\|X_j\|_1$ and $\|X_j\|_2$ which we will express in the form:

$$\frac{\|X_j\|_2^2}{\|X_j\|_1^2} = A_j^2 \quad ; \quad A_j = \frac{1}{\sqrt{N} - s_j\left(\sqrt{N} - 1\right)} \tag{11.168}$$

The complete "*sparsity*" corresponds to $s_j = 1$, that is $A_j = 1$, while the absence of "*sparsity*" corresponds to $s_j = 0$, that is $A_j = \frac{1}{\sqrt{N}}$.

Furthermore, the constraints imposed during the minimization process for a column "j" of X, which are the non-negativity and the fixed sum, will result in $\|X_j\|_1 = \|Y_j\|_1$, at each iteration.
We will express the corresponding penalty term as:

$$DX\left(X\right) = \sum_j \left[\frac{1}{2}\left\|X_j\right\|_2^2 - \frac{A_j^2}{2}\left\|X_j\right\|_1^2\right]^2 \tag{11.169}$$

The gradient with respect to X will be expressed by:

$$\frac{\partial DX\left(X\right)}{\partial X_{nm}} = X_{nm}\left\|X_m\right\|_2^2 + A_m^4\left\|X_m\right\|_1^3 - A_m^2 X_{nm}\left\|X_m\right\|_1^2 - A_m^2\left\|X_m\right\|_1\left\|X_m\right\|_2^2 \tag{11.170}$$

Or also, by taking into account the sum constraint:

$$\frac{\partial DX\left(X\right)}{\partial X_{nm}} = X_{nm}\left\|X_m\right\|_2^2 + A_m^4\left\|Y_m\right\|_1^3 - A_m^2 X_{nm}\left\|Y_m\right\|_1^2 - A_m^2\left\|Y_m\right\|_1\left\|X_m\right\|_2^2 \tag{11.171}$$

11.9 Some examples of regularized algorithms.

Regarding the "*data consistency*" term, we will deal with the "α" divergence and the "β" divergence.
The regularization term on H will be the Euclidean norm of the solution.
The regularization on X will be made in the sense of Hoyer.

11.9.1 Regularized "Alpha" divergence.

*** Régularization on H.**

The basis algorithm is given by (11.161).
The gradients involved in the basic algorithm are given respectively by (11.65) and (11.163).
Taking into account the simplifications related to the property $\sum_i H_{im}^k = 1$, and with $X \equiv X^k$, the regularized algorithm is written according to (11.32):

$$\begin{aligned} H_{nm}^{k+1} = H_{nm}^k + \frac{\delta_m^k}{\lambda} H_{nm}^k \Bigg[& \left(\sum_l \left(\frac{Y_{nl}}{(H^k X)_{nl}} \right)^\lambda X_{lm}^T \right) - \gamma H_{nm}^k \\ & - \sum_i H_{im}^k \left(\sum_l \left(\frac{Y_{il}}{(H^k X)_{il}} \right)^\lambda X_{lm}^T - \gamma H_{im}^k \right) \Bigg] \end{aligned} \tag{11.172}$$

It can also be written in a condensed form:

$$H_{nm}^{k+1} = H_{nm}^{k} + \frac{\delta_m^k}{\lambda} H_{nm}^{k} \left\{ \left[\left(\frac{Y}{(H^k X)} \right)^{\lambda} X^T - \gamma H^k \right]_{nm} - \left[\left(H^k \right)^T \left(\frac{Y}{(H^k X)} \right)^{\lambda} X^T - \gamma \left(H^k \right)^T H^k \right]_{mm} \right\} \quad (11.173)$$

We can easily verify that the $\lambda = 1$ case corresponding to the Kullback-Leibler divergence can be found as well by using the "β" divergences.

- Multiplicative form.

By denoting:

$$U_{nm}^{k} = \left[\left(\frac{Y}{(H^k X)} \right)^{\lambda} X^T - \gamma H^k \right]_{nm} \quad (11.174)$$

The algorithm (11.172) is written:

$$H_{nm}^{k+1} = H_{nm}^{k} + \frac{\delta_m^k}{\lambda} H_{nm}^{k} \left\{ U_{nm}^{k} - \sum_i \left(H_{mi}^{k} \right)^T U_{im}^{k} \right\} \quad (11.175)$$

We proceed to the shift:

$$\left[U_{nm}^{k} \right]_d = U_{nm}^{k} - \min_n \left(U_{nm}^{k} \right) + \epsilon > 0 \quad (11.176)$$

Taking into account the properties of H, the algorithm is written as follows:

$$H_{nm}^{k+1} = H_{nm}^{k} + \frac{\delta_m^k}{\lambda} H_{nm}^{k} \left\{ \left[U_{nm}^{k} \right]_d - \sum_i \left(H_{mi}^{k} \right)^T \left[U_{im}^{k} \right]_d \right\} \quad (11.177)$$

This makes it possible to write an algorithm which will be at the basis of the purely multiplicative algorithm:

$$H_{nm}^{k+1} = H_{nm}^{k} + \frac{\delta_m^k}{\lambda} H_{nm}^{k} \left\{ \frac{\left[U_{nm}^{k} \right]_d}{\sum_i \left(H_{mi}^{k} \right)^T \left[U_{im}^{k} \right]_d} - 1 \right\} \quad (11.178)$$

The purely multiplicative algorithm will be obtained by choosing a descent step size $\frac{\delta_m^k}{\lambda} = 1$.

*** Régularization on X.**

The basic algorithm is given by (11.166) and the expressions of the involved gradients are given by (11.71) and (11.171).
Considering these expressions and the constraint $\sum_l X_{lm}^k = \sum_l Y_{lm}$, some simplifications occur in the expression of the basic algorithm which can be rewritten in the simplified form:

$$X_{nm}^{k+1} = X_{nm}^k + \delta_m^k X_{nm}^k \left[\left(\sum_l Y_{lm} \right) U_{nm}^k - \sum_i X_{im}^k U_{im}^k \right] \tag{11.179}$$

With:

$$\begin{aligned} U_{nm}^k =& \frac{1}{\lambda} \sum_j H_{nj}^T \left(\frac{P_{jm}}{Q_{jm}} \right)^\lambda \\ &+ \mu \left(X_{nm}^k - A_m^2 \|Y_m\|_1 \right) \left(A_m^2 \|Y_m\|_1^2 - \|X_m\|_2^2 \right) \end{aligned} \tag{11.180}$$

In this expression we take into account the fact that:

$$Q = H^{k+1} X^k \tag{11.181}$$

- Multiplicative form.

To obtain such a form, one carries out, as already proposed, the shift which allows to make the expression of U_{nm}^k positive whatever the component "n" considered:

$$\left[U_{nm}^k \right]_d = U_{nm}^k - \min_n \left(U_{nm}^k \right) + \epsilon \tag{11.182}$$

The algorithm (11.179) will be written identically with the shifted quantities. We can then propose a pseudo multiplicative algorithm which is written as follows:

$$X_{nm}^{k+1} = X_{nm}^k + \delta_m^k X_{nm}^k \left[\frac{\left(\sum_l Y_{lm} \right) \left[U_{nm}^k \right]_d}{\sum_i X_{im}^k \left[U_{im}^k \right]_d} - 1 \right] \tag{11.183}$$

This makes it possible to obtain, by taking a descent step size equal to 1, the purely multiplicative form:

$$X_{nm}^{k+1} = X_{nm}^k \left[\frac{\left(\sum_l Y_{lm} \right) \left[U_{nm}^k \right]_d}{\sum_i X_{im}^k \left[U_{im}^k \right]_d} \right] \tag{11.184}$$

11.9.2 Regularized "Beta" divergence.

*** Régularization on H.**

The basis algorithm is given by (11.161).
The gradients involved in this algorithm are given respectively by (11.77) and (11.163).
Using a procedure analogous to that of the previous paragraph, taking into account the simplifications related to the property $\sum_i H_{im}^k = 1$, and with $X \equiv X^k$, we introduce the expression:

$$U_{nm}^k = \sum_j \left[\left(H^k X \right)_{nj}^{\lambda-2} Y_{nj} - \left(H^k X \right)_{nj}^{\lambda-1} \right] X_{jm}^T - \gamma H_{nm}^k \tag{11.185}$$

Then, the regularized algorithm is written according to (11.32):

$$H_{nm}^{k+1} = H_{nm}^k + \delta_m^k H_{nm}^k \left(U_{nm}^k - \sum_i H_{im}^k U_{im}^k \right) \tag{11.186}$$

Or also in condensed form:

$$H_{nm}^{k+1} = H_{nm}^k + \delta_m^k H_{nm}^k \left\{ U_{nm}^k - \left[\left(H^k \right)^T U^k \right]_{mm} \right\} \tag{11.187}$$

The case $\lambda = 1$ corresponding to the Kullback-Leibler divergence allows us to recover the result already obtained for the "α" divergence, while $\lambda = 2$ corresponds to the mean square deviation.
In order to obtain a multiplicative algorithm, we perform the shift:

$$\left[U_{nm}^k \right]_d = U_{nm}^k - \min_n \left(U_{nm}^k \right) + \epsilon \tag{11.188}$$

The multiplicative algorithm then takes a form similar to (11.178), subject to choosing a descent step size equal to $\delta_m^k = 1 \; \forall k, m$.

*** Régularization on X.**

The basis algorithm is given by (11.166) and the expressions for the involved gradients are given by (11.79) and (11.171).
Taking into account these expressions and the constraint $\sum_l X_{lm}^k = \sum_l Y_{lm}$, some simplification occurs in the expression of the basis algorithm which can be rewritten in the simplified form:

$$X_{nm}^{k+1} = X_{nm}^k + \delta_m^k X_{nm}^k \left[\left(\sum_l Y_{lm} \right) U_{nm}^k - \sum_i X_{im}^k U_{im}^k \right] \tag{11.189}$$

With:

$$U_{nm}^{k} = \sum_{j} H_{nj}^{T} \left[P_{jm} \left(Q_{jm} \right)^{\lambda-2} - \left(Q_{jm} \right)^{\lambda-1} \right] + \mu \left(X_{nm}^{k} - A_{m}^{2} \| Y_{m} \|_{1} \right) \left(A_{m}^{2} \| Y_{m} \|_{1}^{2} - \| X_{m} \|_{2}^{2} \right) \tag{11.190}$$

In this expression we take into account the fact that:

$$H \equiv H^{k+1} \quad ; \quad Q = H^{k+1} X^{k} \tag{11.191}$$

- Multiplicative form.

To obtain such a form, one performs, as already proposed, the shift that allows to make the expression of U_{nm}^{k} positive whatever the component "n" considered:

$$\left[U_{nm}^{k} \right]_{d} = U_{nm}^{k} - \min_{n} \left(U_{nm}^{k} \right) + \epsilon \tag{11.192}$$

The algorithm (11.179) will be written identically with the shifted quantities. We can then propose a pseudo multiplicative algorithm which is written as follows:

$$X_{nm}^{k+1} = X_{nm}^{k} + \delta_{m}^{k} X_{nm}^{k} \left[\frac{\left(\sum_{l} Y_{lm} \right) \left[U_{nm}^{k} \right]_{d}}{\sum_{i} X_{im}^{k} \left[U_{im}^{k} \right]_{d}} - 1 \right] \tag{11.193}$$

This allows us to obtain, by taking a descent step size equal to 1, the purely multiplicative form:

$$X_{nm}^{k+1} = X_{nm}^{k} \left[\frac{\left(\sum_{l} Y_{lm} \right) \left[U_{nm}^{k} \right]_{d}}{\sum_{i} X_{im}^{k} \left[U_{im}^{k} \right]_{d}} \right] \tag{11.194}$$

11.10 Regularization. Scale invariant divergences.

In this case, the resolution of the regularized problem implies the minimization with respect to H and X of a functional founded on the use of invariant divergences, which is written:

$$JI\left(H, X\right) = DI\left(Y \| HX\right) + \gamma DIH\left(H\right) + \mu DIX\left(X\right) \tag{11.195}$$

The minimization is carried out under the non-negativity and sum constraints already specified.
The "γ" and "μ" coefficients are the positive regularization factors.

The iterative process of minimization being alternately on H and then on X (for example).
As far as the H is concerned, we operate on a composite divergence:

$$DCIH = DI\left(Y\|HX\right) + \gamma DIH\left(H\right) \tag{11.196}$$

Regarding the "*data consistency*" term $DI\left(Y\|HX\right)$, since the invariance with respect to $Q = HX$ results in an invariance with respect to X, but not with respect to H, there's nothing particular to be expected from using an invariant divergence concerning the iterations on H.

Thus, regardless of the choice of "*data consistency*" term, the minimization with respect to H will necessarily involve a change in variables as already indicated; more generally, concerning H, one can think that it is not even necessary to introduce an invariant divergence inasmuch as the $D\left(Y\|HX\right)$ divergence reflecting the "*data consistency*" and the corresponding invariant divergence $DI\left(Y\|HX\right)$ have the same minimum.
As a consequence, a regularization on the columns of H using an invariant form of the penalty term is not absolutely necessary, but, for the sake of homogeneity, it is still possible to operate by a variable change on H in $DCIH$.

Concerning X, we operate on a composite divergence:

$$DCIX = DI\left(Y\|HX\right) + \mu DIX\left(X\right) \tag{11.197}$$

For these divergences denoted DI, we rely on the properties (11.82)(11.93), and more precisely, in our particular problem, on (11.93)(11.101).

An iterative algorithm on X can be written, with the method developed in **Chapter 10**, in the form given by (11.103):

$$X_{nm}^{k+1} = X_{nm}^{k} + \alpha_{m}^{k} X_{nm}^{k} \left[-\frac{\partial DCIX}{\partial X_{nm}^{k}}\right] \tag{11.198}$$

Of course, this writing implies the use of a penalty term DIX that is invariant with respect to X.
We can easily show that with this algorithm, the constraint $\sum_{n} X_{nm}^{k} = Cte\ \forall k$ is fulfilled.

The algorithm (11.198) is the iterative algorithm on X when minimizing a regularized invariant divergence.

*** Penalty term on X. Invariant form.**

The regularization with respect to X reflects the "*sparsity*" of the components in a column of the X matrix.
An invariant form of the penalty term, inferred from (11.169) is written:

$$DIX\,(X) = \frac{1}{2}\sum_j \left[\frac{\|X_j\|_2^2}{\|X_j\|_1^2} - A^2\right]^2 \tag{11.199}$$

The gradient with respect to X will be written as follows:

$$\frac{\partial DIX\,(X)}{\partial X_{nm}} = \left(A^2 - \frac{\|X_m\|_2^2}{\|X_m\|_1^2}\right)\left(\frac{\|X_m\|_2^2}{\|X_m\|_1^2} - \frac{X_{nm}}{\|X_m\|_1}\right)\frac{1}{\|X_m\|_1} \tag{11.200}$$

We can easily verify that:

$$\sum_n X_{nm}\frac{\partial DIX\,(X)}{\partial X_{nm}} = 0 \tag{11.201}$$

11.10.1 Regularized invariant "Alpha" divergence.

The "*data attachment*" term is the Invariant Alpha divergence (11.116) or (11.117).
With $P \equiv Y$, $Q \equiv HX$ and K_{0m} given by (11.115), of which we reproduce the expression;

$$K_{0m} = \left(\frac{\sum_n P_{nm}^{\lambda} Q_{nm}^{1-\lambda}}{\sum_n Q_{nm}}\right)^{\frac{1}{\lambda}} \tag{11.202}$$

The expressions of the involved gradients are given by (11.125) and (11.200); they lead to:

$$\begin{aligned}\frac{\partial DCIX}{\partial X_{nm}} =& \frac{1}{\lambda}\sum_j H_{nj}^T\left[K_{0m} - K_{0m}^{1-\lambda}P_{jm}^{\lambda}Q_{jm}^{-\lambda}\right] \\ &+ \mu\left(A^2 - \frac{\|X_m\|_2^2}{\|X_m\|_1^2}\right)\left(\frac{\|X_m\|_2^2}{\|X_m\|_1^2} - \frac{X_{nm}}{\|X_m\|_1}\right)\frac{1}{\|X_m\|_1}\end{aligned} \tag{11.203}$$

Taking into account the fact that the invariance factor depends on the iteration, the iterative algorithm on X (11.198), is written:

$$\begin{aligned}X_{nm}^{k+1} = X_{nm}^k + \alpha_m^k X_{nm}^k\Bigg\{&\frac{1}{\lambda}\sum_j H_{nj}^T\left[\left(K_{0m}^k\right)^{1-\lambda}P_{jm}^{\lambda}\left(Q_{jm}^k\right)^{-\lambda} - K_{0m}^k\right] \\ &- \mu\left(A^2 - \frac{\|X_m^k\|_2^2}{\|X_m^k\|_1^2}\right)\left(\frac{\|X_m^k\|_2^2}{\|X_m^k\|_1^2} - \frac{X_{nm}^k}{\|X_m^k\|_1}\right)\frac{1}{\|X_m^k\|_1}\Bigg\}\end{aligned} \tag{11.204}$$

In this expression, the following relationships must be taken into account:
* $H \equiv H^{k+1}$
* $Q^k \equiv H^{k+1} X^k$
* $\|X_m^k\|_1 = \|Y_m\|_1$ by initialization and properties of the algorithm.
* the invariance factor K_{0m}^k is deduced from (11.202) by introducing the expression of Q^k mentioned above.

11.10.2 Regularized invariant "Beta" divergence.

The "*data attachment*" term is the invariant Beta divergence (11.140) or (11.141).
With $P \equiv Y$, $Q \equiv HX$ and K_{0m} given by (11.139)

$$K_{0m} = \frac{\sum_n P_{nm} Q_{nm}^{\lambda-1}}{\sum_n Q_{nm}^{\lambda}} \tag{11.205}$$

The expressions of the involved gradients are given by (11.148) and (11.200); they lead to:

$$\begin{aligned}\frac{\partial DCIX}{\partial X_{nm}} =& \sum_j H_{nj}^T \left[K_{0m}^{\lambda} Q_{jm}^{\lambda-1} - K_{0m}^{\lambda-1} P_{jm} Q_{jm}^{\lambda-2}\right] \\ &+ \mu \left(A^2 - \frac{\|X_m\|_2^2}{\|X_m\|_1^2}\right) \left(\frac{\|X_m\|_2^2}{\|X_m\|_1^2} - \frac{X_{nm}}{\|X_m\|_1}\right) \frac{1}{\|X_m\|_1}\end{aligned} \tag{11.206}$$

Taking into account the variation of the invariance factor during the iteration, we will write:

$$K_{0m}^k = \frac{\sum_n P_{nm} \left(Q_{nm}^k\right)^{\lambda-1}}{\sum_n \left(Q_{nm}^k\right)^{\lambda}} \tag{11.207}$$

The iterative algorithm on X (11.198), is then written:

$$\begin{aligned}X_{nm}^{k+1} = X_{nm}^k + \alpha_m^k X_{nm}^k \Bigg\{ &\sum_j H_{nj}^T \left[\left(K_{0m}^k\right)^{\lambda-1} P_{jm} \left(Q_{jm}\right)^{\lambda-2} - \left(K_{0m}^k\right)^{\lambda} \left(Q_{jm}\right)^{\lambda-1}\right] \\ &- \mu \left(A^2 - \frac{\|X_m^k\|_2^2}{\|X_m^k\|_1^2}\right) \left(\frac{\|X_m^k\|_2^2}{\|X_m^k\|_1^2} - \frac{X_{nm}^k}{\|X_m^k\|_1}\right) \frac{1}{\|X_m^k\|_1} \Bigg\}\end{aligned} \tag{11.208}$$

In this expression, the following relationships must be taken into account:

* $H \equiv H^{k+1}$
* $Q \equiv H^{k+1}X^k$
* $\|X_m^k\|_1 = \|Y_m\|_1$ by initialization and properties of the algorithm.

11.11 Overview of the techniques implemented.

11.11.1 Iterations on H.

For any kind of divergence, with or without regularization:
a) Method of variables change.
b) The obtention of multiplicative algorithms requires shifting the components of the gradient.

11.11.2 Iterations on X.

1 - Non-invariant divergences, with or without a non-invariant regularization term.

a) Method of variables change.
b) The obtention of multiplicative algorithms requires shifting the components of the gradient.

2 - Invariant divergences, with or without an invariant regularization term.

a) Use of the specific property of invariant divergences.
b) Obtaining multiplicative algorithms with fulfilled sum constraint implies the application of the SGM method, followed by normalization.

chapter 12 - Application to Blind Deconvolution.

12.1 Introduction.

In this chapter, we do not attempt to revisit the well-known difficulties of signal deconvolution problems; these aspects have been widely developed in the literature dedicated to inverse problems [31] [16], and more precisely to the deconvolution of images and spectra [3], [50], [51].
In a general context, the use of scale invariant divergences to obtain multiplicative algorithms that simultaneously ensure non-negativity and the sum constraint, implies, at each iteration, to carry out a normalization that will not modify the value of the invariant divergence considered.
However, with regard to particular the problem of blind deconvolution, and more precisely with regard to the respect of the non-negativity constraints, and especially the sum constraints, the use of scale invariant divergences and of their properties leads to particularly well adapted reconstruction algorithms, without the help of a normalization procedure.
On the other hand, whatever the type of divergence used, the method of variables change makes it possible to obtain multiplicative algorithms taking into account the sum constraints.

Consequently, the first part of this chapter is devoted to the use of invariant divergences and to their minimization within the framework of blind deconvolution problems, then the second part is devoted to non-invariant divergences and the associated multiplicative algorithms.
In both cases, we will discuss the smoothness constraint regular-

ization of the solution.

To establish a parallel with the NMF problem of the previous chapter, we will take advantage of the fact that the convolution product is commutative as opposed to the matrix product.
Indeed the basic relationship is given by a Fredholm integral of the first kind:

$$\mathbf{Y}(r,s) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \mathbf{H}(r,s,\xi,\eta)\,\mathbf{X}(\xi,\eta)\,d\xi d\eta \tag{12.1}$$

In the case of convolution, the kernel $\mathbf{H}$ is invariant under translation:

$$\mathbf{H}(r,s,\xi,\eta) = \mathbf{H}(r-\xi, s-\eta) \tag{12.2}$$

We denote:

$$\mathbf{H} \otimes \mathbf{X} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \mathbf{H}(r-\xi, s-\eta)\,\mathbf{X}(\xi,\eta)\,d\xi d\eta \tag{12.3}$$

In imaging, in the case of classical deconvolution $\mathbf{Y}(r,s)$ is the measured image, usually corrupted by noise, $\mathbf{H}(r,s,\xi,\eta)$ is the supposedly known, non-negative system Point Spread Function (PSF), $\mathbf{X}(\xi,\eta)$ is the unknown non-negative image we're trying to reconstruct.
Moreover, the Point Spread Function of the system, i.e. the kernel of the integral equation, has generally an integral equal to 1.

$$\int\int \mathbf{H}(u,v)\,dudv = 1 \tag{12.4}$$

As a result of which:

$$\int\int \mathbf{Y}(r,s)\,drds = \int\int \mathbf{X}(\xi,\eta)\,d\xi d\eta \tag{12.5}$$

The "simple" deconvolution problem is similar to the linear unmixing problem discussed in the previous chapter.
We can notice that in this problem, the non-negativity of the $\mathbf{X}$ solution and the relation (12.5) are the constraints.
On the other hand, in the blind deconvolution problem, we dispose only of the measured image $\mathbf{Y}$, and the unknowns are the kernel $\mathbf{H}$ and the true object $\mathbf{X}$.
The constraints are the non-negativity of $\mathbf{H}$ and $\mathbf{X}$ on the one hand, and the relations (12.4) and (12.5) on the other hand.
Of course, despite the constraints, the problems associated with multiple

solutions are not completely solved.
This problem is similar to the NMF problem in the previous chapter, with the already mentioned advantage of the commutativity of the convolution product. In the convolution problem, the measured image, the unknown object and the impulse response are both pixelated and contained in arrays of the same dimensions.
The basic relationship:

$$\mathbf{Y} = \mathbf{H} \otimes \mathbf{X} \;\Leftrightarrow\; \mathbf{Y}(r,s) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \mathbf{H}(r-\xi, s-\eta)\, \mathbf{X}(\xi,\eta)\, d\xi d\eta \quad (12.6)$$

is written after discretization, in the vector form:

$$y = Hx \quad (12.7)$$

In this matrix form, the measured image of size $(N * N)$ is lexicographically ordered as a vector "y" of length $\left(N^2\right)$; the same applies to the unknown object "x", then, "H" is a circulating Toeplitz block matrix of dimension $\left(N^2 * N^2\right)$.
Given the commutativity of the convolution product, we also have:

$$\mathbf{Y} = \mathbf{X} \otimes \mathbf{H} \Leftrightarrow \mathbf{Y}(r,s) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \mathbf{X}(r-\xi, s-\eta)\, \mathbf{H}(\xi,\eta)\, d\xi d\eta \quad (12.8)$$

which results in:

$$y = Xh \quad (12.9)$$

Here, vectors "h" and "y" of length $\left(N^2\right)$ correspond to the lexicographical organization of the arrays $\mathbf{H}$ and $\mathbf{Y}$, then, X is a circulating Toeplitz block matrix of size $\left(N^2 * N^2\right)$.
By using the notations already introduced in the analysis of the divergences, we will have:

$$p \equiv y \;\;;\;\; q \equiv Hx \equiv Xh \quad (12.10)$$

Referring to the notations of relations (12.7) (12.9) (12.10), the resolution of the blind deconvolution problem consists in recovering the PSF "h" and the object "x" by minimizing a divergence $D(p\|q)$ under the constraints:

$$x_i \geq 0 \;\forall i \;\; ; h_i \geq 0 \;\forall i \;\; ; \sum_i h_i = 1 \;\; ; \sum_i x_i = \sum_i y_i \quad (12.11)$$

Due to the lexicographical organization of the images (tables), all considerations and operations concerning divergences between

vectors are directly applicable in deconvolution or blind deconvolution problems; the image aspects (tables) will only appear in calculations using Fourier transforms.

Note that during calculations, expressions of the form $\left(H^T x\right)$ and $\left(X^T h\right)$ will be used; these operations expressed in the context of continuous functions correspond to:

$$H^T u \equiv \int\int \mathbf{H}\left(-r,-s,\xi,\eta\right)\mathbf{U}\left(\xi,\eta\right)d\xi d\eta \tag{12.12}$$

with, in the case of convolution:

$$\mathbf{H}\left(-r,-s,\xi,\eta\right) = \mathbf{H}\left(-r-\xi,-s-\eta\right) \tag{12.13}$$

and

$$X^T u \equiv \int\int \mathbf{X}\left(-r,-s,\xi,\eta\right)\mathbf{U}\left(\xi,\eta\right)d\xi d\eta \tag{12.14}$$

with, in the case of convolution:

$$\mathbf{X}\left(-r,-s,\xi,\eta\right) = \mathbf{X}\left(-r-\xi,-s-\eta\right) \tag{12.15}$$

12.1.1 Some details related to the calculus.

To perform the calculus of the convolution products, it is interesting to use the Fourier Transforms noted $\mathcal{F}$; indeed, the classical property of the convolution product allows to write [85]:

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\mathbf{H}\left(r-\xi,s-\eta\right)\mathbf{X}\left(\xi,\eta\right)d\xi d\eta = \mathcal{F}^{-1}\left[\mathcal{F}\left(\mathbf{H}\right)\mathcal{F}\left(\mathbf{X}\right)\right] \tag{12.16}$$

and moreover, we have:

$$\mathbf{H}\left(-r,-s\right) = \mathcal{F}\left[\mathcal{F}\left(\mathbf{H}\left(r,s\right)\right)\right] \quad ; \quad \mathbf{X}\left(-r,-s\right) = \mathcal{F}\left[\mathcal{F}\left(\mathbf{X}\left(r,s\right)\right)\right] \tag{12.17}$$

In doing so, we will only handle tables of the size of the object, image, and PSF.
As a result, in the reconstruction iterative process, the unknown images (tables) are updated as a whole.

12.2 Algorithmic method.

The method described applies to all divergences, whether invariant or not. Leaving aside the regularization aspects, the problem is posed as a constrained minimization problem of a divergence between the measures $p \equiv y$ and the convolutional model $q \equiv Hx \equiv Xh \equiv [h \otimes x]$, which reflects the "*data attachment*".

$\underbrace{\min}_{x,h} DI\left(y \| [h \otimes x]\right)$

s.t. $\quad x_i > 0 \ \forall i \ ; \ \sum_i x_i = \sum_i y_i \ \ \text{et} \ \ h_i > 0 \ \forall i \ ; \ \sum_i h_i = 1.$

The proposed blind deconvolution method is structurally inspired by the one initially proposed in [7] and extended by the use of iterative algorithms in [59] [60] for example.
This method, in which we operate alternately on the two unknowns is analogous to the one developed for MMF in **chapter 11**.

It can be synthesized, for example, according to the following scheme:
1 - x^k , $h^k \rightarrow x^k$, h^{k+1}
2 - x^k , $h^{k+1} \rightarrow x^{k+1}$, h^{k+1}

This is the scheme that will be considered in further calculations.
Note that we might as well iterate first on "x" and then on "h".
In blind deconvolution, as in simple deconvolution, the updating is done on all the components of "x" simultaneously; similarly, the updating to "h" is done on all the components simultaneously.
This last point differs from what is done in the case of the NMF where the updating $H^k \rightarrow H^{k+1}$ is carried out successively column by column as well as the updating $X^k \rightarrow X^{k+1}$.
The alternative iterative method inspired by Uzawa's algorithm suggested in **Chapter 11** for NMF, can also be considered for blind deconvolution; it was used in [59], [60]. Such a method will not be developed here.

12.3 Application to scale invariant divergences.

12.3.1 Property of scale invariant divergences with respect to "q".

The use of invariant divergences with respect to "q" is particularly interesting for the blind deconvolution problem, because their use associated with

the commutativity of the convolution product allows to take into account simply the sum constraints on the two unknowns, i.e. $\sum_i h_i = 1$; $\sum_i x_i = \sum_i y_i$.
Let's recall the property of invariant divergences arising here (3.7):

$$\sum_m q_m \left[\frac{\partial DI\left(p\|q\right)}{\partial q}\right]_m = 0 \tag{12.18}$$

*** Initialization.**

The algorithm is initialized with x^0 / $\sum_i x_i^0 = \sum_i y_i$ and h^0 / $\sum_i h_i^0 = 1$.

*** Itérations on "h" - Non-multiplicative form.**

We have at our disposal x^k, h^k; here, $q = X^k h^k$.
For a divergence invariant with respect to "q" denoted $DI\left(p\|q\right)$, the iterative algorithm on "h" is written :

$$h_l^{k+1} = h_l^k + \alpha_l^k h_l^k \left[-\frac{\partial DI\left(p\|q\right)}{\partial h_l^k}\right] \tag{12.19}$$

The gradient with respect to "h" is given by:

$$\frac{\partial DI\left(p\|q\right)}{\partial h_l} = \sum_i \frac{\partial DI\left(p_i\|q_i\right)}{\partial q_i}\frac{\partial q_i}{\partial h_l} \tag{12.20}$$

With

$$\frac{\partial q_i}{\partial h_l} = X_{il} \tag{12.21}$$

We have:

$$\frac{\partial DI\left(p\|q\right)}{\partial h_l} = \sum_i \frac{\partial DI\left(p_i\|q_i\right)}{\partial q_i} X_{il} = \sum_i X_{li}^T \frac{\partial DI\left(p_i\|q_i\right)}{\partial q_i} \tag{12.22}$$

And finally:

$$\frac{\partial DI\left(p\|q\right)}{\partial h_l} = \left[X^T \frac{\partial DI\left(p\|q\right)}{\partial q}\right]_l \tag{12.23}$$

The iterative algorithm on "h" is then written with $x \equiv x^k \Leftrightarrow X \equiv X^k$ and $q = X^k h^k$:

$$h_l^{k+1} = h_l^k + \alpha_l^k h_l^k \left[-X^T \frac{\partial DI\left(p\|q\right)}{\partial q}\right]_l \tag{12.24}$$

As in all analogous algorithms, as indicated in **Chapter 10**, we must in a first step calculate the step size $\left[\alpha_l^k\right]_{Max}$ ensuring the non-negativity of all the components of h^{k+1}, then calculate by a one-dimensional search method over the interval $\left\{0, \left[\alpha_l^k\right]_{Max}\right\}$, the descent step size α^k ensuring the convergence of the algorithm:

$$h_l^{k+1} = h_l^k + \alpha^k h_l^k \left[-X^T \frac{\partial DI\left(p\|q\right)}{\partial q}\right]_l \tag{12.25}$$

For the whole "h" vector, we can write in a more explicit way:

$$h^{k+1} = h^k + \alpha^k h^k \odot \left[-X^T \frac{\partial DI\left(p\|q\right)}{\partial q}\right] \tag{12.26}$$

In this writing the symbol $\odot$ is the pointwise product of two vectors.
With such an algorithm, considering the property (3.7), we obtain $\sum_l h_l^{k+1} = \sum_l h_l^k$, whatever the descent step size α^k, as shown by the following simple calculation.

*** Démonstration of $\sum_l h_l^{k+1} = \sum_l h_l^k$**

From (12.19), we have:

$$\sum_l h_l^{k+1} = \sum_l h_l^k + \alpha^k \sum_l \left\{h_l^k \left[-X^T \frac{\partial DI\left(p\|q\right)}{\partial q}\right]_l\right\} \tag{12.27}$$

By explaining:

$$\left[X^T \frac{\partial DI\left(p\|q\right)}{\partial q}\right]_l = \sum_m X_{lm}^T \left[\frac{\partial DI\left(p\|q\right)}{\partial q}\right]_m \tag{12.28}$$

Putting in the previous equation, it comes:

$$\sum_l h_l^{k+1} = \sum_l h_l^k + \alpha^k \sum_l \left\{h_l^k \left[-\sum_m X_{lm}^T \left[\frac{\partial DI\left(p\|q\right)}{\partial q}\right]_m\right]\right\} \tag{12.29}$$

By swapping the order of the summations, it comes:

$$\sum_l h_l^{k+1} = \sum_l h_l^k + \alpha^k \sum_m \left[-\frac{\partial DI\left(p\|q\right)}{\partial q}\right]_m \sum_l h_l^k X_{lm}^T \tag{12.30}$$

That is:

$$\sum_l h_l^{k+1} = \sum_l h_l^k + \alpha^k \sum_m \left[-\frac{\partial DI\left(p\|q\right)}{\partial q}\right]_m \sum_l X_{ml} h_l^k \tag{12.31}$$

Given the property (3.7) the second term of the second member of the previous equation is zero, hence the announced result.

*** Itérations on "h" - Multiplicative form.**

In order to obtain a multiplicative form of the algorithm, from the relations (12.19) (12.25), we strictly apply the SGM method and we carry out in a first step the decomposition:

$$\left[-\frac{\partial DI\left(p\|q\right)}{\partial h^k}\right]_l = \left[-X^T \frac{\partial DI\left(p\|q\right)}{\partial q}\right]_l = U_l^k - V_l^k \tag{12.32}$$

Avec $U_l^k > 0$, $V_l^k > 0$.
We then write in the next step:

$$\widetilde{h}_l^{k+1} = h_l^k + \alpha^k h_l^k \left[\frac{U_l^k}{V_l^k} - 1\right]_l \tag{12.33}$$

Then, by choosing a descent step size equal to 1 for any "k", we get:

$$\widetilde{h}_l^{k+1} = h_l^k \left[\frac{U_l^k}{V_l^k}\right]_l \tag{12.34}$$

At this point, we don't have $\sum_l \widetilde{h}_l^{k+1} = \sum_l h_l^k = 1$.
However, the invariance properties of the divergences allow to carry out, in a last step, a normalization which does not modify the value of the divergence, and we finally have:

$$h_l^{k+1} = \frac{\widetilde{h}_l^{k+1}}{\sum_l \widetilde{h}_l^{k+1}} \tag{12.35}$$

*** Itérations on "x" - Non multiplicative form.**

At this point we have x^k, h^{k+1}; here, $q = H^{k+1}x^k$.
The iterative algorithm on "x" is written as follows:

$$x_l^{k+1} = x_l^k + \beta_l^k x_l^k \left[-\frac{\partial DI\left(p\|q\right)}{\partial x_l^k}\right] \tag{12.36}$$

The gradient with respect to "x" is written:

$$\frac{\partial DI\left(p\|q\right)}{\partial x_l} = \sum_i \frac{\partial DI\left(p_i\|q_i\right)}{\partial q_i} \frac{\partial q_i}{\partial x_l} \tag{12.37}$$

With:

$$\frac{\partial q_i}{\partial x_l} = H_{il} \tag{12.38}$$

It comes:

$$\frac{\partial DI\left(p\|q\right)}{\partial x_l} = \sum_i \frac{\partial DI\left(p_i\|q_i\right)}{\partial q_i} H_{il} = \sum_i H_{li}^T \frac{\partial DI\left(p_i\|q_i\right)}{\partial q_i} \tag{12.39}$$

And finally:

$$\frac{\partial DI\left(p\|q\right)}{\partial x_l} = \left[H^T \frac{\partial DI\left(p\|q\right)}{\partial q}\right]_l \tag{12.40}$$

The iterative algorithm on "x" is then written with $h \equiv h^{k+1} \Leftrightarrow H \equiv H^{k+1}$ and $q = H^{k+1}x^k$:

$$x_l^{k+1} = x_l^k + \beta_l^k x_l^k \left[-H^T \frac{\partial DI\left(p\|q\right)}{\partial q}\right]_l \tag{12.41}$$

As in all analogous algorithms, as indicated in **Chapter 10**, we must in a first step calculate the step size $\left[\beta_l^k\right]_{Max}$ which ensures the non-negativity of all the components of x^{k+1}, then calculate by a one-dimensional search method over the interval $\left\{0, \left[\beta_l^k\right]_{Max}\right\}$, the descent step size β^k ensuring the convergence of the algorithm:

$$x_l^{k+1} = x_l^k + \beta^k x_l^k \left[-H^T \frac{\partial DI\left(p\|q\right)}{\partial q}\right]_l \tag{12.42}$$

A demonstration similar to the one developed for "h" makes it easy to establish that with this type of algorithm one obtains $\sum_l x_l^{k+1} = \sum_l x_l^k$, regardless of the descent step size β^k.
An expression of this algorithm for the "x" vector as a whole, is:

$$x^{k+1} = x^k + \beta^k x^k \odot \left[-H^T \frac{\partial DI\left(p\|q\right)}{\partial q}\right] \tag{12.43}$$

In this expression, the symbol *odot* represents the pointwise product.
At this point, we have obtained h^{k+1}, x^{k+1}.

*** Iterations on "x" - Multiplicative form.**

To obtain this form, we strictly follow what has been described for "h" (the decomposition here is about $\left[-H^T \frac{\partial DI(p\|q)}{\partial q}\right]_l$), until we obtain $\widetilde{x}_l^{k+1}$, after which, the normalization step is written:

$$x_l^{k+1} = \frac{\widetilde{x}_l^{k+1}}{\sum_l \widetilde{x}_l^{k+1}} \sum_l y_l \tag{12.44}$$

Such normalization will not change the value of the divergence.

12.4 Analysis of 2 particular cases.

12.4.1 Invariant "Beta" divergence.

In **Chapter 5**, we developed the invariance aspects concerning the Beta divergence.
The initial divergence is given by (5.20).
The expression of the invariance factor K is obtained explicitly, it is therefore the nominal value which is written as follows (5.45):

$$K_0 = \frac{\sum_i p_i q_i^{\lambda-1}}{\sum_i q_i^{\lambda}} \tag{12.45}$$

And the corresponding divergence invariant by scale change on "q" is written (5.47):

$$BI\left(p\|q\right) = \frac{1}{\lambda\left(\lambda-1\right)}\left[\sum_i p_i^{\lambda} - \sum_i K_0^{\lambda} q_i^{\lambda}\right] \tag{12.46}$$

Its gradient with respect to "q" has been previously written in the form (5.49):

$$\frac{\partial BI\left(p\|q\right)}{\partial q_j} = K_0^{\lambda}\left[q_j^{\lambda-1} - \left(K_0\right)^{-1} p_j q_j^{\lambda-2}\right] \tag{12.47}$$

The algorithm can then be written:
Initialization:

$$h_i^0 \;\;/\;\; \sum_i h_i^0 = 1 \;\;;\;\;\; x_i^0 \;\;/\;\; \sum_i x_i^0 = \sum_i y_i \tag{12.48}$$

*** Iteration on "h" according to (12.19):**

We know h^k, $x^k \rightarrow X^k \equiv X$, $p = y$, $q = X^k h^k \left(= H^k x^k\right)$.
The invariance factor is iteration dependent and is written as:

$$K_0^k = \frac{\sum_i p_i q_i^{\lambda-1}}{\sum_i q_i^{\lambda}} \tag{12.49}$$

$$h_l^{k+1} = h_l^k + \alpha^k \left(K_0^k\right)^{\lambda} h_l^k \left\{\sum_j X_{lj}^T \left[\left(K_0^k\right)^{-1} p_j^{\lambda} q_j^{\lambda-2} - q_j^{\lambda-1}\right]\right\} \tag{12.50}$$

At this point, we have obtained $h^{k+1} \rightarrow H \equiv H^{k+1}$, we know $p = y$, x^k, $q = H^{k+1} x^k$, then, updating K_0^k according to (12.49), we iterate on "x".

*** Iteration on "x" according to (12.42):**

$$x_l^{k+1} = x_l^k + \alpha^k \left(K_0^k\right)^{\lambda} x_l^k \left\{ \sum_j H_{lj}^T \left[\left(K_0^k\right)^{-1} p_j^{\lambda} q_j^{\lambda-2} - q_j^{\lambda-1} \right] \right\} \tag{12.51}$$

12.4.2 Invariant mean square deviation.

For the root mean square deviation which corresponds to $\lambda = 2$ for the Beta divergence, the nominal invariance factor K_0 can be explicitly computed and is written as follows:

$$K_0 = \frac{\sum_l p_l q_l}{\sum_l q_l^2} \tag{12.52}$$

The corresponding invariant divergence is given by:

$$MCI\left(p \| q\right) = \frac{1}{2} \sum_i \left[p_i^2 - \left(\frac{\sum_l p_l q_l}{\sum_l q_l^2} \right)^2 q_i^2 \right] \tag{12.53}$$

Its gradient with respect to "q" is written as follows:

$$\frac{\partial MCI\left(p \| q\right)}{\partial q_j} = \left(\frac{\sum_l p_l q_l}{\sum_l q_l^2} \right)^2 q_j - \left(\frac{\sum_l p_l q_l}{\sum_l q_l^2} \right) p_j = K_0^2 q_j - K_0 p_j \tag{12.54}$$

According to the relation (12.19), after discussion about the descent step size and taking into account the variarion of K_0 during iterations, it comes for the entire vector "h":

$$h_l^{k+1} = h_l^k + \alpha^k h_l^k \left\{ X^T \left[K_0^k p - \left(K_0^k\right)^2 q \right] \right\}_l \tag{12.55}$$

In this expression, $x \equiv x^k \Leftrightarrow X \equiv X^k$, $q = H^k x^k = X^k h^k$, and K_0^k is expressed as:

$$K_0^k = \frac{\sum_l y_l \left(H^k x^k\right)_l}{\sum_l \left(H^k x^k\right)_l^2} \tag{12.56}$$

We then iterate on "x" according to the relationship (12.42); after discussion about the descent step size and with $h \equiv h^{k+1} \Leftrightarrow H \equiv H^{k+1}$ et $q = H^{k+1} x^k$, it comes for the entire vector "x":

$$x_l^{k+1} = x_l^k + \beta^k x_l^k \left\{ H^T \left[K_0^k p - \left(K_0^k\right)^2 q \right] \right\}_l \tag{12.57}$$

The invariance factor K_0^k variable during iterations, will be expressed by:

$$K_0^k = \frac{\sum_l y_l \left(H^{k+1} x^k\right)_l}{\sum_l \left(H^{k+1} x^k\right)_l^2} \tag{12.58}$$

12.4.3 Invariant Alpha divergence.

In **Chapter 5**, we discussed the invariance aspects concerning the Alpha divergence.
The initial divergence is given by (5.15).
The expression of the invariance factor K is obtained explicitly, so it is the nominal value which is written as (5.27):

$$K_0 = \left(\frac{\sum_i p_i^\lambda q_i^{1-\lambda}}{\sum_i q_i} \right)^{\frac{1}{\lambda}} \tag{12.59}$$

And the corresponding divergence invariant by scale change on "q" is written (5.29):

$$AI\left(p\|q\right) = \frac{1}{\lambda - 1} \left[\sum_i K_0 q_i - \sum_i p_i \right] \tag{12.60}$$

Its gradient with respect to "q" has been written in the form (5.33):

$$\frac{\partial AI\left(p\|q\right)}{\partial q_j} = \frac{1}{\lambda} \left[K_0 - K_0^{1-\lambda} p_j^\lambda q_j^{-\lambda} \right] \tag{12.61}$$

With the initialization:

$$h_i^0 \quad / \quad \sum_i h_i^0 = 1 \quad ; \qquad x_i^0 \quad / \quad \sum_i x_i^0 = \sum_i y_i \tag{12.62}$$

The algorithm can then be described as follows:

*** Iteration on "h" according to (12.19):**

We know h^k, $x^k \rightarrow X^k \equiv X$, $p = y$, $q = X^k h^k$.
The iteration-dependent invariance factor is written as:

$$K_0^k = \left(\frac{\sum_i p_i^\lambda q_i^{1-\lambda}}{\sum_i q_i} \right)^{\frac{1}{\lambda}} \tag{12.63}$$

The iteration is written for "h" as a whole:

$$h^{k+1} = h^k + \frac{\alpha^k K_0^k}{\lambda} h^k \odot \left\{ X^T \left[\left(K_0^k \right)^{-\lambda} p^\lambda q^{-\lambda} - 1 \right] \right\} \tag{12.64}$$

The symbol "$\odot$" corresponds to the pointwise multiplication of the two vectors.

At this point, we obtained h^{k+1}.

*** Iteration on "x" according to (12.42):**

We have $h^{k+1} \rightarrow H^{k+1} \equiv H$, x^k, $p = y$, $q = H^{k+1}x^k$.
The iteration-dependent invariance factor is updated according to (12.63), and it comes:

$$x^{k+1} = x^k + \frac{\alpha^k K_0^k}{\lambda} x^k \odot \left\{ H^T \left[\left(K_0^k \right)^{-\lambda} p^\lambda q^{-\lambda} - 1 \right] \right\} \tag{12.65}$$

The symbol "$\odot$" corresponds to the pointwise multiplication of the two vectors.
At this point, we obtained h^{k+1}, x^{k+1}.

12.4.4 Kullback-Leibler invariant divergence.

This divergence corresponds to the case $\lambda = 1$ of the Alpha divergence as well as the Beta divergence. This is the point in common between these two types of divergence.
The nominal invariance factor K_0 can be explicitly derived and is written:

$$K_0 = \frac{\sum_l p_l}{\sum_l q_l} \tag{12.66}$$

The corresponding invariant divergence is given by:

$$KLI\left(p \| q\right) = \sum_l p_l \sum_i \frac{p_i}{\sum_l p_l} \log \left(\frac{\sum_l q_l}{\sum_l p_l} \right) \frac{p_i}{q_i} \tag{12.67}$$

Its gradient with respect to "q" is written as follows:

$$\frac{\partial KLI\left(p \| q\right)}{\partial q_j} = \frac{\sum_l p_l}{\sum_l q_l} - \frac{p_j}{q_j} = K_0 - \frac{p_j}{q_j} \tag{12.68}$$

According to the relationship (12.19) and after discussing the step size, he comes for "h" as a whole:

$$h^{k+1} = h^k + \alpha^k h^k \odot \left[X^T \left(\frac{p}{q} - K_0^k \right) \right] \tag{12.69}$$

In this expression, $p \equiv y$, $x \equiv x^k \Leftrightarrow X \equiv X^k$ and $q = H^k x^k = X^k h^k$.
The invariance factor K_0^k will be expressed by:

$$K_0^k = \frac{\sum_l y_l}{\sum_l \left(X^k h^k \right)_l} \tag{12.70}$$

We then iterate on "x" according to the relation (12.42); after discussion on the descent step size, it comes for "x" as a whole:

$$x^{k+1} = x^k + \beta^k x^k \odot \left[H^T \left(\frac{p}{q} - K_0^k \right) \right] \tag{12.71}$$

with $p \equiv y$, $h \equiv h^{k+1} \Leftrightarrow H \equiv H^{k+1}$, $q = H^{k+1} x^k$ and K_0^k given by:

$$K_0^k = \frac{\sum_l y_l}{\sum_l \left(H^{k+1} x^k \right)_l} \tag{12.72}$$

12.5 Regularization - Scale invariant divergences.

In the context of deconvolution or blind deconvolution, we classically impose a property of smoothness of the solution; this is the regularization in the sense of Tikhonov [93].
The 2 classical methods used to introduce such a property consist in using as a penalty term either the Euclidean norm of the deviation between the solution and its mean value, or the Euclidean norm of the Laplacian of the solution.
In a first step, we develop the expressions of the penalty terms having the invariance properties, which will be associated with the "*data attachment*" term.

12.5.1 Euclidean norm of the solution.

The non-invariant penalty term for the variable "x" is written as:

$$DRQ_x = \sum_i \left(c_i - x_i \right)^2 \tag{12.73}$$

Where $c_i = 1/N^2 \ \forall i$ and where x_i is the image of dimension $N * N$ lexicographically ordered.
First we need to obtain a form that is invariant with respect to "x".
To do that, we derive the invariance factor, which is written:

$$K_0 = \frac{\sum_j c_j x_j}{\sum_j x_j^2} \tag{12.74}$$

The corresponding invariant divergence is written as follows:

$$DRQI_x = \sum_i \left(c_i - K_0 x_i \right)^2 \tag{12.75}$$

After a few calculations, the gradient with respect to "x" is expressed as follows:

$$\frac{\partial DRQI_x}{\partial x_l} = K_0^2 x_l - K_0 c_l \tag{12.76}$$

We can easily verify that we have:

$$\sum_l x_l \frac{\partial DRQI_x}{\partial x_l} = 0 \tag{12.77}$$

12.5.2 Use of the Laplacian norm of the solution.

The classical non-invariant penalty term for the variable "x" is written as:

$$DRL_x = \|Lx\|^2 = \|x - Tx\|^2 \tag{12.78}$$

The term "Lx" corresponds to the convolution of the table $\mathbf{X}$ by the Laplacian mask: $\begin{pmatrix} 0 & -1/4 & 0 \\ -1/4 & 1 & -1/4 \\ 0 & -1/4 & 0 \end{pmatrix}$.
The resulting table is written in lexicographical order.
We can thus write the vector "Lx" in the form $(x - Tx)$ where the term "Tx" is a vector which, in lexicographic order, corresponds to the convolution of $\mathbf{X}$ by the mask: $\begin{pmatrix} 0 & 1/4 & 0 \\ 1/4 & 0 & 1/4 \\ 0 & 1/4 & 0 \end{pmatrix}$.
The "x" vector corresponds to the lexicographical writing of the table $\mathbf{X}$.
We will note that under these conditions, the matrix T is symmetrical, it is normalized to 1 in columns, i.e. $\sum_i T_{ij} = 1 \ \forall j$, thus $\sum_i (Tx)_i = \sum_i x_i$.
However, the penalty term so described is not invariant with respect to "x". This aspect of the problem was discussed in **Chapter 9**.
In order to have the property of invariance, one can use, to express the discrepancy between "x" and "Tx", the divergences LAI or LBI, which are invariant to both arguments and which are given respectively by (9.11) and (9.13).
The corresponding gradients with respect to "x" are given respectively by (9.12) and (9.14).

We use here their expressions; for $LAI_a\left(x\|Tx\right)$, we have (9.12):

$$\frac{\partial LAI_a\left(x\|Tx\right)}{\partial x_l} = \frac{1}{a}\left[\frac{\sum_i T_{il}}{\sum_i \left(Tx\right)_i} - \frac{\sum_i T_{il} x_i^a \left(Tx\right)_i^{-a}}{\sum_i x_i^a \left(Tx\right)_i^{1-a}}\right] + \frac{1}{a-1}\left[\frac{x_l^{a-1}\left(Tx\right)_l^{1-a}}{\sum_i x_i^a \left(Tx\right)_i^{1-a}} - \frac{1}{\sum_i x_i}\right] \tag{12.79}$$

The case $a = 1$ which corresponds to the basis of a Kullback-Leibler divergence imply the computation of a limit which leads after some calculations to the expression:

$$\frac{\partial LAI_1\left(x\|Tx\right)}{\partial x_l} = \frac{1}{\sum_i x_i}\left[\sum_i T_{il} - \sum_i T_{il}\frac{x_i}{\left(Tx\right)_i}\right] \tag{12.80}$$

Or else by simplifying, taking into account the properties of the matrix T:

$$\frac{\partial LAI_1\left(x\|Tx\right)}{\partial x_l} = \frac{1}{\sum_i x_i}\left[1 - \sum_i T_{il}\frac{x_i}{\left(Tx\right)_i}\right] \tag{12.81}$$

One can observe that, as always with this type of divergence, $\sum_l x_l \frac{\partial}{\partial x_l} = 0$.

Similarly, as regards $LBI_b\left(x\|Tx\right)$, we have (9.14):

$$\frac{\partial LBI_b\left(x\|Tx\right)}{\partial x_l} = \frac{1}{b-1}\left[\frac{x_l^{b-1}}{\sum_i x_i^b} - \frac{\left(Tx\right)_l^{b-1}}{\sum_i x_i \left(Tx\right)_i^{b-1}}\right] + \left[\frac{\sum_i T_{il}\left(Tx\right)_i^{b-1}}{\sum_i \left(Tx\right)_i^b} - \frac{\sum_i T_{il} x_i \left(Tx\right)_i^{b-2}}{\sum_i x_i \left(Tx\right)_i^{b-1}}\right] \tag{12.82}$$

The special case $b = 1$ leads to the same result as the divergence LAI_1, whereas the case $b = 2$ which corresponds to a basis divergence which is the mean square deviation, does not imply a passage to the limit and leads to:

$$\frac{\partial LBI_2\left(x\|Tx\right)}{\partial x_l} = \frac{x_l}{\sum_i x_i^2} - \frac{\sum_i T_{il}\left(Tx\right)_i}{\sum_i \left(Tx\right)_i^2} \tag{12.83}$$

Or else, given the properties of the matrix T ($T = T^T$):

$$\frac{\partial LBI_2\left(x\|Tx\right)}{\partial x_l} = \frac{x_l}{\sum_i x_i^2} - \frac{\left[T\left(Tx\right)\right]_l}{\sum_i \left(Tx\right)_i^2} \tag{12.84}$$

We have, again, as always with this type of divergence: $\sum_l x_l \frac{\partial}{\partial x_l} = 0$.

12.6 Regularized algorithm.

The problem here is to minimize under constraints, with respect to the two unknowns "x" and "h", an invariant composite divergence of the form:

$$DCI = DI\left(y \| [h, x]\right) + \gamma DRI\left(h\right) + \mu DRI\left(x\right) \tag{12.85}$$

The coefficients "γ" and "μ" are the positive regularization factors.
For the $DI\left(y \| [h, x]\right)$ divergence representing the attachment to the data, we can use any divergence invariant with respect to $q \equiv Hx \equiv Xh$.
In contrast, as far as the terms of regularization are concerned, there are two possible cases.
* either this term translates the discrepancy to a constant, which is the case for example of the mean square deviation proposed in the previous section, then an invariant form of this term i.e. (12.75) is usable.
This is valid for any invariant divergence between a constant and the variable considered.
* or the variable is involved in the two arguments of the divergence expressing the penalty, this is the case of a Laplacian type regularization, then one must use invariant divergences with respect to its both arguments, as it is the case for divergences of type LAI or LBI developed in the preceding section.
These considerations being taken into account, the algorithmic method developed in section (12.2) applies to the composite divergence DCI without any change.

12.7 Blind deconvolution - Non-invariant divergences.

In the previous section, it was shown that the use of invariant divergences allows us to obtain multiplicative algorithms while maintaining sum constraints on the unknowns, subject to a normalization step.
In order to propose such algorithms without resorting to this procedure, we develop here Blind Deconvolution algorithms using non-invariant divergences. In order to take into account the sum constraints on "h" and on "x", we rely on the variable change method proposed in the **section 10.3.2**, and more specifically on the algorithm (10.76).
The general structure of the algorithm remains as proposed in the previous section, which consists of iterating alternately on the two unknowns.

Considering a non-invariant divergence:

$$D\left(p\|q\right)=\sum_{i}d\left(p_{i}\|q_{i}\right) \tag{12.86}$$

With $p_i \equiv y_i$ et $q_i \equiv (Hx)_i \equiv (Xh)_i$.

*** Iterations on "h" - Non-multiplicative form.**

We have $x^k \Leftrightarrow X^k$ and h^k, we are searching for h^{k+1}.
After performing the variable change $h_j = \frac{t_j}{\sum_n t_n}$, and according to (10.76), we obtain the algorithm:

$$h_{l}^{k+1}=h_{l}^{k}+\delta^{k}h_{l}^{k}\left\{\left[-\frac{\partial D\left(y\|X^{k}h^{k}\right)}{\partial h_{l}^{k}}\right]-\sum_{m}h_{m}^{k}\left[-\frac{\partial D\left(y\|X^{k}h^{k}\right)}{\partial h_{m}^{k}}\right]\right\} \tag{12.87}$$

We can easily verify that we have $\sum_l h_l^{k+1} = \sum_l h_l^k$; as a consequence, considering the initialization $\sum_j h_j^0 = 1$, we will have $\sum_l h_l^{k+1} = 1$.

*** Iterations on "h" - Multiplicative form.**

Unlike invariant divergences, the use of non-invariant divergences allows multiplicative forms of the algorithms to appear, without the need to use normalization.
This is done by shifting the components of the opposite of the gradient involved in (12.87), in order to make them all positive.

$$\left[-\frac{\partial D\left(y\|X^{k}h^{k}\right)}{\partial h^{k}}\right]_{d}=\left[-\frac{\partial D\left(y\|X^{k}h^{k}\right)}{\partial h^{k}}\right]-\underbrace{\min}_{l}\left[-\frac{\partial D\left(y\|X^{k}h^{k}\right)}{\partial h_{l}^{k}}\right]+\epsilon \tag{12.88}$$

Obviously, given the properties of "h", such an offset will not change the algorithm.
We can then write an algorithm that will allow us to obtain a purely multiplicative form:

$$h_{l}^{k+1}=h_{l}^{k}+\delta^{k}h_{l}^{k}\left\{\frac{\left[-\frac{\partial D\left(y\|X^{k}h^{k}\right)}{\partial h_{l}^{k}}\right]_{d}}{\sum_{m}h_{m}^{k}\left[-\frac{\partial D\left(y\|X^{k}h^{k}\right)}{\partial h_{m}^{k}}\right]_{d}}-1\right\} \tag{12.89}$$

The purely multiplicative form is obtained, as always, by taking a descent step size $\delta^k = 1\ \forall k$, however, there is no guarantee of convergence of such

an algorithm over which we no longer have any control.

*** Iterations on "x"- Non-multiplicative form.**

We know $h^{k+1} \Leftrightarrow H^{k+1}$ and x^k, we want to obtain x^{k+1}.
After having carried out the change of variables $x_j = \frac{u_j}{\sum_n u_n} \sum_i y_i$, referring to (10.76), one obtains the algorithm:

$$x_l^{k+1} = x_l^k + \beta^k x_l^k \left\{ \sum_i y_i \left[-\frac{\partial D\left(y \| H^{k+1} x^k\right)}{\partial x^k} \right]_l - \sum_m x_m^k \left[-\frac{\partial D\left(y \| H^{k+1} x^k\right)}{\partial x^k} \right]_m \right\} \tag{12.90}$$

Taking into account that $\sum_j x_j = \sum_j y_j$, we will have $\sum_l x_l^{k+1} = \sum_l x_l^k = \sum_j y_j$.

*** Iterations on "x" - Multiplicative form.**

As proposed for the variable "h", a multiplicative algorithm can be obtained; to do so, we shift the components of the opposite gradient in involved in (12.90), in order to make them all positive.

$$\left[-\frac{\partial D\left(y \| H^{k+1} x^k\right)}{\partial x^k} \right]_d = \left[-\frac{\partial D\left(y \| H^{k+1} x^k\right)}{\partial x^k} \right] - \underbrace{\min}_{l} \left[-\frac{\partial D\left(y \| H^{k+1} x^k\right)}{\partial x_l^k} \right] + \epsilon \tag{12.91}$$

Obviously, an offset like that will not change the algorithm.
We can then write an algorithm that will allow us to obtain a purely multiplicative form:

$$x_l^{k+1} = x_l^k + \beta^k x_l^k \left\{ \frac{\sum_i y_i \left[-\frac{\partial D\left(y \| H^{k+1} x^k\right)}{\partial x_l^k} \right]_d}{\sum_m x_m^k \left[-\frac{\partial D\left(y \| H^{k+1} x^k\right)}{\partial x_m^k} \right]_d} - 1 \right\} \tag{12.92}$$

The purely multiplicative form is obtained, as always, by taking a descent step size $\beta^k = 1 \; \forall k$, but there's no convergence guarantee for such an algorithm over which we no longer have any control.

12.7.1 Regularization.

In this case, with $q \equiv Hx \equiv Xh$, we have to minimize a composite divergence:

$$DC = D\left(y\|q\right) + \gamma DR_x\left(x\right) + \mu DR_h\left(h\right) \tag{12.93}$$

Decomposing the problem as described in the previous paragraph leads us to:
1 - For known "x", minimize with respect to "h", under constraints, a divergence of the form:

$$DC_h = D\left(y\|Xh\right) + \gamma DR_h\left(h\right) \tag{12.94}$$

2 - For known "h", minimize with respect to "x", under constraints, a divergence of the form:

$$DC_x = D\left(y\|Hx\right) + \mu DR_x\left(x\right) \tag{12.95}$$

Expressions of regularization terms.

For both "h" and "x", we use here a smoothness constraint regularization of the solution which is written:
either in the form (12.73):

$$DRQ_x = \frac{1}{2}\sum_i \left(c - x_i\right)^2 \tag{12.96}$$

Where $c = \frac{\sum_i y_i}{N}$; "N" is the number of components of "y" and/or "x", i.e. the number of pixels in the measured and/or reconstructed image.

with:

$$\frac{\partial DRQ_x}{\partial x_l} = x_l - c \tag{12.97}$$

, or in the form (12.78):

$$DRL_x = \frac{1}{2}\left\|Lx\right\|^2 = \left\|x - Tx\right\|^2 \tag{12.98}$$

With, taking into account the symmetry of the matrix "T":

$$\frac{\partial DRL_x}{\partial x_l} = x_l^2 - 2\left(Tx\right)_l + \left[T\left(Tx\right)\right]_l \tag{12.99}$$

Algorithms.

The regularized algorithms are based on the expressions (12.87) and (12.90), replacing the unregularized divergence D by the composite divergences respectively DC_h and DC_x.
The expressions of the involved gradients are then written as follows:

$$\frac{\partial DC_h\left(y\|X^k h^k\right)}{\partial h_l^k} = \frac{\partial D\left(y\|X^k h^k\right)}{\partial h_l^k} + \gamma \frac{\partial DR_h\left(h^k\right)}{\partial h_l^k} \tag{12.100}$$

and

$$\frac{\partial DC_x\left(y\|H^{k+1} x^k\right)}{\partial x_l^k} = \frac{\partial D\left(y\|H^{k+1} x^k\right)}{\partial x_l^k} + \mu \frac{\partial DR_x\left(x^k\right)}{\partial x_l^k} \tag{12.101}$$

In these expressions, the regularization terms DR are either DRQ or DRL.
So the non-multiplicative regularized algorithms are written in the form:

$$h_l^{k+1} = h_l^k + \delta^k h_l^k \left\{ \left[-\frac{\partial DC_h}{\partial h_l^k} \right] - \sum_m h_m^k \left[-\frac{\partial DC_h}{\partial h_m^k} \right] \right\} \tag{12.102}$$

$$x_l^{k+1} = x_l^k + \beta^k x_l^k \left\{ \sum_i y_i \left[-\frac{\partial DC_x}{\partial x_l^k} \right] - \sum_m x_m^k \left[-\frac{\partial DC_x}{\partial x_m^k} \right] \right\} \tag{12.103}$$

*** Multiplicative forms.**

They are obtained by following the procedure indicated in the previous section; the components of the opposite of the gradients $\left[-\frac{\partial DC_h}{\partial h_l^k} \right]$ and $\left[-\frac{\partial DC_x}{\partial x_l^k} \right]$ are shifted in order to obtain components that are all positive, and we build the algorithms:

$$h_l^{k+1} = h_l^k + \delta^k h_l^k \left\{ \frac{\left[-\frac{\partial DC_h}{\partial h_l^k} \right]_d}{\sum_m h_m^k \left[-\frac{\partial DC_h}{\partial h_m^k} \right]_d} - 1 \right\} \tag{12.104}$$

$$x_l^{k+1} = x_l^k + \beta^k x_l^k \left\{ \frac{\sum_i y_i \left[-\frac{\partial DC_x}{\partial x_l^k} \right]_d}{\sum_m x_m^k \left[-\frac{\partial DC_x}{\partial x_m^k} \right]_d} - 1 \right\} \tag{12.105}$$

Purely multiplicative forms are obtained, as always, by taking the descent step size $\beta^k = \delta^k = 1 \ \forall k$, but there's no guarantee that such algorithms, over which we have no control, will converge.

Appendix 1 - Generalized Logarithm.

A reminder on the Generalized Logarithm function.

The function "Generalized Logarithm" or "Deformed Logarithm" (d-logarithm, Box-Cox transformation) is written as follows:

$$\log_d(x) = \frac{x^{1-d} - 1}{1-d} \tag{1}$$

The limit cases are the function $\log x$ when $d \to 1$ and the linear function $(x-1)$ when $d \to 0$.
The inverse function is the "Generalized exponential" or "Deformed exponential" function:

$$\exp_d(x) = [1 + (1-d)\,x]^{\frac{1}{1-d}} \tag{2}$$

This last function is obviously only defined if the quantity in square brackets is positive or zero.

Appendix 2 - Means.

Recalls on means.

For a set of real numbers $a_j \geq 0 \quad j = 1...N$ and a set of weights $w_j \geq 0$ we define:

A. Weighted arithmetic mean.

$$M_a = \frac{\sum_j w_j a_j}{\sum_{i=1}^N w_i} = \sum_j c_j a_j \quad ; \quad c_j = \frac{w_j}{\sum_{i=1}^N w_i} \quad ; \quad \sum_j c_j = 1 \tag{3}$$

In the particular case where all the weights are equal, we have $c_j = \frac{1}{N}$ and the unweighted arithmetic mean is written as follows:

$$M_a = \frac{\sum_j a_j}{N} \tag{4}$$

B. Weighted geometric mean.

$$M_g = \left[\prod_j a_j^{w_j}\right]^{\frac{1}{\sum_{i=1}^N w_i}} = \exp\left[\frac{1}{\sum_{i=1}^N w_i}\sum_j w_j \log a_j\right] \tag{5}$$

Which can also be written:

$$M_g = \prod_j a_j^{c_j} = \exp\left[\sum_j c_j \log a_j\right] \quad ; \quad c_j = \frac{w_j}{\sum_{i=1}^N w_i} \quad ; \quad \sum_j c_j = 1 \tag{6}$$

In the particular case where all the weights are equal, we have $c_j = \frac{1}{N}$ and the unweighted geometric mean is written as follows:

$$M_g = \prod_j a_j^{\frac{1}{N}} = \left[\prod_j a_j\right]^{\frac{1}{N}} \tag{7}$$

C. Weighted harmonic mean.

$$M_h = \frac{\sum_{i=1}^{N} w_i}{\sum_j \frac{w_j}{a_j}} = \frac{1}{\sum_j \frac{c_j}{a_j}} \quad ; \quad c_j = \frac{w_j}{\sum_{i=1}^{N} w_i} \quad ; \quad \sum_j c_j = 1 \tag{8}$$

In the particular case where all the weights are equal, we have $c_j = \frac{1}{N}$ and the unweighted harmonic mean is written as follows:

$$M_h = \frac{N}{\sum_j \frac{1}{a_j}} \tag{9}$$

D. Generalized mean with exponent "t" - Power mean - Hôlder mean.

For non-zero real "t", we define:

$$M_t = \left[\frac{1}{N} \sum_j a_j^t \right]^{\frac{1}{t}} \tag{10}$$

- $t = 1 \quad \Rightarrow$ Arithmetic mean
- $t = 2 \quad \Rightarrow$ Quadratic mean M_q (Square root mean)
- $t = -1 \quad \Rightarrow$ Harmonic mean
- $t \to 0 \quad \Rightarrow$ one tends towards the geometric mean (caution, it is a passage to the limit.
- $t \to +\infty \quad \Rightarrow$ one tends towards : $\max[a_j]$
- $t \to -\infty \quad \Rightarrow$ one tends towards : $\min[a_j]$

The weighted versions are obtained by the relation:

$$M_t = \left[\frac{1}{\sum_{i=1}^{N} w_i} \sum_j w_j a_j^t \right]^{\frac{1}{t}} \tag{11}$$

or also:

$$M_t = \left[\sum_j c_j a_j^t \right]^{\frac{1}{t}} \quad ; \quad c_j = \frac{w_j}{\sum_{i=1}^{N} w_i} \quad ; \quad \sum_j c_j = 1 \tag{12}$$

An interesting inequality: if $t1 < t2$ we have $M_{t1} \leq M_{t2}$.

This allows us to recover the inequalities between the different means:

$$\min \leq M_h \leq M_g \leq M_a \leq M_q \leq \max \tag{13}$$

This notion can be extended to the generalized mean in the meaning of a function "f".

E. Generalized "f" mean.

Let "f" be an injective and continuous function, the weighted average in the meaning of the function "f" is defined by:

$$M_f = f^{-1}\left[\frac{\sum_j w_j f(a_j)}{\sum_{i=1}^{N} w_i}\right] \tag{14}$$

or also:

$$M_f = f^{-1}\left[\sum_j c_j f(a_j)\right] \quad ; \quad c_j = \frac{w_j}{\sum_{i=1}^{N} w_i} \quad ; \quad \sum_j c_j = 1 \tag{15}$$

If all the weights w_j are equal, we obtain the unweighted version:

$$M_f = f^{-1}\left[\frac{\sum_j f(a_j)}{N}\right] \tag{16}$$

Remarks:
- the function "f" is assumed to be injective to ensure that "f^{-1}" exists.
- the function "f" is assumed to be continuous to ensure that the quantities $\left[\frac{\sum_j f(a_j)}{N}\right]$ or $\left[\frac{\sum_j f(a_j)}{\sum_{i=1}^{N} w_i}\right]$ belong to the domain of "f^{-1}".
- if "f" is injective and continuous, it is strictly monotonous, which results in:

$$\min[a_j] \leq M_f \leq \max[a_j] \tag{17}$$

The various classical means can be considered as special cases:
- if $f(x) = x$ we recover the arithmetic mean.
- if $f(x) = \log x$ we recover the geometric mean.
- if $f(x) = \frac{1}{x}$ we recover the harmonic mean.
- if $f(x) = x^t$ we recover the Hölder mean (generalized mean with exponent "t").

Application to our problem:
In our applications, we have 2 data fields "p" and "q" whose homologous

points are noted p_i and q_i; the means considered are made always between 2 homologous points, and are then extended to the whole field by simply summing the means between homologous points.
Therefore, in all relations concerning means, we always have N=2.
In the weighted means, we will use weights in the form of (α) and $(1-\alpha)$, $0 < \alpha < 1$.
And so we have:
* Weighted arithmetic mean:

$$M_a(p,q) = \sum_i [\alpha p_i + (1-\alpha) q_i] \tag{18}$$

* Weighted geometric mean:

$$M_g(p,q) = \sum_i p_i^{\alpha} q_i^{1-\alpha} \tag{19}$$

* Weighted harmonic mean:

$$M_h(p,q) = \sum_i \frac{p_i q_i}{(1-\alpha) p_i + \alpha q_i} \tag{20}$$

* Weighted quadratic mean:

$$M_q(p,q) = \sqrt{\sum_i \alpha p_i^2 + (1-\alpha) q_i^2} \tag{21}$$

Appendix 3 - Generalization of GHOSH.

Generalization of GHOSH et al. [38]

With the change of denomination of the parameters, and considering that it is necessary to introduce the multiplicative factor already mentioned, the divergence of Ghosh et al. is written as that of BHHJ, by replacing in this one: $1 \Leftrightarrow A$ and $\beta \Leftrightarrow B$, or equivalently, in our divergences: $\lambda \Leftrightarrow A+B$ and $\lambda - 1 \Leftrightarrow B$.
This leads to the generalized divergence:

$$GH(p\|q) = \frac{1}{A\,(A+B)}\left[\sum_i q_i^{A+B} + \frac{A}{B}\sum_i p_i^{A+B} - \frac{A+B}{B}\sum_i p_i^A q_i^B\right] \quad (22)$$

The gradient with respect to "q" will be written as follows:

$$\frac{\partial GH\,(p\|q)}{\partial q_i} = \frac{1}{A} q_i^{B-1}\left(q_i^A - p_i^A\right) \quad (23)$$

The dual divergence will be written:

$$GH(q\|p) = \frac{1}{A\,(A+B)}\left[\sum_i p_i^{A+B} + \frac{A}{B}\sum_i q_i^{A+B} - \frac{A+B}{B}\sum_i q_i^A p_i^B\right] \quad (24)$$

The gradient with respect to "q" will be expressed as:

$$\frac{\partial GH\,(q\|p)}{\partial q_i} = \frac{1}{B} q_i^{A-1}\left(q_i^B - p_i^B\right) \quad (25)$$

If, from the divergence of Ghosh et al. we want to obtain the invariant form of this divergence by scale change on "q", by applying the appropriate

procedure, we obtain the factor K:

$$K = \left(\frac{\sum_i p_i^A q_i^B}{\sum_i q_i^{A+B}} \right)^{\frac{1}{A}} \tag{26}$$

This leads after simplification to the invariant divergence:

$$GHI(p\|q) = \frac{1}{B} \left[\sum_i p_i^{A+B} - \left(\frac{\sum_i p_i^A q_i^B}{\sum_i q_i^{A+B}} \right)^{\frac{A+B}{A}} \sum_i q_i^{A+B} \right] \tag{27}$$

We can easily verify that if we replace "q" by "Kq", nothing changes, so we have well obtained the required result.
This expression is the difference of 2 positive terms (disregarding the constant multiplicative factor $\frac{1}{B}$), one can thus apply on each term of this difference the same increasing function, the "Generalized Logarithm" for example, then the extreme form, the "Log" which will lead to:

$$LGHI(p\|q) = \frac{1}{B} \log \left(\sum_i p_i^{A+B} \right) - \frac{A+B}{AB} \log \left(\sum_i p_i^A q_i^B \right) + \frac{1}{A} \log \left(\sum_i q_i^{A+B} \right) \tag{28}$$

We can see, which is true in all cases, that this divergence is not only invariant by a change of scale on "q", but also by a change of scale on "p".
At this point, we can re-examine the dual Ghosh's divergence which is written:

$$GH(q\|p) = \frac{1}{A\,(A+B)} \left[\sum_i p_i^{A+B} + \frac{A}{B} \sum_i q_i^{A+B} - \frac{A+B}{B} \sum_i q_i^A p_i^B \right] \tag{29}$$

It is rendered invariant by scale change with respect to "q" by introducing the invariance factor:

$$K = \left(\frac{\sum_i q_i^A p_i^B}{\sum_i q_i^{A+B}} \right)^{\frac{1}{B}} \tag{30}$$

We so obtain:

$$GHI(q\|p) = \frac{1}{A\,(A+B)} \left[\sum_i p_i^{A+B} - \left(\frac{\sum_i q_i^A p_i^B}{\sum_i q_i^{A+B}} \right)^{\frac{A+B}{B}} \sum_i q_i^{A+B} \right] \tag{31}$$

Hence, the Logarithmic form:

$$LGHI(q\|p) = \frac{1}{A\,(A+B)} \left[\log \sum_i p_i^{A+B} - \frac{A+B}{B} \log \sum_i q_i^A p_i^B + \frac{A}{B} \log \sum_i q_i^{A+B} \right] \tag{32}$$

Appendix 4 - Gradient of invariant divergences.

Computation of the gradient of an invariant divergence; influence of the invariance factor.

Considering a divergence of $D(p,q)$, which we make invariant with respect to "q" by introducing the invariance factor $K(p,q)$, we attempt here to answer the following question:
What happens to the expression of the gradient with respect to "q" depending on whether we use the *"nominal"* invariance factor K_0, or another acceptable expression of K, in this case $K_1 = \frac{\sum_i p_i}{\sum_i q_i}$?
The two cases to be distinguished are therefore:

* Either $K = K_0$, where K_0 is the "*nominal*" invariance factor , obtained by solving the equation:

$$\sum_i \frac{\partial D(p_i \| K q_i)}{\partial K} = 0 \tag{33}$$

Then, the gradient $\frac{\partial D(p\|Kq)}{\partial q_j}$ is given by (3.59).

$$\frac{\partial D(p\|K_0 q)}{\partial q_j} = K_0 \left[\frac{\partial D(p\|Kq)}{\partial (Kq_j)}\right]_{K=K_0} \tag{34}$$

* Or, $K = K_1$, in this case, the gradient $\frac{\partial D(p\|Kq)}{\partial q_j}$ is given by (3.60)

$$\frac{\partial D(p\|Kq)}{\partial q_j} = \frac{\partial K_1}{\partial q_j}\left[\sum_i \frac{\partial D(p_i\|Kq_i)}{\partial K}\right]_{K=K_1} + K_1\left[\frac{\partial D(p\|Kq)}{\partial (Kq_j)}\right]_{K=K_1} \tag{35}$$

Note that this expression can also be written as follows:

$$\frac{\partial D\left(p\|Kq\right)}{\partial q_j}=\frac{\partial K_1}{\partial q_j}\left[\sum_i q_i\frac{\partial D\left(p_i\|Kq_i\right)}{\partial\left(Kq_i\right)}\right]_{K=K_1}+K_1\left[\frac{\partial D\left(p\|Kq\right)}{\partial\left(Kq_j\right)}\right]_{K=K_1} \tag{36}$$

We observe that when $K\neq K_0$, the first term of the second member of this expression is no longer zero.

In both cases, we still have:

$$\sum_j q_j\frac{\partial D\left(p\|Kq\right)}{\partial q_j}=0 \tag{37}$$

If we consider the case of a linear model, that is to say $q_j=\left(Hx\right)_j$, then:

$$\frac{\partial q_j}{\partial x_l}=h_{jl}\quad\Rightarrow\quad\frac{\partial D\left(p\|Kq\right)}{\partial x_l}=\sum_j\frac{\partial D\left(p\|Kq\right)}{\partial q_j}\frac{\partial q_j}{\partial x_l} \tag{38}$$

and:

$$\frac{\partial D\left(p\|Kq\right)}{\partial x_l}=\sum_j h_{jl}\frac{\partial K}{\partial q_l}\left[\sum_i q_i\frac{\partial D\left(p_i\|Kq_i\right)}{\partial\left(Kq_i\right)}\right]+K\sum_j h_{jl}\left[\frac{\partial D\left(p\|Kq\right)}{\partial\left(Kq_l\right)}\right] \tag{39}$$

If $K=K_0$, the first term will be zero.

Applications to some divergences.
If we use:

$$K=K1=\frac{\sum_j p_j}{\sum_j q_j}\quad\Rightarrow\quad\frac{\partial K}{\partial q_l}=\frac{\partial K_1}{\partial q_l}=-\frac{\sum_j p_j}{\left(\sum_j q_j\right)^2} \tag{40}$$

In a first step, it comes:

$$\frac{\partial D\left(p\|K_1q\right)}{\partial x_l}=\frac{\sum_j p_j}{\sum_j q_j}\left\{-\frac{1}{\sum_j q_j}\left[\sum_i q_i\frac{\partial D\left(p\|K_1q\right)}{\partial\left(K_1q_i\right)}\right]\left[H^T1\right]_l+\left[H^T\frac{\partial D\left(p\|K_1q\right)}{\partial\left(K_1q\right)}\right]_l\right\} \tag{41}$$

We can now specify the divergence under consideration:

1 - Mean Square Deviation.
We have here:

$$D\left(p\|Kq\right)=MC\left(p\|Kq\right)=\frac{1}{2}\sum_i\left(p_i-Kq_i\right)^2 \tag{42}$$

Using the nominal invariance factor: $K = K_0 = \frac{\sum_i p_i q_i}{\sum_i q_i^2}$, we obtain:

$$\frac{\partial MC\left(p\|K_0 q\right)}{\partial q_j} = -K_0\left(p_j - K_0\, q_j\right) = -\frac{\sum_i p_i q_i}{\sum_i q_i^2}\left(p_j - \frac{\sum_i p_i q_i}{\sum_i q_i^2}\, q_j\right) \tag{43}$$

and it comes:

$$\frac{\partial MC\left(p\|K_0 q\right)}{\partial x_l} = -\frac{\sum_i p_i q_i}{\sum_i q_i^2}\left[H^T\left(p - \frac{\sum_i p_i q_i}{\sum_i q_i^2}\, q\right)\right]_l \tag{44}$$

With another invariance factor K_1, we have:

$$MC\left(p\|K_1 q\right) = \frac{1}{2}\sum_j \left(p_j - K_1 q_j\right)^2 \quad \Rightarrow \quad \frac{\partial MC\left(p\|K_1 q\right)}{\partial\left(K_1 q_i\right)} = -\left(p_i - K_1 q_i\right) \tag{45}$$

from where, in a first step, with $K_1 = \frac{\sum_j p_j}{\sum_j q_j}$, we have :

$$\frac{\partial MC\left(p\|K_1 q\right)}{\partial x_l} = \frac{\sum_j p_j}{\sum_j q_j}\left\{\frac{1}{\sum_j q_j}\left[\sum_i q_i\left(p_i - K_1 q_i\right)\right]\left[H^T 1\right]_l - \left[H^T\left(p - K_1 q\right)\right]_l\right\} \tag{46}$$

so, all simplifications made:

$$\frac{\partial MC\left(p\|K_1 q\right)}{\partial x_l} = \frac{\left(\sum_j p_j\right)^2}{\sum_j q_j}\left\{\left[\sum_i \bar{q}_i\left(\bar{p}_i - \bar{q}_i\right)\right]\left[H^T 1\right]_l - \left[H^T\left(\bar{p} - \bar{q}\right)\right]_l\right\} \tag{47}$$

2 - Kullback-Leibler divergence.

In this case, K_0 et K_1 are the same, so:

$$KL\left(p\|K_1 q\right) = \sum_i p_i \log\frac{p_i}{K_1 q_i} + K_1 q_i - p_i = \sum_i d\left(p_i\|K_1 q_i\right) \tag{48}$$

$$\frac{d\left(p_i\|K_1 q_i\right)}{\partial\left(K_1 q_i\right)} = -\frac{p_i}{K_1 q_i} + 1 \tag{49}$$

From the equation (41), we have:

$$\frac{\partial KL\left(p\|K_1 q\right)}{\partial x_l} = \frac{\sum_j p_j}{\sum_j q_j}\left\{-\frac{1}{\sum_j q_j}\left[\sum_i q_i\left(-\frac{p_i}{K_1 q_i} + 1\right)\right]\left[H^T 1\right]_l + \left[H^T\left(-\frac{p}{K_1 q} + 1\right)\right]_l\right\} \tag{50}$$

With $K_1 = \frac{\sum_j p_j}{\sum_j q_j} = K_0$, we have, all simplifications made:

$$\frac{\partial KL\left(p\|K_1 q\right)}{\partial x_l} = \frac{\sum_j p_j}{\sum_j q_j}\left\{ \underbrace{-\left[\sum_i \bar{q}_i\left(1-\frac{\bar{p}_i}{\bar{q}_i}\right)\right]}_{=0}\left[H^T 1\right]_l + \left[H^T\left(1-\frac{\bar{p}}{\bar{q}}\right)\right]_l \right\} \tag{51}$$

We can see that the first term of the expression between braces is zero; this is related to the fact that the invariance factor used here corresponds to the one that would be computed explicitly for this divergence, so it is in fact the *"nominal"* invariance factor K_0 for the Kullback-Leibler divergence and the gradient has a simplified expression (3.59) (34), then, finally:

$$\frac{\partial KL\left(p\|K_1 q\right)}{\partial x_l} = \frac{\sum_j p_j}{\sum_j q_j}\left\{\left[H^T\left(1-\frac{\bar{p}}{\bar{q}}\right)\right]_l\right\} \tag{52}$$

3 - Neyman Chi2.

We have:

$$\chi^2_N\left(p\|Kq\right) = \sum_i \frac{\left(p_i - Kq_i\right)^2}{Kq_i} = \sum_i d\left(p_i\|Kq_i\right) \tag{53}$$

$$\frac{\partial \chi^2_N\left(p\|Kq\right)}{\partial Kq_i} = 1 - \frac{p_i^2}{\left(Kq_i\right)^2} \tag{54}$$

From equation (41), it comes:

$$\frac{\partial \chi^2_N\left(p\|K_1 q\right)}{\partial x_l} = \frac{\sum_j p_j}{\sum_j q_j}\left\{\left[H^T\left(1-\frac{p^2}{\left(K_1 q\right)^2}\right)\right]_l - \frac{1}{\sum_j q_j}\left[\sum_i q_i\left(1-\frac{p_i^2}{\left(K_1 q_i\right)^2}\right)\right]\left[H^T 1\right]_l\right\} \tag{55}$$

With $K_1 = \frac{\sum_j p_j}{\sum_j q_j}$, we have, all simplifications made:

$$\frac{\partial \chi^2_N\left(p\|K_1 q\right)}{\partial x_l} = \frac{\sum_j p_j}{\sum_j q_j}\left\{\left[H^T\left(1-\frac{\bar{p^2}}{\bar{q^2}}\right)\right]_l - \left[\sum_i \bar{q}_i\left(1-\frac{\bar{p_i}^2}{\bar{q_i}^2}\right)\right]\left[H^T 1\right]_l\right\} \tag{56}$$

With $K = K_0 = \sqrt{\frac{\sum_j p_j^2 q_j^{-1}}{\sum_j q_j}}$, we have, from equation (54)

$$\frac{\partial \chi^2_N\left(p\|K_0 q\right)}{\partial K_0 q_i} = 1 - \left(\frac{p_i}{K_0 q_i}\right)^2 \tag{57}$$

Then, with (34):

$$\frac{\partial \chi^2_N\left(p \| K_0 q\right)}{\partial q_j} = K_0 \left[1 - \left(\frac{p_i}{K_0 q_i}\right)^2\right] \tag{58}$$

And finally:

$$\frac{\partial \chi^2_N\left(p \| K_0 q\right)}{\partial x_l} = K_0 \left[H^T.1\right]_l - \frac{1}{K_0}\left[H^T\left(\frac{p^2}{q^2}\right)\right]_l \tag{59}$$

Appendix 5 - Expressions of the invariance factor.

Here we give the expressions of the *"nominal"* invariance factor $K_0(p,q)$ for some particular divergences.

General form:

$$K_0(p,q) = \left(\frac{\sum_j p_j^\alpha q_j^\beta}{\sum_j p_j^\delta q_j^\gamma} \right)^\mu \quad ; \quad \alpha + \beta = \delta + \gamma \quad ; \quad \mu = \frac{1}{\gamma - \beta} \tag{60}$$

In a more synthetic form, one can write, taking into account the relations between the parameters:

$$K_0(p,q) = \left[\sum_i \left(\frac{p_i}{q_i} \right)^{\alpha-\delta} \frac{p_i^\delta q_i^\gamma}{\sum_j p_j^\delta q_j^\gamma} \right]^{\frac{1}{\alpha-\delta}} \tag{61}$$

It is, for quantities of the form (p_i/q_i), a weighted generalized mean of the order "t" with the exponent $t = \alpha - \delta$, and weighting factors w_i such that $\sum_i w_i = 1$, which are written as follows:

$$w_i = \frac{p_i^\delta q_i^\gamma}{\sum_j p_j^\delta q_j^\gamma} \tag{62}$$

In the case of the dual Kullback-Leibler divergence, the invariance factor appears as a generalized mean based on a "ψ" function of the form:

$$K_0 = \psi^{-1} \left[\sum_i w_i \psi \left(\frac{p_i}{q_i} \right) \right] \tag{63}$$

With $\psi(x) = \log x$.
The general form of the weighting factors (62) remains unchanged.

1 - Kullback-Leibler divergence:

$$K_0(p,q) = \frac{\sum_j p_j}{\sum_j q_j} \tag{64}$$

Here, we have: $\alpha = 1, \beta = 0, \gamma = 1, \delta = 0$,
then $\alpha - \delta = 1, \gamma - \delta = 1$.

$$w_i = \frac{q_i}{\sum_j q_j} \tag{65}$$

2 - Dual Kullback-Leibler divergence:

$$K_0 = \psi^{-1}\left[\sum_i w_i \psi\left(\frac{p_i}{q_i}\right)\right] \tag{66}$$

With $\psi(x) = \log x$.

$$w_i = \frac{q_i}{\sum_j q_j} \tag{67}$$

3 - Mean square deviation:

$$K_0(p,q) = \frac{\sum_j p_j q_j}{\sum_j q_j^2} \tag{68}$$

We have here: $\alpha = 1, \beta = 1, \gamma = 2, \delta = 0$,
so, $\alpha - \delta = 1, \gamma - \delta = 2$.

$$w_i = \frac{q_i^2}{\sum_j q_j^2} \tag{69}$$

4 - Neyman's Chi2:

$$K_0(p,q) = \sqrt{\frac{\sum_j p_j^2 q_j^{-1}}{\sum_j q_j}} \tag{70}$$

Here: $\alpha = 2, \beta = -1, \gamma = 1, \delta = 0$,
then $\alpha - \delta = 2, \gamma - \delta = 1$.

$$w_i = \frac{q_i}{\sum_j q_j} \tag{71}$$

5 - Pearson's Chi2:

$$K_0\left(p,q\right)=\frac{\sum_j q_j}{\sum_j p_j^{-1}q_j^2} \tag{72}$$

We have here: $\alpha=0,\beta=1,\gamma=2,\delta=-1$,
then $\alpha-\delta=1,\gamma-\delta=3$.

$$w_i=\frac{p_i^{-1}q_i^2}{\sum_j p_j^{-1}q_j^2} \tag{73}$$

6 - Arithmetic-Geometric mean

$$K_0\left(p,q\right)=\left(\frac{\sum_j \sqrt{p_j q_j}}{\sum_j q_j}\right)^2 \tag{74}$$

We have: $\alpha=1/2,\beta=1/2,\gamma=1,\delta=0$,
then $\alpha-\delta=1/2,\gamma-\delta=1$.

$$w_i=\frac{q_i^3}{\sum_j q_j^3} \tag{75}$$

7 - Alpha divergence:

$$K_0\left(p,q\right)=\left(\frac{\sum_j p_j^{\lambda_a}q_j^{1-\lambda_a}}{\sum_j q_j}\right)^{\frac{1}{\lambda_a}} \tag{76}$$

If $\lambda_a=1$ we recover K_0 corresponding to the K.L. divergence.
We have here: $\alpha=\lambda_a,\beta=1-\lambda_a,\gamma=1,\delta=0$,
so: $\alpha-\delta=\lambda_a,\gamma-\delta=1$.

$$w_i=\frac{q_i}{\sum_j q_j} \tag{77}$$

8 - Beta divergence:

$$K_0\left(p,q\right)=\frac{\sum_j p_j q_j^{\lambda_b-1}}{\sum_j q_j^{\lambda_b}} \tag{78}$$

If $\lambda_b = 1$ we recover K_0 corresponding to the K.L. divergence.
If $\lambda_b = 2$ we recover K_0 corresponding to the mean square deviation.
Here: $\alpha = 1, \beta = \lambda_b - 1, \gamma = \lambda_b, \delta = 0$,
then: $\alpha - \delta = 1 - \lambda_b, \gamma - \delta = \lambda_b$.

$$w_i = \frac{q_i^{\lambda_b}}{\sum_j q_j^{\lambda_b}} \tag{79}$$

Appendix 6 - Effects of the invariance factor.

Here we show the effects of using an invariance factorsuch as $K = \frac{\sum_j p_j}{\sum_j q_j}$ on some classical divergences.

These effects can be summarized by:

Rule:
Except for a multiplicative coefficient, which depends only on $\sum_j p_j$, the invariant divergence can be obtained by replacing in the initial divergence, the variables p_i and q_i by the normalized variables $\bar{p}_i = \frac{p_i}{\sum_j p_j}$ and $\bar{q}_i = \frac{q_i}{\sum_j q_j}$.

1 - Mean square deviation:
We have the base divergence:

$$EQM\left(p\|q\right) = \sum_i \left(p_i - q_i\right)^2 \tag{80}$$

With $K = \frac{\sum_j p_j}{\sum_j q_j}$, we have:

$$EQM\left(p\|Kq\right) = \sum_i \left(p_i - \frac{\sum_j p_j}{\sum_j q_j} q_i\right)^2 \tag{81}$$

then:

$$EQMI\left(p\|q\right) = \left(\sum_j p_j\right)^2 \sum_i \left(\bar{p}_i - \bar{q}_i\right)^2 \tag{82}$$

which can be reduced to an invariant divergence with respect to both arguments:

$$EQMI\left(p\|q\right)=\sum_{i}\left(\bar{p}_{i}-\bar{q}_{i}\right)^{2} \tag{83}$$

2 - Kullback-Leibler divergence:
The initial divergence is written:

$$KL\left(p\|q\right)=\sum_{i}p_{i}\log\frac{p_{i}}{q_{i}}+q_{i}-p_{i} \tag{84}$$

With $K=\frac{\sum_{j}p_{j}}{\sum_{j}q_{j}}$, it comes:

$$KL\left(p\|Kq\right)=\sum_{i}p_{i}\log\frac{p_{i}}{\frac{\sum_{j}p_{j}}{\sum_{j}q_{j}}q_{i}}+\frac{\sum_{j}p_{j}}{\sum_{j}q_{j}}q_{i}-p_{i} \tag{85}$$

That is:

$$KL\left(p\|Kq\right)=KLI\left(p\|q\right)=\sum_{j}p_{j}\left[\sum_{i}\bar{p}_{i}\log\frac{\bar{p}_{i}}{\bar{q}_{i}}+\bar{q}_{i}-\bar{p}_{i}\right] \tag{86}$$

which can be reduced to an invariant divergence with respect to both arguments:

$$KLI\left(p\|q\right)=\left[\sum_{i}\bar{p}_{i}\log\frac{\bar{p}_{i}}{\bar{q}_{i}}+\bar{q}_{i}-\bar{p}_{i}\right]=\left[\sum_{i}\bar{p}_{i}\log\frac{\bar{p}_{i}}{\bar{q}_{i}}\right] \tag{87}$$

It is a Kullback-Leibler between normalized variables.

3 - Neyman's Chi2:
The initial divergence is written:

$$\chi_{N}^{2}\left(p\|q\right)=\sum_{i}\frac{\left(p_{i}-q_{i}\right)^{2}}{q_{i}} \tag{88}$$

With $K=\frac{\sum_{j}p_{j}}{\sum_{j}q_{j}}$, we have:

$$\chi_{N}^{2}\left(p\|Kq\right)=\sum_{i}\frac{\left(p_{i}-\frac{\sum_{j}p_{j}}{\sum_{j}q_{j}}q_{i}\right)^{2}}{\frac{\sum_{j}p_{j}}{\sum_{j}q_{j}}q_{i}} \tag{89}$$

That is:

$$\chi_N^2 I\left(p\|q\right) = \sum_j p_j \sum_i \frac{\left(\bar{p}_i - \bar{q}_i\right)^2}{\bar{q}_i} \tag{90}$$

which can be reduced to an invariant divergence with respect to both arguments:

$$\chi_N^2 I\left(p\|q\right) = \sum_i \frac{\left(\bar{p}_i - \bar{q}_i\right)^2}{\bar{q}_i} \tag{91}$$

4 - Pearson's Chi2:
The initial divergence is written:

$$\chi_P^2\left(p\|q\right) = \sum_i \frac{\left(q_i - p_i\right)^2}{p_i} \tag{92}$$

With $K = \frac{\sum_j p_j}{\sum_j q_j}$, we obtain:

$$\chi_P^2 I\left(p\|q\right) = \sum_i \frac{\left(\frac{\sum_j p_j}{\sum_j q_j} q_i - p_i\right)^2}{p_i} \tag{93}$$

That is:

$$\chi_P^2 I\left(p\|q\right) = \sum_j p_j \sum_i \frac{\left(\bar{q}_i - \bar{p}_i\right)^2}{\bar{p}_i} \tag{94}$$

which can be reduced to an invariant divergence with respect to both arguments:

$$\chi_P^2 I\left(p\|q\right) = \sum_i \frac{\left(\bar{q}_i - \bar{p}_i\right)^2}{\bar{p}_i} \tag{95}$$

Appendix 7 - Inequalities between invariant divergences.

We show here, for some classic divergences denoted $D(p,q)$, that if $K_0(p,q)$ is the *"nominal"* invariance factor, and if $K_1(p,q) \neq K_0(p,q)$ is another possible invariance factor, we have:

$$D(p\|K_0 q) \leq D(p\|K_1 q) \tag{96}$$

This is an obvious consequence of the definition of the *"nominal"* invariance factor.
On the other hand, we have the relationship:

$$D(p\|K_1 q) - D(p\|K_0 q) = D(K_0\|K_1) \tag{97}$$

This is the basic divergence taken between the scalar quantities K_0 and K_1.

1 - Mean square deviation:
We have here:

$$MC(p\|K_0 q) = \sum_i (p_i - K_0 q_i)^2, \quad MC(p\|K_1 q) = \sum_i (p_i - K_1 q_i)^2 \tag{98}$$

We can write:

$$MC_{K_1} - MC_{K_0} = (K_1 - K_0)\left[(K_0 + K_1)\left(\sum_i q_i^2\right) - 2\sum_i p_i q_i\right] \tag{99}$$

For this divergence the nominal invariance factor K_0 is computed analytically and is written as follows:

$$K_0 = \frac{\sum_j p_j q_j}{\sum_j q_j^2} \tag{100}$$

This immediately leads to:

$$MC_{K_1} - MC_{K_0} = \sum_i q_i^2 \left(K_1 - K_0\right)^2 \geq 0 \tag{101}$$

Consequently:

$$MC\left(p \| K_0 q\right) \leq MC\left(p \| K_1 q\right) \tag{102}$$

With an equality (at 0 moreover) for $p_i = q_i \ \forall i$, with then: $K_0 = K_1 = 1$.
We also note, moreover, that:

$$MC\left(p \| K_1 q\right) - MC\left(p \| K_0 q\right) \approx MC\left(K_0 \| K_1\right) \tag{103}$$

2 - Kullback-Leibler divergence:
We have here:

$$KL\left(p \| K_0 q\right) = \sum_i \left(p_i \log \frac{p_i}{K_0 q_i} + K_0 q_i - p_i\right) \tag{104}$$

$$KL\left(p \| K_1 q\right) = \sum_i \left(p_i \log \frac{p_i}{K_1 q_i} + K_1 q_i - p_i\right) \tag{105}$$

We can write:

$$KL_{K_1} - KL_{K_0} = \sum_i p_i \left[\log \frac{K_0}{K_1} + \left(K_1 - K_0\right) \frac{\sum_i q_i}{\sum_i p_i}\right] \tag{106}$$

For this divergence the nominal invariance factor K_0 is derived analytically and is written as follows:

$$K_0 = \frac{\sum_j p_j}{\sum_j q_j} \tag{107}$$

This allows to write:

$$KL_{K_1} - KL_{K_0} = \sum_i q_i \left[K_0 \log \frac{K_0}{K_1} + K_1 - K_0\right] \tag{108}$$

So $[KL_{KL_1} - KL_{K_0}]$ is, with the exception of the multiplicative factor $\sum_i q_i$, equal to the Kullback-Leibler divergence between the scalars factors K_0 et K_1 denoted $KL\left(K_0 \| K_1\right) \geq 0$.

Consequently:

$$KL_{K_0} \leq KL_{K_1} \tag{109}$$

With equality (to zero) if $p_i = q_i \ \forall i$, i.e. equivalently $K_0 = K_1 = 1$.

3 - Pearson's Chi2:
In this case, we have:

$$\chi^2_P\left(p\|K_0 q\right) = \sum_i \frac{\left(p_i - K_0 q_i\right)^2}{p_i}, \quad \chi^2_P\left(p\|K_1 q\right) = \sum_i \frac{\left(p_i - K_1 q_i\right)^2}{p_i} \tag{110}$$

We can write:

$$\chi^2_{P,K_1} - \chi^2_{P,K_0} = \left(K_1 - K_0\right) \sum_i \frac{\left(K_0 + K_1\right) q_i^2 - 2 p_i q_i}{p_i} \tag{111}$$

For this divergence the nominal invariance factor K_0 is derived analytically and is written as follows:

$$K_0 = \frac{\sum_j q_j}{\sum_j q_j^2 p_j^{-1}} \tag{112}$$

This leads to:

$$\chi^2_{P,K_1} - \chi^2_{P,K_0} = \sum_i q_i \frac{\left(K_0 - K_1\right)^2}{K_0} \tag{113}$$

This expression is obviously positive, and therefore:

$$\chi^2_{P,K_0} \leq \chi^2_{P,K_1} \tag{114}$$

With equality (to zero) if $p_i = q_i \ \forall i$, then $K_0 = K_1 = 1$.
It can be seen that with the exception of a multiplicative factor, we have:

$$\chi^2_{P,K_1} - \chi^2_{P,K_0} \approx \chi^2_{K_0,K_1} \tag{115}$$

4 - Neyman's Chi2:
In this case, we have:

$$\chi^2_N\left(p\|K_0 q\right) = \sum_i \frac{\left(p_i - K_0 q_i\right)^2}{K_0 q_i}, \quad \chi^2_N\left(p\|K_1 q\right) = \sum_i \frac{\left(p_i - K_1 q_i\right)^2}{K_1 q_i} \tag{116}$$

We can write:

$$\chi^2_{N,K_1} - \chi^2_{N,K_0} = \frac{\left(K_0 - K_1\right)}{K_0 K_1} \left(\sum_i p_i^2 q_i^{-1} - K_0 K_1 \sum_i q_i \right) \tag{117}$$

For this divergence the nominal invariance factor K_0 is derived analytically and is written as follows:

$$K_0 = \left(\frac{\sum_j p_j^2 q_j^{-1}}{\sum_j q_j} \right)^{\frac{1}{2}} \tag{118}$$

We obtain immediately:

$$\chi^2_{N,K_1} - \chi^2_{N,K_0} = \sum_j q_j \frac{(K_0 - K_1)}{K_0 K_1} \left(K_0^2 - K_0 K_1 \right) = \sum_j q_j \frac{(K_0 - K_1)^2}{K_1} \tag{119}$$

This expression is obviously positive, and therefore:

$$\chi^2_{N,K_0} \leq \chi^2_{N,K_1} \tag{120}$$

With equality (to zero) if $p_i = q_i \ \forall i$, then $K_0 = K_1 = 1$.
As in the previous examples, we have:

$$\chi^2_N \left(p \| K_1 q \right) - \chi^2_N \left(p \| K_0 q \right) \approx \chi^2_N \left(K_0 \| K_1 \right) \tag{121}$$

Appendix 8 - Order relations between divergences.

We show on a few examples that for a divergence $D(p\|q)$ rendered invariant by using its nominal factor K_0, that is $D(p\|K_0 q)$, we have:

$$D(p\|K_0 q) \leq D(p\|q) \tag{122}$$

with equality to "0" if $p_i = q_i \quad \forall i$.
This is a consequence of the definition of the nominal invariance factor, because $D(p\|q)$ corresponds to $K = 1$.

1 - Mean square deviation.
Here, we have:

$$MC(p\|q) = \sum_i (p_i - q_i)^2, \quad MC(p\|K_0 q) = \sum_i (p_i - K_0 q_i)^2 \tag{123}$$

We can write:

$$MC - MC_{K_0} = (1 - K_0)\left[(1 + K_0)\left(\sum_i q_i^2\right) - 2\sum_i p_i q_i\right] \tag{124}$$

With:

$$K_0 = \frac{\sum_i p_i q_i}{\sum_i q_i^2} \tag{125}$$

it comes:

$$MC - MC_{K_0} = \sum_i q_i^2 (1 - K_0)^2 \geq 0 \tag{126}$$

2 - Kullback-Leibler divergence.
We have:

$$KL(p\|q) = \sum_i \left(p_i \log \frac{p_i}{q_i} + q_i - p_i\right), \quad KL(p\|K_0 q) = \sum_i \left(p_i \log \frac{p_i}{K_0 q_i} + K_0 q_i - p_i\right) \tag{127}$$

We can write:

$$KL - KL_{K_0} = \sum_i q_i \left[\frac{\sum_i p_i}{\sum_i q_i} \log K_0 + (1 - K_0) \right] \tag{128}$$

For this divergence the nominal invariance factor K_0 is derived analytically and is written as follows:

$$K_0 = \frac{\sum_j p_j}{\sum_j q_j} > 0 \tag{129}$$

Then:

$$KL - KL_{K_0} = \sum_i q_i \left(K_0 \log K_0 + 1 - K_0 \right) \geq 0 \tag{130}$$

3 - Pearson's Chi2.
In this case, we have:

$$\chi^2_P \left(p \| q \right) = \sum_i \frac{\left(p_i - q_i \right)^2}{p_i}, \quad \chi^2_P \left(p \| K_0 q \right) = \sum_i \frac{\left(p_i - K_0 q_i \right)^2}{p_i} \tag{131}$$

We can write:

$$\chi^2_P - \chi^2_{P,K_0} = (1 - K_0) \sum_i \frac{(1 + K_0)\, q_i^2 - 2 p_i q_i}{p_i} \tag{132}$$

For this divergence the nominal invariance factor K_0 is derived analytically and is written as follows:

$$K_0 = \frac{\sum_j q_j}{\sum_j q_j^2 p_j^{-1}} > 0 \tag{133}$$

Then:

$$\chi^2_P - \chi^2_{P,K_0} = \left(\sum_i q_i^2 p_i^{-1} \right) (1 - K_0) \left[(1 + K_0) - 2K_0 \right] \tag{134}$$

Or, also:

$$\chi^2_P - \chi^2_{P,K_0} = \left(\sum_i q_i^2 p_i^{-1} \right) (1 - K_0)^2 \tag{135}$$

And:

$$\chi^2_P - \chi^2_{P,K_0} = \sum_i q_i \frac{1}{K_0} (1 - K_0)^2 \geq 0 \tag{136}$$

4 - Neyman's Chi2.
We have, in this case:

$$\chi_N^2\left(p\|q\right)=\sum_i\frac{\left(p_i-q_i\right)^2}{q_i},\quad \chi_N^2\left(p\|K_0q\right)=\sum_i\frac{\left(p_i-K_0q_i\right)^2}{K_0q_i} \tag{137}$$

We can write:

$$\chi_N^2-\chi_{N,K_0}^2=\left(1-\frac{1}{K_0}\right)\sum_i p_i^2q_i^{-1}+\left(1-K_0\right)\sum_i q_i \tag{138}$$

Then:

$$\chi_N^2-\chi_{N,K_0}^2=\sum_i q_i\left[\frac{\sum_i p_i^2q_i^{-1}}{\sum_i q_i}\left(1-\frac{1}{K_0}\right)+\left(1-K_0\right)\right] \tag{139}$$

For this divergence the nominal invariance factor K_0 is derived analytically and is written as follows:

$$K_0=\left(\frac{\sum_j p_j^2q_j^{-1}}{\sum_j q_j}\right)^{\frac{1}{2}}>0 \tag{140}$$

Consequently, we have:

$$\chi_N^2-\chi_{N,K_0}^2=\sum_i q_i\left[K_0^2\left(1-\frac{1}{K_0}\right)+\left(1-K_0\right)\right]=\sum_i q_i\left(1-K_0\right)^2\geq 0 \tag{141}$$

4 - Alpha Divergence.
Using the expression Alpha divergence (5.15), it comes:

$$AC-AC_{K_0}=\frac{1}{\lambda_a\left(\lambda_a-1\right)}\left[\sum_i p_i^{\lambda_a}q_i^{1-\lambda_a}\left(1-K_0^{1-\lambda_a}\right)-\left(1-K_0\right)\right] \tag{142}$$

For this divergence the nominal invariance factor K_0 is derived analytically and is written as follows:

$$K_0=\left(\frac{\sum_j p_j^{\lambda_a}q_j^{1-\lambda_a}}{\sum_j q_j}\right)^{\frac{1}{\lambda_a}}>0 \tag{143}$$

Finally, we have:

$$AC-AC_{K_0}=\frac{\sum_i q_i}{\lambda_a\left(\lambda_a-1\right)}\left[K_0^{\lambda_a}-\lambda_aK_0-1+\lambda_a\right] \tag{144}$$

That expression is always positive if $K_0 \geq 0$.
The specific case of $\lambda_a \to 0$ leads to:

$$K_0 - \log K_0 - 1 \geq 0 \tag{145}$$

In a similar way, $\lambda_a \to 1$ leads to:

$$K_0 \log K_0 - K_0 + 1 \geq 0 \tag{146}$$

4 - Beta Divergence.
Using the expression of the Beta divergence (5.20), the derivation of the nominal invariance factor leads to:

$$K_0 = \frac{\sum_i p_i q_i^{\lambda_b - 1}}{\sum_i q_i^{\lambda_b}} \tag{147}$$

and we obtain after a few calculations:

$$BC - BC_{K_0} = \frac{\sum_i q_i^{\lambda_b}}{\lambda_b \left(\lambda_b - 1\right)} \left[K_0^{\lambda_b} - \lambda_b K_0 - 1 + \lambda_b\right] \geq 0 \tag{148}$$

This expression is analogous to that obtained for the Alpha Divergence and the particular cases can be derived from it immediately.

Appendix 9 - Regularization problems for the NMF.

In this appendix, we indicate the difficulties that arise when constructing purely multiplicative regularized algorithms for NMF, as they have been proposed in the literature [25] [79] [15].
Non-regularized multiplicative algorithms as they are proposed in the literature dedicated to NMF [25] [67] are immediately deduced from the SGM method described in Chapter 10.
All these algorithms only take into account the non-negativity constraint.
The method proposed in the literature can be summarized as follows:
For a divergence $D\left(Y\|HX\right)$ representing the data attachment term, the algorithmic method always consists, as previously indicated, of alternately performing a H descent step, followed by a X descent step.
Considering that we have obtained H^{k+1} from H^k, X^k, the basic algorithm can be written, with respect to the variable X for example, with $X^0 > 0$, in the form:

$$X_{nm}^{k+1} = X_{nm}^k + \alpha^k X_{nm}^k \left[-\frac{\partial D}{\partial X^k}\right]_{nm} \tag{149}$$

With $U_{nm}^k > 0$ and $V_{nm}^k > 0$, it is always possible to write:

$$\left[-\frac{\partial D}{\partial X^k}\right]_{nm} = U_{nm}^k - V_{nm}^k \tag{150}$$

This leads to:

$$X_{nm}^{k+1} = X_{nm}^k + \alpha^k X_{nm}^k \left[U_{nm}^k - V_{nm}^k\right] \tag{151}$$

From this, we deduce an algorithm which is at the base of purely multiplicative algorithms:

$$X_{nm}^{k+1} = X_{nm}^k + \alpha^k X_{nm}^k \left[\frac{U_{nm}^k}{V_{nm}^k} - 1\right] \tag{152}$$

The initial direction of displacement $X^k \left[-\frac{\partial D}{\partial X^k}\right]$ has of course been modified, and becomes $\frac{X^k}{V_k}\left[-\frac{\partial D}{\partial X^k}\right]$, but remains a descent direction because V^k is positive. After which, by choosing a descent step size $\alpha^k = 1\ \forall k$, we obtain the purely multiplicative form:

$$X_{nm}^{k+1} = X_{nm}^k \left[\frac{U_{nm}^k}{V_{nm}^k}\right] \tag{153}$$

In the case where the term of attachment to the data is given by the mean square (RMS) deviation:

$$D\left(Y \| HX\right) = \sum_{ij} \left[Y_{ij} - \left(HX\right)_{ij}\right]^2 \tag{154}$$

The multiplicative algorithm is written from (153), with $H \equiv H^{k+1}$:

$$X_{nm}^{k+1} = X_{nm}^k \left[\frac{\left(H^T Y\right)}{\left(H^T H X^k\right)}\right]_{nm} \tag{155}$$

This algorithm was proposed by Lee and Seung [67]. It is the direct transposition of the ISRA algorithm [29], well known in the field of deconvolution. Similarly, for a data attachment term, which is the Kullback-Leibler divergence, we obtain a purely multiplicative algorithm, which is the immediate transposition of the Richardson-Lucy (RL) algorithm [83] [69] well known in the field of deconvolution, which is written, with $H \equiv H^{k+1}$:

$$X_{nm}^{k+1} = \frac{X_{nm}^k}{\sum_l H_{nl}^T} \left[H^T \frac{Y}{HX^k}\right]_{nm} \tag{156}$$

As in all descent algorithms, the convergence is ensured by a computation of the descent step size, so that with a fixed descent step size for all iterations, and equal to 1, nothing is guaranteed for convergence, even if non-negativity is ensured subject to the choice of a positive initial iterate.
In the case of deconvolution, the convergence of such algorithms has been demonstrated for the two divergences mentioned above.
In the case of NMF, multiplicative algorithms have been proposed for many of the divergences listed in the book by Cichocki et al. [25].

*** Problems related to regularization.**

When we consider the regularized algorithms, things get a little worse. For example, again with the X variable, it is now a matter of minimizing under a non-negativity constraint, a composite divergence:

$$J\left(Y,HX\right)=D\left(Y\|HX\right)+\mu\ DR\left(X\right) \tag{157}$$

where $DR\left(X\right)$ is the penalty term, and "μ" the positive regularization factor. Basically, knowing X^k and H^{k+1}, the algorithm for the variable X is written as follows:

$$X_{nm}^{k+1}=X_{nm}^{k}+\alpha^k X_{nm}^{k}\left[-\frac{\partial D}{\partial X^k}+\mu\left(-\frac{\partial DR}{\partial X^k}\right)\right]_{nm} \tag{158}$$

The method generally proposed in the literature consists of decomposing $-\frac{\partial D}{\partial X^k}$ in a difference of 2 positive terms U^k and V^k: $-\frac{\partial D}{\partial X^k}=U^k-V^k$ as previously proposed, but on the other hand such a decomposition is not applied to the regularization term $-\frac{\partial DR}{\partial X^k}$, which means to implicitly consider that this term is always positive.

This is false. It is only true for certain penalty functions (such as $DR\left(X\right)=\|X\|^2$ for example).

From there, one thing leads to another; the proposed algorithm is first written:

$$X_{nm}^{k+1}=X_{nm}^{k}+\alpha^k X_{nm}^{k}\left[U_{nm}^{k}-V_{nm}^{k}+\mu\left(-\frac{\partial DR}{\partial X^k}\right)\right] \tag{159}$$

Then:

$$X_{nm}^{k+1}=X_{nm}^{k}+\alpha^k X_{nm}^{k}\left[\frac{U_{nm}^{k}}{V_{nm}^{k}+\mu\left(\frac{\partial DR}{\partial X^k}\right)_{nm}}-1\right] \tag{160}$$

And finally, by taking a descent step size equal to 1, we obtain a multiplicative algorithm which is written as follows:

$$X_{nm}^{k+1}=X_{nm}^{k}+\alpha^k X_{nm}^{k}\left[\frac{U_{nm}^{k}}{V_{nm}^{k}+\mu\left(\frac{\partial DR}{\partial X^k}\right)_{nm}}\right] \tag{161}$$

It is absolutely clear (in principle) that all this is only valid if $\left[V_{nm}^{k}+\mu\left(\frac{\partial DR}{\partial X^k}\right)_{nm}\right]$ is strictly positive; otherwise we have transformed a descent algorithm into an ascent algorithm.....

Authors who apply this type of algorithm prudently point out that if the regularization factor "μ" is small, everything is fine; it is just as clear that everything is fine regardless of "μ" if the penalty term is $DR(X) = \|X\|^2$, because in this case the gradient of the penalty term is positive.
It should be mentioned that this method is the transposition to the NMF of a rather old algorithm designated in the literature as One Step Late (OSL) by P.J.Green [39], in which the author clearly mentions the limits of use of his algorithm.

In fact, in order for this to work in all cases, it is necessary to proceed as described in Chapter 10 regarding the SGM method.

$*$ **The opposite of the gradient of the penalty term must be decomposed into the form:**

$$\left(-\frac{\partial DR}{\partial X^k}\right)_{nm} = UR^k_{nm} - VR^k_{nm} \tag{162}$$

With $UR^k_{nm} \geq 0$ and $VR^k_{nm} \geq 0$.
Then, we first write the algorithm in the form:

$$X^{k+1}_{nm} = X^k_{nm} + \alpha^k X^k_{nm}\left[U^k_{nm} - V^k_{nm} + \mu\, UR^k_{nm} - \mu\, VR^k_{nm} + \epsilon - \epsilon\right] \tag{163}$$

We have: $U^k_{nm} + \mu\, UR^k_{nm} + \epsilon > 0$ and $V^k_{nm} + \mu\, VR^k_{nm} + \epsilon > 0$.

From this, we deduce the algorithm that will be the basis of the purely multiplicative algorithms:

$$X^{k+1}_{nm} = X^k_{nm} + \alpha^k X^k_{nm}\left[\frac{U^k_{nm} + \mu\, UR^k_{nm} + \epsilon}{V^k_{nm} + \mu\, VR^k_{nm} + \epsilon} - 1\right] \tag{164}$$

Which does not cause any problems with regard to the descent properties since $V^k_{nm} + \mu\, VR^k_{nm} + \epsilon > 0$.
After that, a purely multiplicative algorithm is obtained by taking a descent step size equal to 1 for all the iterations:

$$X^{k+1}_{nm} = X^k_{nm}\left[\frac{U^k_{nm} + \mu\, UR^k_{nm} + \epsilon}{V^k_{nm} + \mu\, VR^k_{nm} + \epsilon}\right] \tag{165}$$

Obviously, the same reasoning applies for iterations on H, leading to H^{k+1} knowing H^k and X^k.
Nevertheless, the convergence of this type of algorithm is not demonstrated in all generality because of the particular choice of the descent step size.

Bibliography

[1] S.M. Ali and S.D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 131–142, 1966.

[2] S.I. Amari. Alpha divergence is unique, belonging to both f-divergence and Bregman divergence classes. *Information Theory, IEEE Transactions on*, 55(11):4925–4931, 2009.

[3] H.C. Andrews and B. R. Hunt. Digital image restoration. *Prentice-Hall Signal Processing Series, Englewood Cliffs: Prentice-Hall, 1977*, 1977.

[4] S. Arimoto. Information-theoretical considerations on estimation problems. *Information and Control*, 19(3):181–194, 1971.

[5] L. Armijo. Minimization of functions having Lipschitz continuous first partial derivatives. *Pacific Journal of mathematics*, 16(1):1–3, 1966.

[6] C. Arndt. *Information measures; information and its description in science and engineering*. Springer, 2001.

[7] G.R. Ayers and J.C. Dainty. Iterative blind deconvolution method and its applications. *Optics letters*, 13(7):547–549, 1988.

[8] M. Basseville. Distance measures for signal processing and pattern recognition. *Signal processing*, 18(4):349–369, 1989.

[9] M. Basseville. Divergence measures for statistical data processing - an annotated bibliography. *Signal Processing*, 93(4):621–633, 2013.

[10] M. Basseville. Information: entropies, divergences et moyennes. Technical report, IRISA 1996.

[11] A. Basu, I.R. Harris, N.L. Hjort, and M.C. Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559, 1998.

[12] A. Ben-Tal, A. Charnes, and M. Teboulle. Entropic means. *Journal of Mathematical Analysis and Applications*, 139(2):537–551, 1989.

[13] R. Beran. Minimum Hellinger distance estimates for parametric models. *The Annals of Statistics*, pages 445–463, 1977.

[14] J-F. Bercher. On some entropy functionals derived from Rényi information divergence. *Information Sciences*, 178(12):2489–2506, 2008.

[15] M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca, and R. J. Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. *Computational statistics & data analysis*, 52(1):155–173, 2007.

[16] M. Bertero and P. Boccacci. *Introduction to inverse problems in imaging*. CRC press, 1998.

[17] D.P. Bertsekas. *Non linear programming*. Athena scientific, Belmont, Mass., 1995.

[18] P.K. Bhatia and S. Singh. On a new csiszär´s f-divergence measure. *Cybernetics and information technologies*, 13(2):43–57, 2013.

[19] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[20] L.M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217, 1967.

[21] J. Burbea and C.R. Rao. On the convexity of higher order Jensen differences based on entropy functions (corresp.). *Information Theory, IEEE Transactions on*, 28(6):961–963, 1982.

[22] Y. Censor and S.A. Zenios. *Parallel optimization: Theory, algorithms, and applications*. Oxford University Press on Demand, 1997.

[23] A. Cichocki and S.I. Amari. Families of α-β-and γ-divergences: Flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568, 2010.

[24] A. Cichocki, S. Cruces, and S.I. Amari. Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization. *Entropy*, 13(1):134–170, 2011.

[25] A. Cichocki, R. Zdunek, A.H. Phan, and S.I. Amari. *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009.

[26] I. Csiszär. Information measures: A critical survey. In *Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions, Random Processes*, pages 73–86, 1974.

[27] I. Csiszär et al. Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318, 1967.

[28] J.C. Culioli. *Introduction à l'optimisation*. Ed.Ellipses, 2012.

[29] M.E. Daube-Witherspoon and G. Muehllehner. An iterative image space reconstruction algorthm suitable for volume ECT. *Medical Imaging, IEEE Transactions on*, 5(2):61–66, 1986.

[30] A.R. De Pierro. On the convergence of the iterative image space reconstruction algorithm for volume ECT. *IEEE TRANS. MED. IMAG.*, 6(2):174–175, 1987.

[31] G. Demoment. Déconvolution des signaux. Technical report, Cours Supelec 1985.

[32] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[33] S.S. Dragomir, V. Gluscevic, and C.E.M. Pearce. Approximations for the Csiszär's f divergence via mid point inequalities. *Inequality Theory and Applications*, 1:139–154, 2001.

[34] S. Eguchi and S. Kato. Entropy and divergence associated with power function and the statistical application. *Entropy*, 12(2):262–274, 2010.

[35] C. Févotte and J. Idier. Algorithms for nonnegative matrix factorization with the β-divergence. *Neural computation*, 23(9):2421–2456, 2011.

[36] H. Fujisawa and S. Eguchi. Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, 99(9):2053–2081, 2008.

[37] S. Furuichi and F.C. Mitroi. Mathematical inequalities for some divergences. *Physica A: Statistical Mechanics and its Applications*, 391(1):388–400, 2012.

[38] A. Ghosh, I.R. Harris, A. Maji, A. Basu, and L. Pardo. A generalized divergence for statistical inference. Technical report, BIRU/2013/3, Bayesian and Interdisciplinary Research Unit, Indian Statistical Institute, Kolkata, India, 2013.

[39] P. J. Green. Bayesian reconstructions from emission tomography data using a modified em algorithm. *IEEE transactions on medical imaging*, 9(1):84–93, 1990.

[40] M. Grendár and R.K. Niven. The Pólya information divergence. *Information Sciences*, 180(21):4189–4194, 2010.

[41] G.H. Hardy, J.E. Littlewood, and G. Pólya. *Inequalities*. Cambridge university press, 1952.

[42] J. Havrda and F. Charvát. Quantification method of classification processes. concept of structural α-entropy. *Kybernetika*, 3(1):30–35, 1967.

[43] E. Hellinger. Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die reine und angewandte Mathematik*, 136:210–271, 1909.

[44] N. Henze and M.D. Penrose. On the multivariate runs test. *Annals of statistics*, pages 290–298, 1999.

[45] F.B. Hildebrand. *Introduction to numerical analysis*. Courier Corporation, 1987.

[46] J-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of convex analysis*. Springer Science & Business Media, 2012.

[47] P.O. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of machine learning research*, 5(Nov):1457–1469, 2004.

[48] J. Idier, editor. *Approche bayésienne pour les problèmes inverses*. Traité IC2, Série traitement du signal et de l'image, Hermès, Paris, nov. 2001.

[49] F. Itakura and S. Saito. Analysis synthesis telephony based on the maximum likelihood method. In *Proceedings of the 6th International Congress on Acoustics*, volume 17, pages C17–C20. pp. C17–C20, 1968.

[50] P. A. Jansson. *Deconvolution. With applications in spectroscopy.* New York: Academic Press, 1984, edited by Jansson, Peter A., 1984.

[51] P. A. Jansson. *Deconvolution of images and spectra.* Courier Corporation, 2014.

[52] H. Jeffreys. An invariant form for the prior probability in estimation problems. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 186, pages 453–461. The Royal Society, 1946.

[53] J.L. Jensen. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica*, 30(1):175–193, 1906.

[54] J.N. Kapur. Generalized entropy of order α and type β. In *The Math. Seminar*, volume 4, pages 78–82, 1967.

[55] J.N. Kapur. Nonadditive measures of entropy and distributions of statistical mechanics. *Indian J. Pure Appl. Math*, 14:1372–1387, 1983.

[56] L. Knockaert. A class of statistical and spectral distance measures based on Bose-Einstein statistics. *IEEE Trans. on Signal Processing*, 11:3171–3174, 1993.

[57] S. Kullback and R.A. Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[58] K. Lange, R. Carson, et al. EM reconstruction algorithms for emission and transmission tomography. *J Comput Assist Tomogr*, 8(2):306–16, 1984.

[59] H. Lanteri, C. Aime, H. Beaumont, and P. Gaucherel. Blind deconvolution using the richardson-lucy algorithm. In *Optics in Atmospheric Propagation and Random Phenomena*, volume 2312, pages 182–192. International Society for Optics and Photonics, 1994.

[60] H. Lanteri, M. Barilli, H. Beaumont, C. Aime, P. Gaucherel, and H. Touma. Comparison of several algorithms for blind deconvolution: analysis of noise effects. In *Optics in Atmospheric Propagation and Adaptive Systems*, volume 2580, pages 275–287. International Society for Optics and Photonics, 1995.

[61] H. Lanteri, M. Roche, and C. Aime. Penalized maximum likelihood image restoration with positivity constraints: multiplicative algorithms. *Inverse problems*, 18(5):1397, 2002.

[62] H. Lantéri, M. Roche, O. Cuevas, and C. Aime. A general method to devise maximum-likelihood signal restoration multiplicative algorithms with non-negativity constraints. *Signal Processing*, 81(5):945–974, 2001.

[63] H. Lantéri and C. Theys. La Gamma divergence de Fujisawa et Eguchi. Une alternative pour imposer la contrainte de somme. Application au démélange linéaire. In *Actes du 24e Colloque GRETSI sur le Traitement du Signal et des Images*, 2013.

[64] H. Lantéri, C. Theys, and C. Aime. Divergences invariantes par changement d'échelle pour la reconstruction de signaux ou d'images . Application á l'interférométrie optique. In *Actes du 25e Colloque GRETSI sur le Traitement du Signal et des Images*, 2015.

[65] H. Lantéri, C. Theys, and C. Aime. Scale invariant divergences for signal and image reconstruction. In *Signal Processing Conference (EUSIPCO), 2015 23rd European*, pages 2162–2166. IEEE, 2015.

[66] H. Lantéri, C. Theys, and C. Richard. Minimisation d'une forme générale de divergence sous contrainte de non-négativité application à la factorisation en matrices non-négatives. In *Actes du 23e Colloque GRETSI sur le Traitement du Signal et des Images*, 2011.

[67] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[68] J. Lin. Divergence measures based on the Shannon entropy. *Information Theory, IEEE Transactions on*, 37(1):145–151, 1991.

[69] L.B. Lucy. An iterative technique for the rectification of observed distributions. *The astronomical journal*, 79:745, 1974.

[70] O.J.J. Michel, R.G. Baraniuk, and P. Flandrin. Time-frequency based distance and divergence measures. In *Time-Frequency and Time-Scale Analysis, 1994., Proceedings of the IEEE-SP International Symposium on*, pages 64–67. IEEE, 1994.

[71] M. Mihoko and S. Eguchi. Robust blind source separation by β divergence. *Neural computation*, 14(8):1859–1886, 2002.

[72] H. Neemuchwala, A. Hero, S. Zabuawala, and P. Carson. Image registration methods in high-dimensional space. *International Journal of Imaging Systems and Technology*, 16(5):130–145, 2007.

[73] J. Neyman. Contribution to the theory of the x2 test. In *Proceedings of the Berkeley symposium on mathematical statistics and probability*, volume 1, pages 239–273. University of California Press, 1949.

[74] F. Österreicher. On a class of perimeter-type distances of probability distributions. *Kybernetika*, 32(4):389–393, 1996.

[75] F. Österreicher. Csiszär´s f-divergences-basic properties. *RGMIA Res. Rep. Coll*, 2002.

[76] F. Österreicher. Distances based on the perimeter of the risk set of a testing problem. *Austrian Journal of Statistics*, 42(1):3–19, 2013.

[77] J.A. Pardo and M.C. Pardo. Small-sample comparisons for the Rukhin goodness-of-fit-statistics. *Statistical Papers*, 40(2):159–174, 1999.

[78] L. Pardo. *Statistical inference based on divergence measures*. CRC Press, 2005.

[79] V. P. Pauca, J. Piper, and R. J. Plemmons. Nonnegative matrix factorization for spectral data analysis. *Linear algebra and its applications*, 416(1):29–47, 2006.

[80] A.D. Polyanin, V.F. Zaitsev, and A. Moussiaux. *Handbook of First-Order Partial Differential Equations*. CRC Press, November 2001.

[81] A. Rényi. On a new axiomatic theory of probability. *Acta Mathematica Hungarica*, 6(3-4):285–335, 1955.

[82] A. Rényi. Fourth berkeley symp. *Math. Stat. Probability*, 1:547, 1961.

[83] W.H. Richardson. Bayesian-based iterative method of image restoration. *JOSA*, 62(1):55–59, 1972.

[84] R.T. Rockafellar. *Convex analysis*. Princeton university press, 2015.

[85] F. Roddier. *Distributions et transformations de Fourier à l'usage des physiciens et des ingénieurs*. Ediscience, 1971.

[86] A.L. Rukhin. Optimal estimator for the mixture parameter by the method of moments and information affinity. In *Transactions of the 12th Prague Conference on Information Theory*, pages 214–219, 1994.

[87] M. Salicrú. Measures of information associated with Csiszär's divergences. *Kybernetika*, 30(5):563–573, 1994.

[88] C. E. Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948.

[89] B.D. Sharma and D.P. Mittal. New nonadditive measures of entropy for discrete probability distributions. *J. Math. Sci*, 10:28–40, 1975.

[90] L.A. Shepp and Y. Vardi. Maximum likelihood reconstruction for emission tomography. *Medical Imaging, IEEE Transactions on*, 1(2):113–122, 1982.

[91] I.J. Taneja. Generalized information measures and their applications. on-line book, 2001. *URL www. mtm. ufsc. br/taneja/book/book. html.*

[92] I.J. Taneja. On generalized information measures and their applications. *Advances in Electronics and Electron Physics*, 76:327–413, 1989.

[93] A.N. Tikhonov and V.Y. Arsenin. *Methods of solving incorrect problems*. Nauka, Moscow, 1974.

[94] D.M. Titterington. General structure of regularization procedures in image reconstruction. *Astronomy and Astrophysics*, 144:381, 1985.

[95] C. Tsallis. Possible generalization of Boltzmann-Gibbs statistics. *Journal of statistical physics*, 52(1-2):479–487, 1988.

[96] M. Uchida and H. Shioya. An extended formula for divergence measures using invariance. *Electronics and Communications in Japan (Part III: Fundamental Electronic Science)*, 88(4):35–42, 2005.

[97] T.S. Zaccheo and R.A. Gonsalves. Iterative maximum-likelihood estimators for positively constrained objects. *JOSA A*, 13(2):236–242, 1996.

[98] J. Zhang. Divergence function, duality, and convex analysis. *Neural Computation*, 16(1):159–195, 2004.