Generalized Maximum Entropy, Convexity and Machine Learning

Timothy D. Sears

A thesis submitted for the degree of
Doctor of Philosophy at
The Australian National University

May 2008

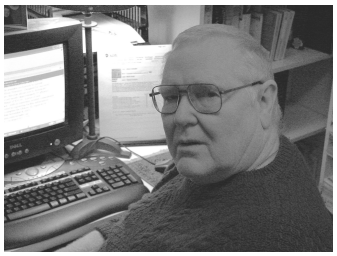
Typeset in Computer Modern by TEX and LATEX 2ε.

Except where otherwise indicated, this thesis is my own original work.

Publications completed during candidacy: Lloyd and Sears (2005); Sears (2007); Sears and Sunehag (2007); Sears and Vishwanathan (2007); Takeuchi et al. (2006). Chapters 3 to 5 are based in part on Sears and Vishwanathan (2007) and Chapter 7 is based on Sears and Sunehag (2007).

Timothy D. Sears
12 May 2008

In Memoriam



John Edward Sears
1935-2006

To Beth, who made it possible.

Acknowledgements

It's the end of a long road and I am happy to be able to recognize and thank the many people who have helped me along the way. First I wish to thank my family for their love and support. They make it hard to have a tough day. Thanks Ian, T.J., Corinne and Beth.

Sometimes people, even Aussies, have asked me: Why Australia? or Why ANU? They would do better to ask themselves: Why not Australia? Fantastic country, beautiful climate, and the friendliest people on the planet (except when I cut them off in traffic). I hope nobody else finds out about it. But the real inspiration for coming to ANU here in Canberra, came from reading two books by ANU researchers "Logic and Learning" and "Learning with Kernels". The author of the first, John Lloyd, and a co-author of the second, Alex Smola, served on my supervisory panel, provided valuable advice and support. Thanks Alex for your endless flow of ideas and energy, and thanks John for your encouragement, sound judgment, and advice. Thanks to both for according me the honor of collaborating on papers along the way. A special thanks to the third member of my panel S.V.N. Vishwanathan, known to all simply as Vishy, who has been my chief collaborator during my research here. Always upbeat, always ready to engage on topics from beekeeping to Banach spaces, Vishy has been invaluable to me. Thanks Vishy. Thanks also go to Peter Sunehag who has patiently read virtually all of this thesis and coauthored a version of Chapter 7, which appeared elsewhere. I appreciate your collaboration, Peter. Special thanks to Simon Günter, who proofread a number of chapters. I take credit for the remaining mistakes.

I have also benefited from many discussions with researchers, students and visitors at ANU's College of Engineering and Computer Science and the Canberra NICTA lab. Their enthusiasm, ideas, and support have helped my work in countless ways. And so to Doug Aberdeen, Yasemin Altun, Marconi Barbosa, Justin Bedo, Olivier Buffet, Wray Buntine, Tiberio Caetano, Stéphane Canu, Jon Cohen,Li Cheng, Evan Greensmith, Charles Gretton, Omri Guttman, Markus Hegland, Knut Hüper, Marcus Hutter, Risi Kondor, Adam Kowalczyk, Quoc Li, Shahar Mendelson, Kee Siong Ng, Cheng Soon Ong, Robert Orsi,Petra Philips, Mark Reed, Scott Sanner, Nic Schraudolph, Le Song, Choon Hui Teo, Jochen Trumpf, Manfred Warmuth, Chris Webers, Bob Williamson, Jin Yu, Xinhua Zhang, I say thank you all.

Special thanks go to Kishor Gawande who provided programming support

via Elefant and great, timely programming advice for the experiments. On the RSISE IT staff, thank you to James Ashton and the rest of the IT team, not forgetting the late Joe Elso. On the RSISE administrative staff thank you to Michelle Moravec, Deb Pioch and Di Kossatz for keeping me on track and on time insofar as that is possible!

Thanks go to my examiners for taking the trouble to read this work, providing detailed comments, and finding mistakes.

I gratefully acknowledge financial sponsorship from NICTA. National ICT Australia is funded by the Australian Government's Department of Communications, Information Technology and the Arts and the Australian Research Council through Backing Australia's Ability and the ICT Center of Excellence program.

Timothy D. Sears
Canberra, Australia
May, 2007

Abstract

This thesis identifies and extends techniques that can be linked to the principle of maximum entropy (maxent) and applied to parameter estimation in machine learning and statistics. Entropy functions based on deformed logarithms are used to construct Bregman divergences, and together these represent a generalization of relative entropy. The framework is analyzed using convex analysis to characterize generalized forms of exponential family distributions. Various connections to the existing machine learning literature are discussed and the techniques are applied to the problem of non-negative matrix factorization (NMF).

Keywords: Maximum entropy, Bregman divergence, exponential family, deformed logarithm, escort distribution, non-negative matrix factorization.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter gives an overview of some of the major ideas developed in the thesis in condensed form. The technical aspects are downplayed as far as possible.

1.1 Maxent and Machine Learning

This thesis is mainly concerned with identifying and extending techniques that can be linked to the principle of maximum entropy (maxent) and applied to parameter estimation in machine learning and statistics. Why maxent? A major focus of machine learning is to extract information from data and use it to automatically perform some task. Only performance on the task really matters, not necessarily an understanding of the task. Nevertheless to design algorithms it seems reasonable to model the 'true' state of the world in order to achieve a successful outcome. Inevitably, models can only include some, not all, potentially relevant information. Models always leave us in doubt about the exact state of the world. Therefore, to a machine learning researcher, the following quotation from Jaynes (1957a) makes maxent sound like it might be a useful design principle:

> The principles and mathematical methods of statistical mechanics are seen to be of much more general applicability... In the problem of prediction, the maximization of entropy is not an application of a law of physics, but merely a method of reasoning which ensures that no unconscious arbitrary assumptions have been introduced.

Jaynes was recommending the use of maxent as an induction principle for statistics and other fields beyond statistical physics. This occurred in 1957, a time which predates the field of machine learning. Since then, maxent has been applied in a number of disciplines, including finance, econometrics, astrophysics, imaging, and philosophy (Besnerais et al., 1999; Borwein et al., 2003; Stutzer, 2000; Golan et al., 1996; Donoho et al., 1992; Williamson, 2005). Today, directly acknowledged inspiration from maxent occurs in just a few areas of machine learning. The use of the maxent principle in Natural Language Processing (NLP) (e.g.

Pietra et al., 1995; Berger et al., 1996) is fairly straightforward to trace. Environmental modeling (Phillips et al., 2006, 2004) uses similar techniques. If one looks elsewhere in the machine learning literature it would be hard to conclude that the maxent principle is widely applied. However, if one takes a broader view, it turns out that a generalized form of maxent can often be found in the field.

Taking a broader view is facilitated by expressing and analyzing models in the terminology of convex analysis. Some of the work here is in the vein of establishing this connection. In the end, however, I do not claim some special status for maxent. Putting induction principles into practice makes them start to look rather similar. In fact, the field of machine learning has accommodated many induction principles with a variety of origins, all of which have had their successes. Instead, the terminology and mathematics of maxent make it a useful point of departure for developing other models, some of which can be seen as generalizations of the original maxent principle. This is also demonstrated in this work.

1.1.1 Classic to General

First a brief introduction to entropy. For the purposes of this thesis entropy is scalar valued function of a vector. In particular, the function

$$S(\boldsymbol{p}) := \sum_{i=1}^{N} p_i \log(p_i)$$

is often called Boltzmann-Gibbs entropy in physics, Shannon entropy in information theory or just plain entropy. In the the cross-disciplinary spirit of the thesis we will call it Shannon-Boltzmann-Gibbs (SBG) entropy. It is useful to have a specific name for this function, since a major focus of the thesis is on generalizations of entropy .

Let us now introduce the classic maxent problem, as a prelude to generalizing it. The classic maxent problem is a finite-dimensional optimization problem. Its solution is a vector $\boldsymbol{p}$ that solves

$$\underset{\boldsymbol{p}}{\text{minimize}} \quad S(\boldsymbol{p}) \;\; \text{subject to} \;\; \boldsymbol{A}\boldsymbol{p} = \boldsymbol{b}, \;\; \text{and } p_i \geq 0,$$

where $\boldsymbol{A}$ is an $M \times N$ matrix, with $M < N$, and the constraint matrix is usually assumed to have full (row) rank and take the form $\boldsymbol{A} = \begin{pmatrix} \boldsymbol{B} \\ \mathbf{1}^{\mathrm{T}} \end{pmatrix}$, *i.e.,* it includes a normalization constraint. In this thesis, we will maintain a focus on discrete models throughout, since they are sufficient to highlight the main points.

Using notation from convex analysis it is possible to write this problem down more compactly. More importantly, convex analysis allows us to reason about

optimization problems involving non-differentiable functions and hard constraints in a manner similar to the familiar techniques for smooth functions, as long as certain technicalities are addressed along the way. This approach is explained in detail in mathematical references such as Hiriart-Urruty and Lemarechal (2001); Rockafellar and Wets (1998).

The problem can be re-expressed as

$$\underset{\boldsymbol{p}}{\text{minimize}} \quad S(\boldsymbol{p}) + \delta_{\{\mathbf{0}\}}(\boldsymbol{A}\boldsymbol{p} - \boldsymbol{b}). \tag{1.1}$$

Explicit reference to the non-negativity constraint can be dropped by assuming S takes on the value $+\infty$ outside its effective domain. Here also δ_C is the convex indicator function of the set C, which takes on the value 0 on the set and $+\infty$ outside the set. In this case the set in question is simply the singleton set $\{0\}$, since exact matching is required.[1]

This formulation of the problem lends itself to an elegant form of duality known as Fenchel duality. Fenchel duality applies when analyzing the inf-convolution of two convex functions as in $\inf_x (f(x) + g(x))$. The Fenchel dual is given by $\sup_{x*} (f^*(x^*) + g^*(-x^*))$, where f^* and g^* are Fenchel conjugates, defined as $f^*(x^*) := \sup_x \langle x, x^* \rangle - f(x)$. The reader can see that the dual problem is constructed by separately conjugating the components of the primal. This modularity is a hallmark of Fenchel duality. A slightly more complicated version of Fenchel duality is needed to handle the constraint matrix $\boldsymbol{A}$ (e.g. Borwein and Lewis, 2000, Theorem 3.3.5). It can be shown that the Fenchel dual for (1.1) turns out to be:

$$\underset{\boldsymbol{\mu}}{\text{minimize}} \quad S^*(A^{\mathrm{T}}\boldsymbol{\mu}) + \langle \boldsymbol{\mu}, \boldsymbol{b} \rangle \,. \tag{1.2}$$

Here $\langle \cdot, \cdot \rangle$ represents the inner product in $\mathbb{R}^M$, and S^* is the Fenchel conjugate of S. This latter function is given by: $S^*(\boldsymbol{p}^*) = \sum_{n=1}^{N} \exp(p_m^* - 1)$. Also, a given solution to this problem, $\bar{\boldsymbol{\mu}}$, can be used to recover the primal solution $\bar{\boldsymbol{p}}$ via

$$\bar{\boldsymbol{p}} = \nabla S^*(A^{\mathrm{T}}\boldsymbol{\mu}).$$

Letting $\boldsymbol{p}^* = A^{\mathrm{T}}\boldsymbol{\mu}$, and applying some algebra allows an entry of $\bar{\boldsymbol{p}}$ to be expressed as

$$\bar{p}_n \propto \exp(p_n^*).$$

[1] Often, the indicator of this singleton set would be written as $\delta_{\mathbf{0}}$.

or in the notation preferred in this thesis

$$\bar{\boldsymbol{p}} \propto \exp[A^{\mathrm{T}}\bar{\boldsymbol{\mu}}]. \tag{1.3}$$

In this manner, the machine learning folklore that maxent leads to exponential families can be confirmed. Although this explanation is not mathematically difficult, machine learners often acquire this bit of folklore in a different way. In section 1.2 we will briefly digress on a more typical explanation of exponential families. But first, we will generalize the setup in (1.1) a bit.

The first generalization is to relax the exact matching constraints to instead match within an epsilon-ball around zero, defined by some P-norm. In the notation of 1.1, $\delta_{\{\mathbf{0}\}}$ becomes $\delta_{\epsilon\mathcal{B}_P}$ Next, the notion of entropy can be generalized. It is well known that Kullback-Leibler (KL) divergence, or relative entropy, is a natural generalization of SBG entropy (*e.g.* Cover and Thomas (1991)). KL-divergence is in turn a special case of Bregman divergence. Bregman divergence between two distributions is defined as $\Delta F(\boldsymbol{p}, \boldsymbol{p}_0) := F(\boldsymbol{p}) - F(\boldsymbol{p}_0) - \langle \nabla F(\boldsymbol{p}_0), \boldsymbol{p} - \boldsymbol{p}_0 \rangle$, using a convex function F defined on the non-negative orthant.[2] Bregman divergences appear in a number of machine learning algorithms. (See (Azoury and Warmuth, 2001, e.g.) for further details, as well as Chapter 2.) Briefly, the resulting function acts something like a square-distance measure. The generalized problem is now:

$$\underset{\boldsymbol{p}}{\text{minimize}} \quad \Delta F(\boldsymbol{p}, \boldsymbol{p}_0) + \delta_{\epsilon B_p}(\boldsymbol{A}\boldsymbol{p} - \boldsymbol{b}). \tag{1.4}$$

In this setting the classical objective function can be recovered by specifying that $F = S$ and $\boldsymbol{p}_0 = \mathbf{1}/N$, the uniform distribution over the state space. Of course this is only one of many routes for generalization, some of the others will be mentioned later on. A key property of the original problem that has been preserved is that of convexity. Also, the specific form of the problem is still well-matched to Fenchel duality.

The Fenchel dual for (1.4) turns out to be:

$$\underset{\boldsymbol{\mu}}{\text{minimize}} \quad F^*(A^{\mathrm{T}}\boldsymbol{\mu} + \boldsymbol{p}_0^*) + \langle \boldsymbol{\mu}, \boldsymbol{b} \rangle + \epsilon \, ||\boldsymbol{\mu}||_Q \,, \tag{1.5}$$

where the P-norm and Q-norm are dual to each other ($1/P + 1/Q = 1$), and $\boldsymbol{p}_0^* = \nabla F(\boldsymbol{p}_0)$. Because of the shape of $\boldsymbol{A}$, the dual problem is typically much easier to solve than the primal problem in practice.

The resulting formulation (1.5) can be seen as a more general form of the maximum a-posteriori (MAP) problem for exponential families and can be shown to be equivalent to versions involving the log-partition function in the classic ver-

[2] More properly, with effective domain equal to the non-negative orthant.

sion of the problem. The key difference is that the objective does not involve a partition function and instead contains an explicit parameter for the normalization. After solving (in practice, almost always numerically) for the optimal dual solution, $\bar{\boldsymbol{\mu}}$, the primal solution, $\bar{\boldsymbol{p}}$, can be recovered via:

$$\bar{\boldsymbol{p}} = \underbrace{\nabla F^*}_{\text{'Family'}}(\overbrace{A^{\mathrm{T}}\bar{\boldsymbol{\mu}}}^{\text{'Score'}} + \boldsymbol{p}_0^*) \tag{1.6}$$

The gradient of the conjugate of the entropy function (F) determines the family (∇F^*) of the distribution. In the classic case, the negative entropy function, S, yields the familiar scalar function $\exp(\cdot)$ as the gradient map, hence an exponential family distribution. The freedom introduced in (1.4) can be used to employ a variety of entropy functions, including those based on so-called deformed logarithms (Naudts, 2004b). These entropy functions yield the corresponding deformed exponential family as distributions and will be taken up in Chapter 4.

1.1.2 Non-SBG entropy examples

An important example of generalized entropy is based on the q-logarithm. This introduces a parameter, q, usually restricted to the interval $(0, 2)$. The definition of the q-logarithm, deformed exponential $\exp_q$, and corresponding negative entropy S_q are:[3]

$$\begin{aligned}
\log_q(p) &= \frac{x^{1-q} - 1}{1-q} \\
\exp_q(p) &= (1 + (1-q)p)_+^{\frac{1}{1-q}} \\
S_q(\boldsymbol{p}) &= -\sum_{i=1}^{N} p_i \log_q(\frac{1}{p_i}),
\end{aligned}$$

where $(\cdot)_+ = \max(\cdot, 0)$. Taking limits in the case $q = 1$, the standard definitions of all three functions are recovered. When used in the generalized maxent problem (1.4), the solution has the form of a q-exponential family.

A graphical comparison of the $\exp_q$ function (Figure 1.1) suggests how q-exponentials can yield distributions with power-law behavior. For values of $q < 1$, $\exp_q$ can reach the value zero, in contrast to $q = 1$, while values of $q > 1$ cause the function to approach zero slowly, leading to 'fat-tailed' distributions.

[3]The entropy function described is closely related to Tsallis entropy, except that it cleanly delivers the $\exp_q$ family as the result, whereas Tsallis entropy does so after a reparameterization. More precisely, a term of the negative entropy function used here is $-p\log_q(1/p)$ while Tsallis entropy is defined with $p\log_q(p)$as the negative entropy term.

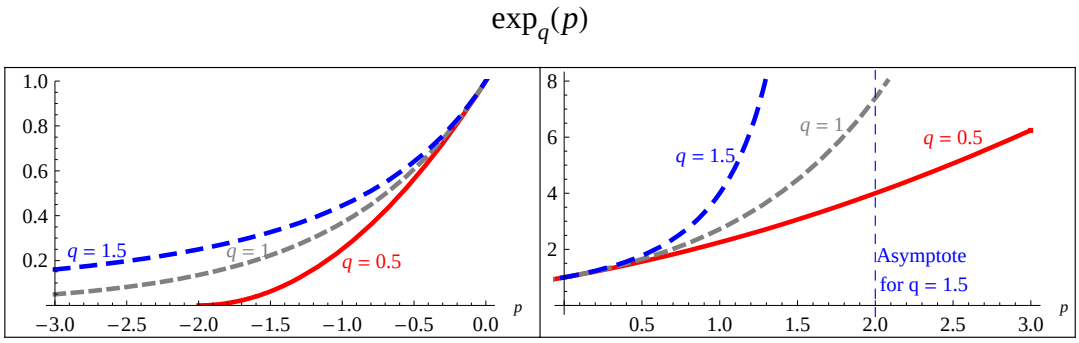


Figure 1.1: Left-hand graph indicates tail behavior. $q > 1 \rightarrow$ fat tail. $q < 1 \rightarrow$ truncated tail.

The role that distributions based on the $\exp_q$ functions might play in applications is perhaps better suggested by a couple of examples.

Example: The loaded die problem. The loaded die problem Jaynes (1982a) is familiar to students of maxent and a useful example to indicate what role q might play in a model. To recapitulate, in that problem the matrix and data vector are

$$\boldsymbol{A} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \qquad \boldsymbol{b} = \begin{pmatrix} 4.5 \\ 1 \end{pmatrix}.$$

The key feature of the example is that the expectation of the die is set much higher than the fair value of 3.5. Together with normalization, this forms the constraint matrix, clearly ruling out the uniform distribution associated with a fair die. Using the problem formulation of (1.4), set $\epsilon = 0$ (exact constraint matching), and vary q. Each value of q produces a different value for the probability of a particular face, while matching the input expectations. The solutions are illustrated in Figure 1.2. Lower values of q depress the probabilities of the extreme events (rolling a 1 or a 6), while higher values of q accentuate them. The classic solution is in the middle of the pack for each face of the die. Figure 1.2 also illustrates that the sensitivity of the probability to the parameter q varies. From a machine learning perspective it is hard to say much more about this model, since no task has been defined. However, the value of the estimate for face 2 appears to be more stable than the others. In a finance setting one might say the task of making a two-way market on whether a 2 appears next looks to be somewhat easier than doing so for the other faces.

In the loaded die example the focus was on exact constraint matching and exploring the freedom to choose an entropy function. In the next example, constraint relaxations are also used.

Example: The Dantzig Selector. Recently, Candes and Tao (2007) reported

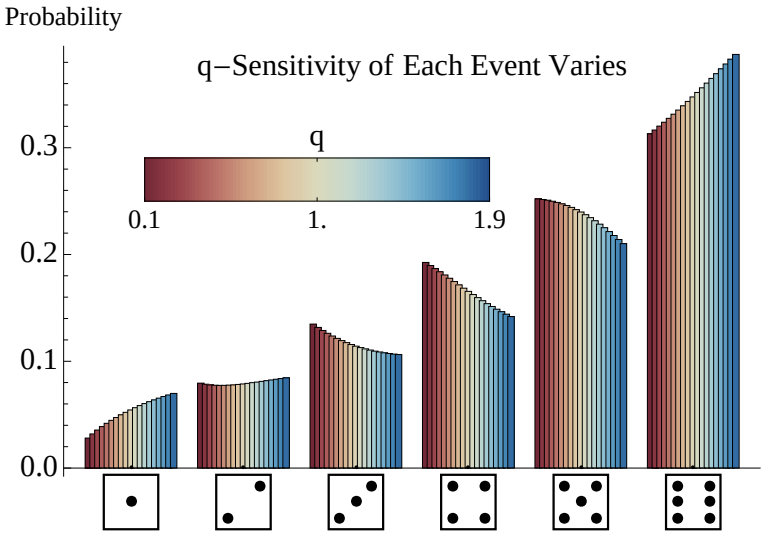


Figure 1.2: The loaded die problem. Mean input value: 4.5 instead of the 'fair' value of 3.5

that for a regression model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}$, the solution to the problem

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \;\; ||\boldsymbol{\beta}||_1 + \delta_{\epsilon B_\infty}(\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}))$$

exactly finds the correct set of regressors, with high probability, under special conditions of low noise and a sparse 'true' model $\widehat{\boldsymbol{\beta}}$. The use of the 1-norm as a regularizer is well known to produce sparse models, and this is one case where a sparse model is known to be optimal. With the help of a partitioned matrix we can encode the regression coefficients in a non-negative vector using the relation:

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{I} & | -\boldsymbol{I} \end{pmatrix} \boldsymbol{p}.$$

This simple change makes it easy to The regularizer in the Dantzig selector can be approached by S_q in the limit as $q \to 0$, and therefore fits the generalized maxent format of (1.4).

The entropy function used here reflects a very specific kind of prior knowledge, suited to the particular task of discovering the active set of regressors. Both of these examples suggest that other entropy functions might provide an interesting set of tools precisely because they lead to non-exponential families. However in machine learning a large stock of models and techniques are built upon the unique properties of exponential families.

1.2 Exponential Families and Machine Learning

In the maxent literature, the equivalence between classic maxent and maximum likelihood is shown in a number of places (e.g. Jaynes, 1982a). The presentation leading to (1.3) and given in more detail in Chapter 3, was originally inspired by Borwein and Limber (1996), which uses conjugate functions and Fenchel duality to calculate a dual problem. However, in most presentations of exponential family models, the dual problem is established via Lagrangian duality and involves the log partition function, or normalizing function. In the machine learning literature, an exponential family distribution is often assumed at the start, and the connection via duality to SBG entropy is often omitted. It is therefore worthwhile to walk through a more conventional presentation to underline that in certain special cases we are really doing nothing different.

Let X be a vector-valued (M-dimensional) random variable that can take on one of N values $x_1 \dots x_N$ from some index set $\mathcal{I}$. Suppose also that X has a log-linear, or exponential family distribution. This means the density (likelihood) of a particular value for X is given by

$$Pr(X = x_n \mid \boldsymbol{\theta}) = p(x_n) = \exp(\, \langle \boldsymbol{\theta}, \boldsymbol{\phi}(x_n) \rangle \,)/Z(\boldsymbol{\theta}) \text{ where} \tag{1.7}$$

$$Z(\boldsymbol{\theta}) = \sum_{i=1}^{N} \exp(\, \langle \boldsymbol{\theta}, \boldsymbol{\phi}(x_n) \rangle \,),$$

where Z is a normalizing function and $\boldsymbol{\phi} : \mathcal{I} \to \mathbb{R}^M$ is a vector-valued function representing a statistical measurement.

The likelihood (1.7) can also be written as

$$\exp(\, \langle \boldsymbol{\theta}, \boldsymbol{\phi}(x_n) \rangle - T(\boldsymbol{\theta}) \,), \tag{1.8}$$

with the function $T = \log Z$. In the exponential family it does not matter whether normalization is inside or outside of the exp function. (This small fact becomes more important when other entropy functions are introduced.)

To consider more than one observation, say K of them, let x_{n_k} denote the case where observation k is in the n-th state. A joint setting of the K random variables is then a vector $\boldsymbol{x} = (x_{n_1}, \dots, x_{n_K})$. If the K observations are i.i.d. the

joint density is

$$\begin{aligned} Pr(X_1 = x_{n_1}, \ldots, X_K = x_{n_K} \mid \boldsymbol{\theta}) & \qquad (1.9) \\ = P(\boldsymbol{x}) &= \prod_k p(x_{n_k}) \\ &= \prod_k \exp(\, \langle \boldsymbol{\theta}, \boldsymbol{\phi}(x_{n_k}) \rangle - T(\boldsymbol{\theta})\,) \\ &= \exp\left(\sum_k \langle \boldsymbol{\theta}, x_{n_k} \rangle - K\, T(\boldsymbol{\theta}) \right). \qquad (1.10) \end{aligned}$$

Suppose now an experiment has taken place and we can fix $\boldsymbol{x} = \widehat{\boldsymbol{x}}$. Maximizing this expression for the joint likelihood gives the *maximum likelihood* (ML) estimate of $\boldsymbol{\theta}$. Typically this is done by instead minimizing $-\log P$, which can be written as:

$$\begin{aligned} -\log P(\widehat{\boldsymbol{x}}) &= -\sum_k \langle\, \boldsymbol{\theta}, \phi(\widehat{\boldsymbol{x}})\, \rangle + K\, T(\boldsymbol{\theta}) \\ &= -\left\langle \boldsymbol{\theta}, \sum_k \phi(\widehat{\boldsymbol{x}}) \right\rangle + K\, T(\boldsymbol{\theta}). \qquad (1.11) \end{aligned}$$

Now defining $\widehat{\boldsymbol{b}} = \sum_k \phi(\widehat{\boldsymbol{x}})/K$, the empirical mean of the observed statistics makes the last expression equal to

$$-K \left\langle \boldsymbol{\theta}, \widehat{\boldsymbol{b}} \right\rangle + K\, \log Z(\boldsymbol{\theta}) \;\propto\; -\left\langle \boldsymbol{\theta}, \widehat{\boldsymbol{b}} \right\rangle + \log Z(\boldsymbol{\theta}).$$

The factor K is typically dropped since it does not affect optimization.

The ML problem then becomes:

$$\boxed{\underset{\boldsymbol{\theta}}{\text{minimize}} -\left\langle \boldsymbol{\theta}, \widehat{\boldsymbol{b}} \right\rangle + \log Z(\boldsymbol{\theta})}. \qquad (1.12)$$

The relationship between this formalism and the one represented by (1.2) stems from the fact that the latter yields the *same* parameters. To make the connection clear, first set

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{B} \\ \mathbf{1}^{\mathrm{T}} \end{pmatrix},$$

i.e., include a normalization constraint in $\boldsymbol{A}$. Next identify the statistics ϕ with the matrix $\boldsymbol{B}$ in the following way: $\boldsymbol{B}_{\cdot n} = \phi(x_n)$, that is, the n-th column of $\boldsymbol{B}$.

Also split $\boldsymbol{\mu}$ conformably with $\boldsymbol{A}$, letting

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\theta} \\ \theta_{\mathbf{1}} \end{pmatrix},$$

where $\boldsymbol{\theta}$ is the vector of parameters associated with the feature constraints and $\theta_{\mathbf{1}}$ is a scalar parameter associated with the normalization constraint. Assume for the moment that we can set $\theta_{\mathbf{1}} = -T(\boldsymbol{\theta})$ as necessary (Sufficient conditions for the existence and properties of the function T is discussed in Chapter 3.) Using the foregoing transformations, the dual problem objective can be reworked as follows:

$$\begin{aligned}
&S^*(\boldsymbol{A}^*\boldsymbol{\mu}) - \langle \boldsymbol{b}, \boldsymbol{\mu} \rangle \\
&= S^*(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{\theta} + \mathbf{1}\theta_{\mathbf{1}}) - \left\langle \widehat{\boldsymbol{b}}, \boldsymbol{\theta} \right\rangle - \theta_{\mathbf{1}} \cdot 1 \\
&= \sum s^*(\langle \phi(x_n), \boldsymbol{\theta} \rangle - T(\boldsymbol{\theta})) - \left\langle \widehat{\boldsymbol{b}}, \boldsymbol{\theta} \right\rangle + T(\boldsymbol{\theta}) \\
&= \sum \exp(\langle \phi(x_n), \boldsymbol{\theta} \rangle - T(\boldsymbol{\theta}) - 1) - \left\langle \widehat{\boldsymbol{b}}, \boldsymbol{\theta} \right\rangle + T(\boldsymbol{\theta}) \\
&= C \left(\sum \exp(\langle \phi(x_n), \boldsymbol{\theta} \rangle - T(\boldsymbol{\theta})) \right) - \left\langle \widehat{\boldsymbol{b}}, \boldsymbol{\theta} \right\rangle + T(\boldsymbol{\theta}) \\
&= C \cdot 1 - \left\langle \widehat{\boldsymbol{b}}, \boldsymbol{\theta} \right\rangle + T(\boldsymbol{\theta}). \qquad (1.13)
\end{aligned}$$

Dropping the constant C from the last line and comparing this to (1.12) we can see that (1.2) will indeed yield the same parameters, as long as the data supplied to the maxent problem is $\boldsymbol{b} = (\widehat{\boldsymbol{b}}, 1)$. Note that in this case the data vector consists of the empirical mean of the sample, along with an explicit setting for the sum of the state probabilities. In maxent problems from other disciplines outside of statistics, the data vector $\boldsymbol{b}$ can be set by other measurements, e.g. prices in a finance example.

In practice, maximum likelihood is often associated with overfitting. The *maximum a posteriori* (MAP) estimation procedure attempts to address this problem by incorporating what is termed a prior distribution or regularizer for $\boldsymbol{\theta}$. It is only slightly more complicated to redo the foregoing with such a prior. A sketch should suffice. Instead of starting with (1.9), we begin with $P(\boldsymbol{x}, \boldsymbol{\theta}) = P(\boldsymbol{x}|\theta)P(\boldsymbol{\theta})$ and assume the first term is an exponential family distribution, parameterized by θ. Following the same analysis that led to the ML problem shows that the MAP problem can stated as:

$$\boxed{\underset{\boldsymbol{\theta}}{\text{minimize}} - \left\langle \boldsymbol{\theta}, \widehat{\boldsymbol{b}} \right\rangle + \log Z(\boldsymbol{\theta}) - \log P(\boldsymbol{\theta})} \qquad (1.14)$$

In the Bayesian terminology, the last term is the negative log prior, while in the

approximation literature, the same term is said to play the role of a regularizer. It is possible to map both the ML and the MAP problem, into the generalized maxent framework. A fairly general form of constraint relaxation has the form $G(\boldsymbol{A}\boldsymbol{p} - \boldsymbol{b})$ (see Chapter 3), which leads to the dual problem:

$$\underset{\boldsymbol{\mu}}{\text{minimize}}\, S^*(\boldsymbol{A}^*\boldsymbol{\mu} + \boldsymbol{p}_0^*) - \langle \boldsymbol{b}, \boldsymbol{\mu} \rangle + G^*(-\boldsymbol{\mu}). \tag{1.15}$$

Now comparing the objective (1.13) with (1.12) and (1.14), the correspondence comes down to whether the relaxation term of the generalized maxent problem corresponds to the negative log prior. For arbitrary priors the correspondence will break down. Usually, however, the prior is chosen to maintain the convexity of the optimization problem, therefore this issue is not often encountered.

The foregoing also shows why it is a good idea to distinguish clearly between the initial guess and the prior. In the maxent literature the initial guess is sometimes confusingly called a prior distribution. The latter is a term perhaps better applied to a distribution over parameters rather than states. This distinction will be maintained. Looking ahead, priors will be better identified with regularization functions for the parameters of $\bar{\boldsymbol{p}}$, arising out of constraint relaxations, not the initial guess $\boldsymbol{p}_0$, which is a distribution over states, not parameters.

In the maximum likelihood problem (1.12), the log partition function features prominently, and a great deal of machinery is available to estimate such models (Wainwright and Jordan, 2003b). Under fairly mild conditions it can be shown to be C^∞, convex, and the cumulant generating function of the resulting model. In particular its gradient matches the expectation of the features under the maximum likelihood model.[4] See (Proposition 2 Wainwright and Jordan, 2003b)

Having sketched the correspondence between the generalized maxent framework and both ML and MAP estimation, a few comments are in order. First the exponential family approach to model-building hardwires several assumptions. The first of course is SBG entropy, since this is what gives rise to the exponential family. Another assumption is normalization. This is implicit in the use of the log partition function Z in the objective function. If one builds a model starting with the log-partition function it is hardly possible to relax this assumption. In the generalized maxent framework various relaxations are possible and normalized models can be regarded as belonging to a continuum of models, the rest of which are not normalized. Later, when we relax the SBG entropy/exponential family assumption, a form of the log partition function survives for some entropy functions, but generally it will be more natural to work with the conjugate of the entropy function, as in (1.5).

[4] Or, indeed the resulting MAP model, if that estimation method is employed.

1.3 Outline and Contributions

This thesis draws on results from statistical physics, information geometry and machine learning. Given this diversity, some effort was expended in unifying notation as far as possible. A great deal of this was made possible by using the tools and notation of convex analysis. To provide the common ground needed, Chapter 2 establishes notation and presents what I hope is a useful primer on convex analysis including information on 'conjugate calculus'. This topic is very helpful in deriving dual representations of optimization problems.

Chapter 3 establishes the duality theory utilized in the rest of the thesis. It is based upon a number of sources, most notably Rockafellar and Wets (1998). In that chapter we also look once again at the normalizing function discussed in Section 1.2 and establish the connection to escort probabilities. Escort probabilities are known in statistical physics and are briefly touched in a machine learning paper (Lafferty, 1999), but a presentation of the connection based on Fenchel duality I have not seen elsewhere. The ability to get a 'nice' characterization of escorts is dependent on whether the entropy enjoys the Legendre property. I suspect that this fact is known in some form or another to researchers in statistical physics who take the ability to perform the Legendre transform for granted in their work, but the chapter makes an explicit connection.

As mentioned before, this thesis explores the freedom to build machine learning models with different entropy functions. At this point it is natural to ask on what basis such alternative entropies should be chosen. Or, even what exactly *is* entropy? These questions are somewhat beyond the scope of the thesis and are active areas of research (see e.g. Niven, 2007). Instead we focus here on the more direct question of what different entropy functions *do* and how to work with them. In Chapter 4 we follow the suggestion offered by Naudts in a series of papers (Naudts, 2004a,b, 2002) and use the notion of deformed logarithms to construct parameterized families of entropy functions. These lead to the study of functions called ϕ-logarithms and ϕ-exponentials. A number of notable objective functions are in this class, including SBG entropy, and Tsallis entropy (Gell-Mann and Tsallis, 2004). It turns out that deformed logarithms do not always enjoy the Legendre property. The chapter provides a simple test of when they do. The related entropy functions are useful tools to model heavy-tailed distributions or distributions with limited support (light tails), where one may need to look beyond exponential families.

In Chapter 5 we explore a series of extended examples to suggest that the framework is in fact useful for understanding and extending existing approaches to model building in machine learning. In the first three sections we find that the generalized maxent problem has been studied under various names in different fields. One aim of this thesis is to show that several well-studied problems

in convex analysis (e.g. Pietra et al., 2002a; Bauschke, 2003), and information geometry (e.g. Murata et al., 2004; Eguchi, 2005) are instances of the generalized maxent problem (3.23) with exact constraint matching. In each case we show that the main theorems in these papers can be reconstructed simply by using basic concepts from convex analysis along with Fenchel duality. We also offer a perspective on boosting and regression models. Boosting is an example of what could be termed *conditional-empirical models*. It does two things to alter the classic problem. One is to utilize the product rule to factorize the distribution: $p(x, y) = p(y|x)p(x)$. The second is to replace one of the factors, say $p(x)$, with its empirical counterpart. This is a strong modeling assumption, since it places zero weight on a vast portion of the state space. After imposing it, the original problem can be reformulated as a much smaller one, with considerable practical benefits. In machine learning this is referred to as the *discriminative* approach, as opposed to the *generative* one (Jebara, 2003).

It should also be stressed that some of the most interesting models come from studying non-Legendre entropy functions. As seen in the Dantzig selector examples, these can be associated with sparse models. One such example, involving an application to non-negative matrix factorization (NMF) is covered in Chapter 6. NMF is an interesting and slightly surprising application as all the well-known approaches to this problem involve non-convex optimization. The model introduced here is also non-convex, except that the algorithm uses two alternating generalized maxent problems as substeps. However, the theory remains relevant because the dual characterization of the substeps shows clearly how sparsity in the solution is achieved. This example also provides an opportunity to exercise the computer code that was written for the thesis on two real-world data sets. Hopefully the diversity of the data sets gives a hint of the wide range of potential applications.

Another important reason that exponential families are popular in machine learning is that they represent a 'perfect' fit with graphical models. This fit is established by the Hammersley-Clifford Theorem. For undirected graphical models the Hammersley-Clifford theorem (see Lauritzen, 1996) is fundamental for relating the structure of the graph to the factorizations of the probability density. This theorem links the conditional independence assumptions encoded in the structure of a graph to the factorization properties of a probability distribution. Exploiting the factorization property is often the key to developing an efficient estimation algorithm, while conditional independence is often considered a form of prior information available to the modeler.

Still more complicated factorizations can be expressed using graphical models. Among these are two basic types: directed graphs and undirected graphs. The former are associated with Bayesian Networks while the latter yield Markov Random Fields (MRFs). Hidden Markov Models (HMMs), Maximum Entropy

Markov Models (MEMMs) and Conditional Random Fields (CRFs), are all well-known instances of graphical models (Lafferty et al., 2001). At the moment all the examples the author knows of employ exponential families, and are therefore based on SBG entropy. Is this happy marriage destroyed by moving outside the exponential family? Chapter 7, takes a first step to address the question. It presents a generalization of the Hammersley-Clifford Theorem in which, by adjusting our view about the semantics of the edges in a graph, factorization of the probability distribution is restored. Chapter 8 concludes with a summary and a discussion of possible future research directions.

Chapter 2

Mathematical Background

This chapter sets out much of the notation and covers some of the results needed to make the thesis as self-contained as possible.

2.1 Essential Convex Analysis Concepts

Why study convex analysis? One compelling reason is that convex analysis provides tools that unify the expression of smooth and non-smooth optimization problems, as well as constrained and unconstrained problems. In this section, a primer on convex analysis is presented. An attempt has been made to use standard notation as far as possible. Where confusion of terminology or notation may arise, a footnote of explanation has been added. Many of the results in this chapter can be found in, or distilled from, one or more of the following excellent texts Rockafellar (1970); Hiriart-Urruty and Lemarechal (2001); Borwein and Lewis (2000); Rockafellar and Wets (1998); Luenberger (1969). Readers can scan the checklist of topics covered in this section to see if they can skip ahead:

- Set attributes: int, ri, bdry.
- The extended reals.
- Convex set, convex function.
- For a convex function F set $\operatorname{dom} F$.
- Epigraph, sublevel set.
- Level-bounded function, coercive function.
- The values $\inf F$, $\sup F$, $\min F$. The set $\operatorname{argmin} F$.
- Closed, proper, lower semicontinuous functions.
- Continuity and differentiability of convex functions.

- Subgradient of a convex function.
- The Fenchel-Legendre conjugate.
- Operations preserving convexity *e.g.* epi-addition, epi-multiplication, epi-composition. definitions preserve convexity.
- Function properties transformed in duality.
- The Legendre property, essential strict convexity, essential smoothness.
- 'Convex calculus', examples of convex functions and their conjugates.

2.1.1 Basic Notation

$\mathcal{X}$ and its dual, $\mathcal{X}^*$, denote copies of the Euclidean space $\mathbb{R}^N$. It is useful to distinguish between $\mathcal{X}$ and its dual, even though they are isomorphic. The variables $\boldsymbol{p}, \boldsymbol{q}, \boldsymbol{x}$, etc. denote elements of $\mathcal{X}$, and $\boldsymbol{p}^*, \boldsymbol{q}^*, \boldsymbol{x}^*$ etc. denote elements of the dual space $\mathcal{X}^*$. The two spaces are connected via the usual dot product $\langle \boldsymbol{p}, \boldsymbol{p}^* \rangle = \sum_n p_n p_n^*$.

For sets $\mathcal{C}, \mathcal{D} \subseteq \mathcal{X}$ and $\alpha, \beta \in \mathbb{R}$, the Minkowski sum $\alpha\mathcal{C} + \beta\mathcal{D}$ is given by

$$\alpha\mathcal{C} + \beta\mathcal{D} = \{\alpha\boldsymbol{p} + \beta\boldsymbol{q} : \boldsymbol{p} \in \mathcal{C}, \boldsymbol{q} \in \mathcal{D}\}.$$

The meaning of the notation $\boldsymbol{p} + \mathcal{B}$ is that of $\{\boldsymbol{p}\} + \mathcal{B}$. The meaning of $\mathcal{A} - \mathcal{B}$ corresponds to the same definition, of taking $\alpha = 1$ and $\beta = -1$. The definition also straightforwardly resolves statements like the following.

$$\mathbf{0} \in \mathcal{A} - \mathcal{B} \Leftrightarrow \exists \boldsymbol{a} \in \mathcal{A}, \boldsymbol{b} \in \mathcal{B} \text{ such that } \boldsymbol{a} - \boldsymbol{b} = \mathbf{0} \Leftrightarrow \mathcal{A} \cap \mathcal{B} \neq \emptyset$$

The interior $\operatorname{int}\mathcal{C}$ of a set $\mathcal{C} \subseteq \mathcal{X}$ consists of all $\boldsymbol{p} \in \mathcal{C}$ for which some ϵ-ball centered at $\boldsymbol{p}$ is contained in $\mathcal{C}$. Interiors of Minkowski sums are sometimes encountered.

Fact 2.1 (Interiors) $\mathbf{0} \in \mathcal{A} \cap \operatorname{int}\mathcal{B}$ *or* $\mathbf{0} \in \operatorname{int}\mathcal{A} \cap \mathcal{B} \implies \mathbf{0} \in \operatorname{int}(\mathcal{A} - \mathcal{B})$ *and the reverse implication does not hold.*

Some intuition for this result is that the set $\mathcal{A} - \mathcal{B}$ is actually larger than its two components.

Images of sets are used as well. If $\boldsymbol{A}$ is a linear map between $\mathcal{X}$ and $\mathcal{M}$ and $\mathcal{C} \subset \mathcal{X}$, then $\boldsymbol{A}\mathcal{C} = \{\, \boldsymbol{y} \in \mathcal{M} \mid \boldsymbol{y} = \boldsymbol{A}\boldsymbol{p} \text{ for some } \boldsymbol{p} \in \mathcal{C} \,\}$.

A set $\mathcal{L} \subseteq \mathcal{X}$ is called a *linear subspace* if $\mathcal{L}$ is nonempty and $\alpha\mathcal{L} + \beta\mathcal{L} \subseteq \mathcal{L}$ for all $\alpha, \beta \in \mathbb{R}$. An affine subspace, or *hyperplane* is a translated linear subspace. It has the form $\mathcal{L} + \boldsymbol{x}_0$, for some linear subspace, $\mathcal{L}$, and a vector, $\boldsymbol{x}_0$. The affine

hull of a set consists of the smallest affine subspace containing it. In particular the affine hull of $\mathcal{C}$ is the set:

$$\{\, \boldsymbol{v} \mid \alpha \boldsymbol{v}_1 + (1-\alpha)\boldsymbol{v}_2, \alpha \in \mathbb{R}, \boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathcal{C} \,\},$$

containing all the points which can be expressed as convex combinations of points from the original set. We often think of a hyperplane as having dimension one less than the space it is embedded in, but this need not be the case for an affine hull, which will inherit the dimensionality of the set generating it.

The relative interior $\operatorname{ri}\mathcal{C}$ of a convex set $\mathcal{C}$ consists of all points $\boldsymbol{p} \in \mathcal{C}$ for which the intersection of the affine hull of $\mathcal{C}$ with some ϵ-ball centered at $\boldsymbol{p}$ is contained in $\mathcal{C}$. The boundary, $\operatorname{bdry}\mathcal{C}$, is the set of non-interior points of $\mathcal{C}$.

It is useful to allow a function to take on the values $+\infty$, and/or $-\infty$. This is helpful in defining functions over the whole of a space, which are normally restricted to certain sets. For this purpose define the *extended reals*: $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$. There is some downside to this practice. Arithmetic on the extended reals requires some care. It is important to avoid the occurrence of expressions equivalent to $\infty - \infty$ and ∞/∞, which are not defined.

2.1.2 Convexity

A set $\mathcal{C}$ is *convex* if $\forall x \in \mathcal{C}$, $y \in \mathcal{C}$ and $\alpha \in (0,1)$, then $\alpha x + (1-\alpha)y \in \mathcal{C}$. The concept is illustrated in Figure 2.1.

Fact 2.2 *Convexity of sets is preserved under arbitrary intersections, affine maps, Cartesian products, direct sums, taking the interior and taking the closure.*

Figure 2.1 provides a simple illustration of convexity based on convex combinations.

For the rest of this chapter F refers to a function $F : \mathcal{X} \to \bar{\mathbb{R}}$. F has *effective domain*, $\operatorname{dom} F = \{\boldsymbol{x} \mid F(\boldsymbol{x}) < +\infty\}$. F is *proper* if there is at least one $\boldsymbol{x}$ such that $F(\boldsymbol{x}) < \infty$ and $F(\boldsymbol{x}) > -\infty, \forall \boldsymbol{x} \in \mathcal{X}$. *In other words, a proper function has a non-empty effective domain, where it is finite (*i.e., *bounded below) and outside of which it takes on the value* $+\infty$.

The *epigraph* of F is the set $\operatorname{epi} F = \{(\boldsymbol{x}, \alpha) \in \mathcal{X} \times \mathbb{R} : F(\boldsymbol{x}) \geq \alpha\}$. Examples of epigraphs appear in in Figures 2.2 and 2.3.

F is called a *convex function* if $\operatorname{epi} F$ is a convex set. This is equivalent to Jensen's inequality: $\forall \boldsymbol{x}, \boldsymbol{y} \in \operatorname{dom} F$ and $\alpha \in (0,1)$, then $\alpha F(\boldsymbol{x}) + (1-\alpha)F(\boldsymbol{y}) \geq F(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y})$. These two alternative characterizations of convexity form a typical pattern of convex analysis. Operations which may yield convex functions can be verified by examining the resulting function or its epigraph, whichever is easier.

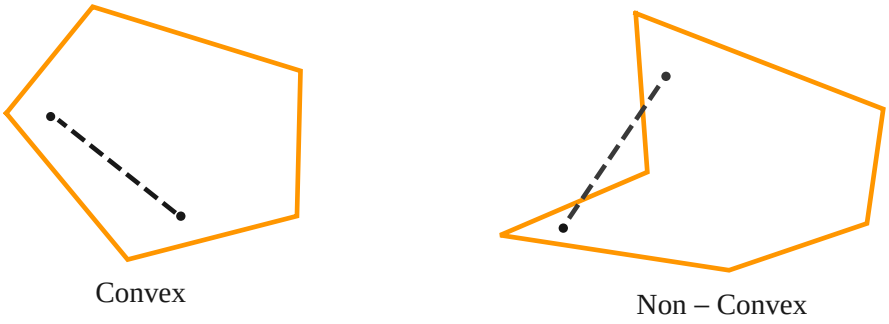


Figure 2.1: Set convexity.

A function is *strictly convex* if '$\leq$' is replaced with"$<$" in Jensen's inequality, The difference is illustrated in Figure 2.2 Proper convex functions are of primary

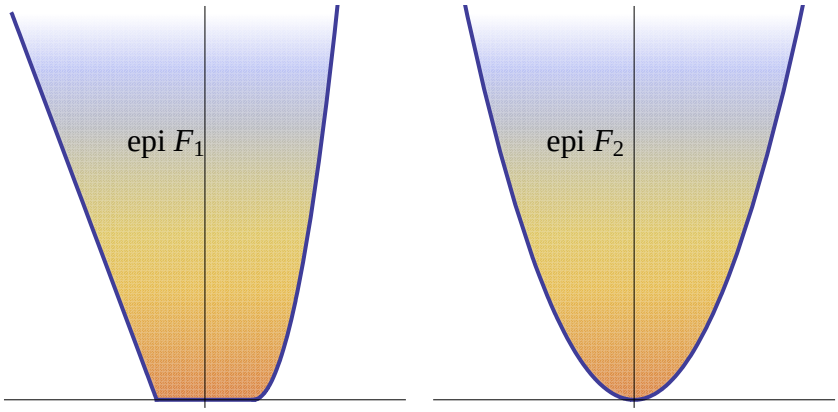


Figure 2.2: Strict convexity.

interest in this work.

Theorem 2.3 (Convex domain) *The effective domain of a proper convex function is a convex set.*

Proof Suppose $\boldsymbol{p}, \boldsymbol{q} \in \operatorname{dom} F$. Then for $\alpha \in (0,1)$, let $\boldsymbol{x} = \alpha\boldsymbol{p} + (1-\alpha)\boldsymbol{q}$. In the following, the first inequality is due to convexity and the second is due to the assumed finite values of $F(\boldsymbol{p})$ and $F(\boldsymbol{q})$:

$$F(\boldsymbol{x}) = F(\alpha\boldsymbol{p} + (1-\alpha)\boldsymbol{q}) \leq \alpha F(\boldsymbol{p}) + (1-\alpha)F(\boldsymbol{q}) < \infty,$$

which implies the result. ■

Fact 2.4 (Continuity) *A convex proper function is continuous on the interior of its effective domain. Proof: (Sec. 7.9. Corollary 1, Luenberger, 1969).*

Many properties of convex functions hold on the interior of its effective domain. If a convex function with $\dim(\operatorname{dom} F) < \dim \mathcal{X}$ is encountered, the interior of its domain is actually empty, but its relative interior, denoted $\operatorname{ri}(\operatorname{dom} F)$ may not be empty. If the effective domain has full dimension, the relative interior and the interior coincide. For this reason, statements involving the relative interior are of greater generality, since they place no restriction on the dimensionality of $\operatorname{dom} F$. In practice, the loss of generality is very mild. For simplicity, the discussion in this chapter uses int instead of ri. In later chapters, if the distinction is crucial, ri will be used.

2.1.3 Closure

Closure is another important property. A convex proper function F is said to be closed if its epigraph is a closed set. Equivalently,

Theorem 2.5 (Closure) *F is closed if, for any fixed $\boldsymbol{p}_0 \in \operatorname{int}(\operatorname{dom} F)$, and any $\boldsymbol{p} \in \mathcal{X}$,*

$$F(\boldsymbol{p}) = \lim_{\lambda \uparrow 1} F((1-\lambda)\boldsymbol{p}_0 + \lambda \boldsymbol{p}).$$

A closed function is therefore well-behaved at the boundary of its domain. To see the difference between a non-closed versus a closed convex function compare

$$F(x) = \begin{cases} \left(x - \frac{1}{2}\right)^2 \text{ if } x > \frac{1}{2} \\ +\infty \text{ if } x \leq \frac{1}{2} \end{cases} \quad \text{to} \quad F(x) = \begin{cases} \left(x - \frac{1}{2}\right)^2 \text{ if } x \geq \frac{1}{2} \\ +\infty \text{ if } x < \frac{1}{2} \end{cases} \tag{2.1}$$

The two functions are depicted in Figure 2.3. The heavy line indicates points included in the epigraph. The first function is not closed because $F(1/2) = +\infty$ (instead of 0). Therefore there is a convergent sequence $\{(x_n, F(x_n)\} \to (1/2, 0)$, which does not match the value $(1/2, F(1/2)) = (1/2, +\infty)$.

A closed function behaves better as the objective of a minimization problem, since it attains its infimum while the non-closed function does not. In optimization it is highly desirable to work with functions that are closed, convex and proper. This conjunction will be denoted with the shorthand *ccp*.

The operation to close the set $\operatorname{epi} F$ by adding its limit points is familiar from real analysis. The new set $\operatorname{cl} \operatorname{epi} F$ corresponds to the epigraph of a function $\operatorname{cl} F$,

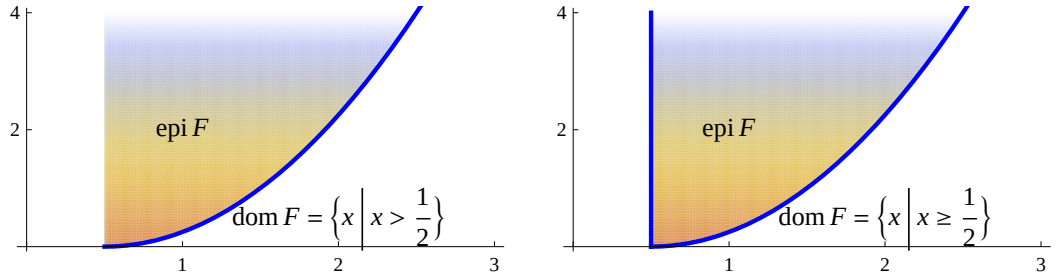


Figure 2.3: Open versus closed convex function (Equation 2.1.)

which is obviously closed and convex. The new function satisfies: $\mathrm{cl}\, F \leq F$, with equality holding if F was closed already.

Often an introduced function is assumed to be ccp. If a function arises as the result of some operations on ccp functions, the resulting function may or may not remain ccp. A common strategy is to introduce side assumptions to maintain the ccp property. Closure is typically the frailest of the properties involved, so additional assumptions are often aimed specifically at ensuring the resulting function is closed.

2.1.4 Subgradients

The concept of a subgradient is a generalization of the gradient concept from calculus. It turns out to be especially natural for convex functions, including those which are not necessarily smooth. For smooth functions, the concept of a gradient can be used to define a line, or in higher dimensions, a hyperplane which supports (is tangent to) the function at a point. The gradient represents the slope of the line, or more generally, the parameters defining the hyperplane. Conversely this generalization allows any nonvertical hyperplane, if any exists, to serve as a subgradient at a given point.

The *one-sided directional derivative* of F at $\boldsymbol{p}$ in the direction $\boldsymbol{d}$ is defined as:

$$F'(\boldsymbol{p};\boldsymbol{d}) := \lim_{t \downarrow 0} \frac{F(\boldsymbol{p}+t\boldsymbol{d}) - F(\boldsymbol{p})}{t}. \tag{2.2}$$

Note that this is a less restrictive version of the Gateaux derivative, since the limit is not taken from either direction. If it exists and is equal to the same value in every direction $\boldsymbol{d}$, the function is (Frechet) differentiable in the usual sense (see *e.g.* Sec. 7.2 Luenberger, 1969). This definition of a derivative turns out to be closely related to the subgradient concept, defined next.

If F is proper and convex and $\boldsymbol{p}$ is a point where F is finite, that is, $\boldsymbol{p} \in \mathrm{dom}\, F$,

then $\boldsymbol{p}^* \in \mathcal{X}^*$ is called a *subgradient* of F at $\boldsymbol{p}$ if

$$F(\boldsymbol{q}) \geq F(\boldsymbol{p}) + \langle \boldsymbol{p}^*, \boldsymbol{q} - \boldsymbol{p} \rangle, \forall \boldsymbol{q} \in \mathcal{X}. \tag{2.3}$$

The set of all subgradients at a point is called the *subdifferential* of F at $\boldsymbol{p}$ and is denoted by $\partial F(\boldsymbol{p})$. If $F(\boldsymbol{p}) = +\infty$, that is, if $\boldsymbol{p} \notin \operatorname{dom} F$, then $\partial F(\boldsymbol{p}) := \emptyset$.

If $\partial F(\boldsymbol{x})$ is not empty then F is said to be *subdifferentiable* at $\boldsymbol{x}$. If the set $\partial F(\boldsymbol{x})$ is a singleton, then the function is said to be *differentiable* at $\boldsymbol{p}$ and we use $\nabla F(\boldsymbol{p})$ to denote its gradient at $\boldsymbol{p}$. The subdifferential mapping, ∂F, also denotes a set-valued function. Thinking of ∂F this way, its domain is defined as

$$\operatorname{dom} \partial F = \{ \boldsymbol{p} \in \operatorname{dom} F \mid \partial F \neq \emptyset \}.$$

Subgradients often exist when gradients are not defined. For example, the function $f(x) = |x|$ is differentiable everywhere except zero. However, its subdifferential is not empty there. Instead it consists of the interval $[-1, +1]$, corresponding to the slopes of the lines which support the function at $x = 0$. Each such supporting function must pass through the shaded area shown in Figure 2.4.

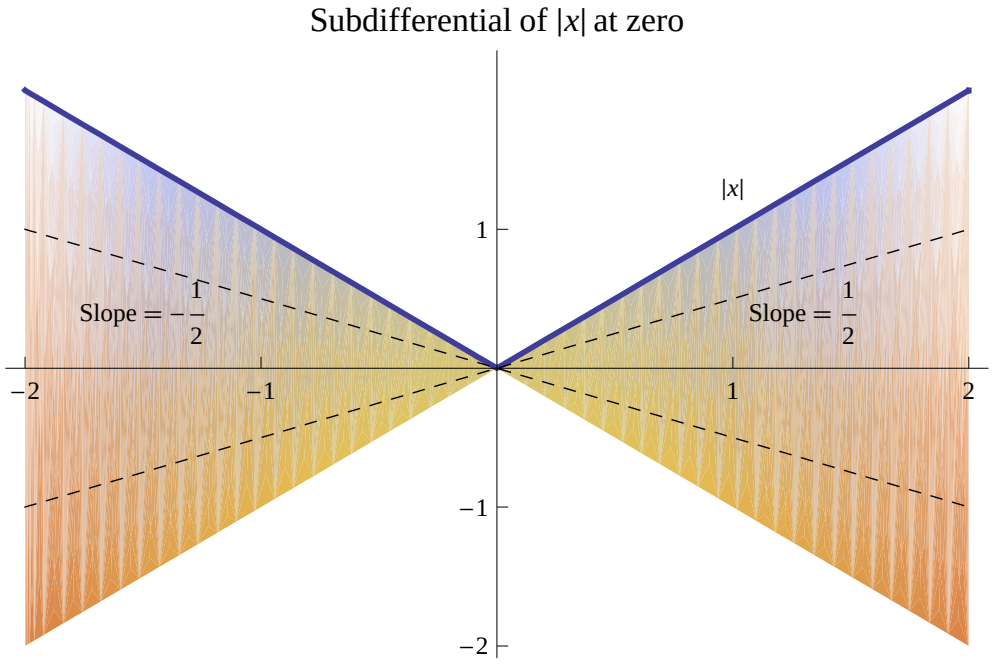


Figure 2.4: Subgradient of $|x|$ at zero consists of the set: $[-1, 1]$ corresponding to the slopes of affine functions supporting $|x|$ at zero. Each supporting affine function corresponding to a subgradient passes through the shaded area.

Note that in contrast to the one-sided directional derivative (2.2), the defini-

tion of a subgradient does not even mention limits. However, the following Fact illustrates that the two concepts are in close agreement.

Fact 2.6 (Subgradient properties) *If F is convex and proper.*

i. $\boldsymbol{p} \in \text{int}(\text{dom}\, F) \Rightarrow \partial F(\boldsymbol{p}) \neq \emptyset$.

ii. $\boldsymbol{p} \in \text{int}(\text{dom}\, F) \Rightarrow F'(\boldsymbol{p}; \boldsymbol{d}) = \sup_{\boldsymbol{p}^* \in \partial F(\boldsymbol{p})} \langle \boldsymbol{p}^*, \boldsymbol{d} \rangle$

iii. $\boldsymbol{p} \in \text{dom}(\partial F)$ *and* $\partial F(\boldsymbol{p})$ *is bounded* $\Leftrightarrow \boldsymbol{p} \in \text{int dom}\, F$.

iv. $\boldsymbol{p} \in \text{dom}\, F$ *and* $\partial F(\boldsymbol{p}) = \emptyset \Rightarrow F'(\boldsymbol{p}; \boldsymbol{d}) = -\infty$

Proof: (Thms. 23.3-4 Rockafellar, 1970).

Note (iv) implied that the supremum in (ii) is achieved since ∂F is also closed, having been defined with non-strict inequalities.

As the Fact suggests, the subdifferential may or may not be empty on the boundary of its domain. Consider the function $f : \mathbb{R} \to \bar{\mathbb{R}}$, where

$$f(p) = \begin{cases} p \log(p), & p > 0 \\ 0, & p = 0 \\ +\infty, & p < 0 \end{cases}.$$

This function has $\text{dom}\, f = [0, +\infty)$ (and is ccp). At $p = 0, \partial f = \emptyset$. Contrast this with the function

$$g(p) = \begin{cases} p^{3/2} - p, & p > 0 \\ 0, & p = 0 \\ +\infty, & p < 0 \end{cases}.$$

It has the same domain as f, but $\partial g(0) = (-\infty, -1]$. Fact 2.6 and these two examples indicate that $\text{int}(\text{dom}\, F) \subset \text{dom}\, \partial F \subset \text{dom}\, F$, sometimes strictly so. A similar situation is illustrated in Figure 2.5

Fact 2.7 (Differentiability of strictly convex functions) *A proper, strictly convex function is differentiable on the interior of its effective domain. Proof: Use Fact 2.6 and strict inequality to rule out non-unique subgradients.*

Some other properties of ∂F relate to the conjugate of F and are deferred to the next subsection.

One of the most useful features of calculus involves the rules for calculating derivatives leading up to the chain rule. With due regard for the set operations involved, subgradient calculus rules can be established.

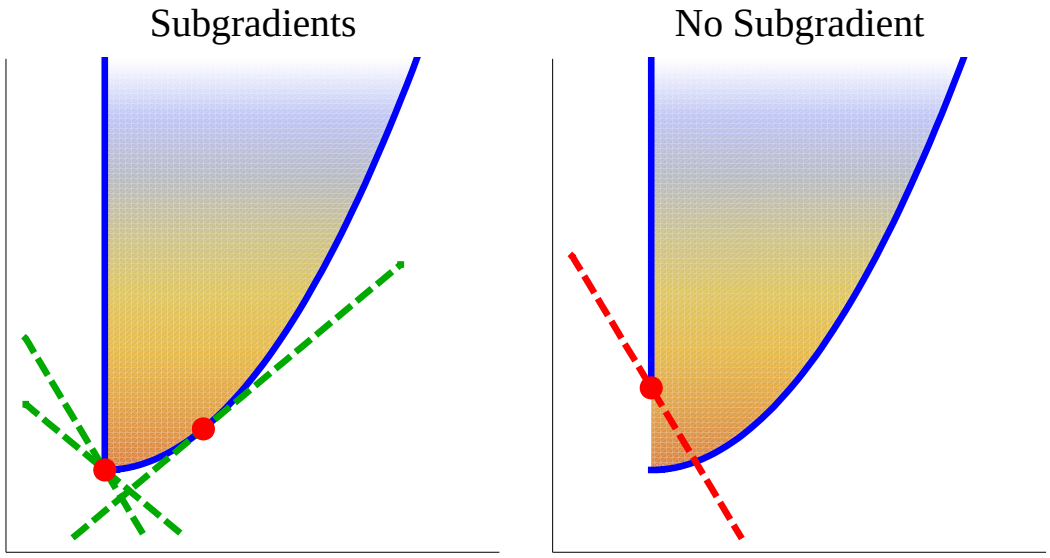


Figure 2.5: Subgradients at the boundary may or may not exist.

Fact 2.8 *Let $F, G : \mathcal{X} \to \bar{\mathbb{R}}$ be ccp and $\boldsymbol{A} : \mathcal{X} \to \mathcal{M}$ be a linear map with $(A \operatorname{dom} F) \cap \operatorname{dom} G \neq \emptyset$. Then if $H(\boldsymbol{p}) = F(\boldsymbol{p}) + G(\boldsymbol{A}\boldsymbol{p})$*

$$\partial H(\boldsymbol{p}) = \partial F(\boldsymbol{p}) + \boldsymbol{A}^* \partial \boldsymbol{G}^*(\boldsymbol{p}^*),$$

where the '+' represents the Minkowski sum.
Proof: (Thm 23.9 Rockafellar, 1970)

2.1.5 Fenchel Conjugation

The concept of conjugation is very important for establishing duality results later on. It is also the basis of much of the 'convex calculus' used in this work. The Fenchel conjugate $F^* : \mathcal{X}^* \to (-\infty, \infty]$, also known as the Fenchel-Legendre conjugate, is defined as

$$F^*(\boldsymbol{p}^*) := \sup_{\boldsymbol{p} \in \mathcal{X}} \left\{ \langle \boldsymbol{p}^*, \boldsymbol{p} \rangle - F(\boldsymbol{p}) \right\}. \tag{2.4}$$

F^* may be referred to simply as a conjugate function. There are many important connections between functions and their conjugates.

Theorem 2.9 *If F is proper, F^* is ccp. Proof: (B.2.1.3, Hiriart-Urruty and Lemarechal, 2001).*

Proof (Alternate) From the definition (2.4), F^* is constructed from a set of affine functions, indexed by $\boldsymbol{p}$, each of which is closed and convex. In (2.4),

taking the sup corresponds to arbitrary intersections of the epigraphs of the linear functions in the set. This operation preserves closure and convexity of epi F^*. Since F is proper, $\operatorname{dom} F \neq \emptyset$, therefore $F(\boldsymbol{p})$ is finite somewhere, and therefore $F^*(p^*) > -\infty, \forall \boldsymbol{p}^*$ and $F^*(\boldsymbol{p}^*) < +\infty$ for some $\boldsymbol{p}^*$. ■

Note that this Fact does not require F to be closed or convex.

Fact 2.10 *[Biconjugation] If F is proper $F^{**} \leq F$.*
Proof: (E.1.3.5, Hiriart-Urruty and Lemarechal, 2001).

The epigraph of a biconjugate is equal to the closed convex hull of the original function. This only adds points to the epigraph and any additional points must correspond to lower function values. The two functions coincide exactly when the function can be represented by an envelope of affine functions.

Fact 2.11 (Biconjugation of ccp functions) *[Biconjugation of ccp functions] If F is ccp then $F = F^{**}$.*
Proof: (E.1.3.6, Hiriart-Urruty and Lemarechal, 2001).

Theorem 2.12 (Fenchel-Young Inequality) *For all $\boldsymbol{p} \in \mathcal{X}$ and $\boldsymbol{p}^* \in \mathcal{X}^*$,*

$$F^*(\boldsymbol{p}^*) + F(\boldsymbol{p}) - \langle \boldsymbol{p}^*, \boldsymbol{p} \rangle \geq 0$$

Proof For all $\boldsymbol{p}^* \in \mathcal{X}^*$, the following holds:

$$F^*(\boldsymbol{p}^*) = \sup_{\boldsymbol{p} \in \mathcal{X}} \{\langle \boldsymbol{p}, \boldsymbol{p}^* \rangle - F(\boldsymbol{p})\} \geq \langle \boldsymbol{p}, \boldsymbol{p}^* \rangle - F(\boldsymbol{p}) \forall \boldsymbol{p} \in \mathcal{X}$$

Rearranging gives the result. ■

Among the most useful connections are those relating subgradients of F and F^*. To appreciate the following it is essential to bear in mind that ∂F and argmax refer to sets, not necessarily a single point.

Theorem 2.13 (F-Y equality and subgradients) *For $\boldsymbol{p} \in \operatorname{dom} F$, $\boldsymbol{p}^* \in \partial F(\boldsymbol{p})$ iff Fenchel-Young holds with equality.*

Proof From the definition of a subgradient, for a fixed $\boldsymbol{p} \in \operatorname{dom} F$ and $\boldsymbol{p}^* \in \partial F(\boldsymbol{p})$, the following holds $\forall \boldsymbol{q} \in \mathcal{X}$:

$$\begin{aligned} F(\boldsymbol{q}) &\geq F(\boldsymbol{p}) + \langle \boldsymbol{p}^*, \boldsymbol{q} - \boldsymbol{p} \rangle \\ -F(\boldsymbol{q}) + \langle \boldsymbol{p}^*, \boldsymbol{q} \rangle &\leq -F(\boldsymbol{p}) + \langle \boldsymbol{p}^*, \boldsymbol{p} \rangle \end{aligned}$$

Since the inequality holds for all $\boldsymbol{q}$ it holds for the supremum over $\boldsymbol{q}$ as well. But that makes the left-hand side equal to $F^*(\boldsymbol{p}^*)$. The last expression then becomes

$$F^*(\boldsymbol{p}^*) \leq \langle \boldsymbol{p}, \boldsymbol{p}^* \rangle - F(\boldsymbol{p}).$$

This is the reverse of the Fenchel-Young Inequality, therefore it can only hold with equality. ■

From this follows:

Fact 2.14 (Subgradient as maximizer)

$$\partial F(\boldsymbol{p}) = \underset{\boldsymbol{p}^*}{\operatorname{argmax}} \{ \langle \boldsymbol{p}^*, \boldsymbol{p} \rangle - F^*(\boldsymbol{p}^*) \}$$

The subgradient of F and F^* are related as well.

Theorem 2.15 (Subgradient inversion) *The subdifferential of a closed, convex, proper function, F and its conjugate F^* are related through the* inversion rule*: $\boldsymbol{p} \in \partial F^*(\boldsymbol{p}^*) \iff \boldsymbol{p}^* \in \partial F(\boldsymbol{p})$. In function/relation notation this is*

$$\partial F^* = (\partial F)^{-1} \text{ and } \partial F = (\partial F^*)^{-1}$$

Proof From two applications of the previous theorem we have:

$$\begin{aligned} \boldsymbol{p}^* \in \partial F(\boldsymbol{p}) &\Leftrightarrow 0 = F(\boldsymbol{p}) + F^*(\boldsymbol{p}^*) - \langle \boldsymbol{p}^*, \boldsymbol{p} \rangle \text{ and} \\ \boldsymbol{p} \in \partial F^*(\boldsymbol{p}^*) &\Leftrightarrow 0 = F^*(\boldsymbol{p}^*) + F^{**}(\boldsymbol{p}) - \langle \boldsymbol{p}^*, \boldsymbol{p} \rangle . \end{aligned}$$

From Theorem 2.9, $F^{**} = F$, so the right-hand conditions are identical, which gives the result. ■

These results are important in the optimization of convex functions, discussed in the next section.

2.1.6 Sublevel Sets

An *α-sublevel set* of a convex function F is denoted $\operatorname{lev}_\alpha F := \{\boldsymbol{p} \in \mathcal{X} \times \mathbb{R} \mid F(\boldsymbol{p}) \leq \alpha\}$. Sublevel sets are formed by the intersection of an epigraph and a half-space. Both are convex. Both are closed, being defined with a non-strict inequality. Hence the intersection is convex.

A ccp function, F, is *level-bounded* if $\forall \alpha \in \mathbb{R}, \operatorname{lev}_\alpha F$ is compact. Since sublevel sets are already closed, the additional requirement is that of boundedness. An example of a ccp function that is *not* level-bounded is $f(x) = \frac{1}{x}$(with $\operatorname{dom} f := \mathbb{R}_{++}$). A comparison with the level-bounded function $f(x) = x \log x$ (with $\operatorname{dom} F := \mathbb{R}_+$) is given in Figure 2.6. Level-boundedness is one of the

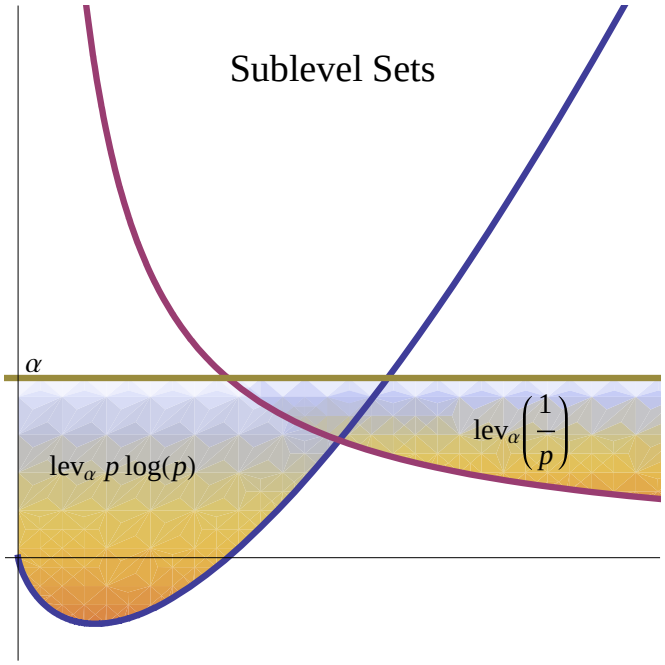


Figure 2.6: Sublevel sets for two functions, one level-bounded, the other not.

requirements for a convex function to attain its minimum. This is discussed further in the next section. First, two growth properties of functions turn out to be related to level-boundedness.

0-coercive and 1-coercive:

i. F is 0-coercive if

$$||\boldsymbol{p}|| \to +\infty \quad \Rightarrow \quad F(\boldsymbol{p}) \to +\infty.$$

ii. F is 1-coercive if

$$||\boldsymbol{p}|| \to +\infty \quad \Rightarrow \quad \frac{F(\boldsymbol{p})}{||\boldsymbol{p}||} \to +\infty.$$

Fact 2.16 *Assume F is ccp.*

i. F is level-bounded iff F is 0-coercive.

ii. $\boldsymbol{p} \in \operatorname{dom} F$ *iff* $F^*(\boldsymbol{p}^*) - \langle \boldsymbol{p}, \boldsymbol{p}^* \rangle$ *is 0-coercive.*

iii. F is 1-coercive iff $F - \langle \boldsymbol{p}^*, \cdot \rangle, \forall \boldsymbol{p}^* \in \mathcal{X}^*$ *is 0-coercive.*

iv. $\operatorname{dom} F^* = \mathcal{X}$ *iff F^* is 1-coercive.*

v. In particular, F is 1-coercive $\Rightarrow \mathbf{0} \in \operatorname{int}(\operatorname{dom} F^*)$.

Proof: (E.1.3.10, Hiriart-Urruty and Lemarechal, 2001).

The 1-coercive property corresponds to a stronger version of level-boundedness. In other words F is coercive if it remains level-bounded even after subtracting any linear function (ii above). Note also that if one first has information about a conjugate function, F^*, and F is assumed to be ccp, then its level-boundedness and coercivity can be inferred directly from F^* using Fact 2.16.

2.1.7 Abstract Minimization

Abstract minimization refers to the case where F is unconstrained, or when constraints are implicit in the definition of $\operatorname{dom} F$. In real analysis, students learn that a continuous function achieves its extreme values on a compact set. This is known as the Weierstrass theorem (see *e.g.* Chap. 2, Theorem 1, Luenberger, 1969). In many optimization problems, compactness of the constraint set cannot be assumed, yet we still want guarantees that optimal solutions exist. Convexity and level-boundedness provide a clean answer to the problem of abstract minimization, with level-boundedness playing a role analogous to compactness in the Weierstrass theorem.

Theorem 2.17 (Abstract minimization) *Let F be a ccp and level-bounded.*

i. The set $\operatorname{argmin} F$ *is non-empty,* i.e., *F attains its infimum. It is also compact and convex.*

ii. If $\bar{\boldsymbol{p}} \in \operatorname{argmin} F \Leftrightarrow \mathbf{0} \in \partial F(\bar{\boldsymbol{p}})$.

Proof Let $\bar{\alpha} = \inf F$. (i) Since F is proper $-\infty < \bar{\alpha} < +\infty$. Also because F is level-bounded, $\forall \alpha > \bar{\alpha}$, $\operatorname{lev}_\alpha F$ is non-empty. This implies attainment. Since each $\operatorname{lev}_{\alpha_n} F$ is compact and non-empty, and intersection preserves closedness, compactness and convexity, $\operatorname{argmin} F$ is also shares those properties.

For (ii), suppose $\bar{\boldsymbol{p}} \in \operatorname{argmin} F$ and $\bar{\alpha} = F(\bar{\boldsymbol{p}})$. Then

$$\partial F(\bar{\boldsymbol{p}}) = \{\boldsymbol{p}^* \in \mathcal{X}^* \mid F(\boldsymbol{q}) \geq \langle \boldsymbol{p}^*, \boldsymbol{q} - \bar{\boldsymbol{p}} \rangle + F(\bar{\boldsymbol{p}}), \forall \boldsymbol{q} \in \mathcal{X}\}.$$

Also $\boldsymbol{p}^* \in \partial F^*(\bar{\boldsymbol{p}})$, since the defining condition for the set reduces to $F(\boldsymbol{q}) \geq \bar{\alpha}, \forall \boldsymbol{q} \in \mathcal{X}$. For the reverse implication, simply note that $F(\boldsymbol{q}) \geq F(\bar{\boldsymbol{p}}), \forall \boldsymbol{q} \in \mathcal{X}$, is the definition of a global minimum. ■

2.1.8 The Legendre Property

Two properties: essential smoothness and essential strict convexity are also related under conjugation.

A ccp function F is *essentially smooth* if $\text{int}(\text{dom}\, F)$ is non-empty, ∂F is a singleton on that set and empty outside it. In other words the function is differentiable on $\text{int}(\text{dom}\, F)$.

A ccp function F is *essentially strictly convex* if it is strictly convex on every convex subset of $\text{dom}\, \partial F$.

Functions which possess both properties are said to have the Legendre property. Functions with this property form an especially convenient class to work with.

Fact 2.18 (Legendre property under conjugation) *A proper, convex function F is essentially smooth iff F^* is essentially strictly convex. Therefore F is Legendre iff F^* is Legendre.*
Proof: (11.13, Rockafellar and Wets, 1998) or (Thm. 26.3 Rockafellar, 1970).

Fact 2.19 (Conjugate formula for Legendre function) *If F is ccp and Legendre, then F^* is given by the formula:*

$$F^*(x^*) = \left\langle x^*, (\nabla F)^{-1}(x^*) \right\rangle - F((\nabla F)^{-1}(x^*)). \tag{2.5}$$

Proof: See (11.9 Rockafellar and Wets, 1998).

The foregoing is a kind of best case. The requirement of essential smoothness will not be satisfied by some of the functions looked at later in this work. Nevertheless the formula of Fact 2.19 applies with some extra care. A useful relaxation is to assume that F is not essentially smooth. Instead:

Fact 2.20 (Formula for non-Legendre functions) *Assume F is differentiable on $\mathcal{C} = \text{int}(\text{dom}\, F)$ and $\partial F(\boldsymbol{p}) \neq \emptyset$ for some $\boldsymbol{p} \in \text{int}(\text{dom}\, F)$. Let $\mathcal{D} = \nabla F(\mathcal{C})$. Then for $\boldsymbol{p}^* \in \mathcal{D}$, $F^*(\boldsymbol{p})$ is given by Fact 2.19. Proof: (Thm. 26.4 Rockafellar, 1970).*

The value of F^* outside of $\mathcal{D}$ requires function-specific analysis.

2.1.9 Operations and Conjugation

In this section a number of important types of convex functions are presented. In addition, there are many operations which preserve convexity. Most treatments of convex functions introduce these operations right away. By waiting until after Fenchel conjugation has been discussed another perspective can be given. A number of convexity-preserving operations appear as duals to each other under conjugation, albeit sometimes with the addition of some technical assumptions to maintain closure.

Dual operations

Addition and *epi-addition*:

1. Addition: $F_1 + F_2$ is simply pointwise addition of two functions.

2. Epi-addition[1]: If F_1 and F_2 are ccp, then their *epi-sum* denoted by the expression $F_1 \uplus F_2$, is defined by

$$(F_1 \uplus F_2)(\boldsymbol{p}) = \inf_{\boldsymbol{q}}\{F_1(\boldsymbol{p}) + F_2(\boldsymbol{p}-\boldsymbol{q})\}$$

Fact 2.21 (Sum/Epi-sum) *If F_1 and F_2 are ccp, and $\mathbf{0} \in \operatorname{int}(\operatorname{dom} F_1 - \operatorname{dom} F_2)$ then*

$$(F_1 + F_2)^* = F_1^* \uplus F_2^* \text{ and}$$
$$(F_1 \uplus F_2)^* = F_1^* + F_2^* .$$

The assumption of this Fact can be replaced with either

$$\operatorname{dom} F_1 \cap \operatorname{int}(\operatorname{dom} F_2) \neq \emptyset \text{ or } \operatorname{int}(\operatorname{dom} F_1) \cap \operatorname{dom} F_2 \neq \emptyset.$$

Proof: (11.23(a), Rockafellar and Wets, 1998)

In addition to $\mathcal{X}$ some functions involve another Euclidean space $\mathcal{M} = \mathbb{R}^M$. A typical element of this space will be denoted $\boldsymbol{b}$ or $\boldsymbol{y}$. A connection between $\mathcal{X}$ and $\mathcal{M}$ is made via the linear map $\boldsymbol{A} : \mathcal{X} \to \mathcal{M}$. Usually this map is assumed to be onto, or full row rank when represented by a matrix. The variables $\boldsymbol{u}$ and $\boldsymbol{u}^*$ are used to denote elements of $\mathcal{M}$ and $\mathcal{M}^*$. When the star notation becomes tedious, we drop it and switch to Greek letters, e.g. converting $\boldsymbol{u}^*$ to $\boldsymbol{\mu} \in \mathcal{M}^*$.

Composition and epi-composition with a linear map:
Let $\boldsymbol{A} : \mathcal{X} \to \mathcal{M}$ and $F : \mathcal{X} \to \bar{\mathbb{R}}$.

1. Composition: The composition $F \circ \boldsymbol{A}$ is given by

$$(F \circ \boldsymbol{A})(\boldsymbol{p}) = F(\boldsymbol{A}\boldsymbol{p}).$$

2. Image function[2].The image function, denoted $\boldsymbol{A}F : \mathcal{M} \to \bar{\mathbb{R}}$, is defined by

$$\boldsymbol{A}F(\boldsymbol{y}) = \inf_{\boldsymbol{p}}\{F(\boldsymbol{p}) + \delta_0(\boldsymbol{A}\boldsymbol{p} - \boldsymbol{y})\}.$$

[1] Also called inf-convolution. The name used here is epi-addition, since it corresponds to the Minkowski sum of two epigraphs.

[2] Also called the epi-composition of $\boldsymbol{A}$ and F.

Note that the image function $\boldsymbol{A}F$ has the same target space as F, while changing the argument from $\boldsymbol{p}$ to $\boldsymbol{y}$. The value assigned to $\boldsymbol{y}$ under $\boldsymbol{A}F$ is the same one achieved by a minimizer of F subject to $\boldsymbol{A}\boldsymbol{p} = \boldsymbol{y}$. This function is closely connected to the classic maxent problem taken up in Section 3.5, which requires the same calculation.

Fact 2.22 (Composition/Image function) *Assume that* $\operatorname{range} \boldsymbol{A} \cap \operatorname{int}(\operatorname{dom} F)$, *then*

$$(F \circ \boldsymbol{A})^* = \boldsymbol{A}^* F^* \text{ and}$$
$$(\boldsymbol{A}F)^* = F^* \circ \boldsymbol{A}^*.$$

Proof: (11.23(b), Rockafellar and Wets, 1998).

The next pair of operations are useful in establishing a symmetric form of duality for optimization problems discussed in the next chapter.

Zero-restriction/inf-projection: Consider a function $F : \mathcal{X} \times \mathcal{M} \to \bar{\mathbb{R}}$, which is ccp in the variable $(\boldsymbol{p}, \boldsymbol{u})$, *i.e.,* the function is jointly convex in its two variables.

1. Inf-projection. The inf-projection of F is $G(\boldsymbol{u}) = \inf_{\boldsymbol{p}} F(\boldsymbol{p}, \boldsymbol{u})$.

2. Zero-restriction. The zero-restriction of F is $H(\boldsymbol{p}) = F(\boldsymbol{p}, \boldsymbol{0})$.

Fact 2.23 (Zero-restriction/Inf-projection) *If F is proper, then its inf-projection G is convex and*

$$G^*(\boldsymbol{u}^*) = F^*(\boldsymbol{0}, \boldsymbol{u}^*).$$

If F is ccp and there exists $\boldsymbol{p}$ such that $(\boldsymbol{p}, \boldsymbol{0}) \in \operatorname{int}(\operatorname{dom} F)$ then its zero restriction H is ccp and

$$H^*(\boldsymbol{p}^*) = \inf_{\boldsymbol{u}^*} F^*(\boldsymbol{p}^*, \boldsymbol{u}^*).$$

Proof: (11.23(c), Rockafellar and Wets, 1998).

The above Fact simplifies the general situation somewhat. The two operations are clearly dual to each other. However conjugating a zero restriction requires a closure operation that can be ignored when the sufficient condition stated is in force. In fact the 'zero' in zero restriction can be generalized as well, but it will not be pursued here.

Indicator and Support Functions

The *indicator function* of a convex set $\mathcal{C}$ in $\mathcal{X}$ is defined as

$$\delta_{\mathcal{C}}(\boldsymbol{p}) := \begin{cases} 0, & \boldsymbol{p} \in \mathcal{C} \\ \infty, & \boldsymbol{p} \notin \mathcal{C} \end{cases}, \tag{2.6}$$

while its conjugate, known as the *support function*, is

$$\delta^*_{\mathcal{C}}(\boldsymbol{p}^*) := \sup_{\boldsymbol{p} \in \mathcal{C}} \langle \boldsymbol{p}, \boldsymbol{p}^* \rangle \,. \tag{2.7}$$

Often the function symbol σ is used instead of δ^*. When $\mathcal{C}$ consists of a single point $\boldsymbol{c}$, $\delta^*_{\mathcal{C}}(\boldsymbol{p}^*) = \langle \boldsymbol{c}, \boldsymbol{p}^* \rangle$. Indicators of certain types of sets arise frequently and their support functions are worth remembering. We have already seen one support function. Using Fact 2.6, the directional derivative is

$$\begin{aligned} F'(\boldsymbol{p}; \boldsymbol{d}) &= \sup_{\boldsymbol{p}^* \in \partial F(\boldsymbol{p})} \langle \boldsymbol{p}^*, \boldsymbol{d} \rangle \\ &= \sigma_{\partial F(\boldsymbol{p})}(\boldsymbol{d}), \end{aligned}$$

which is the support function of the set $\partial F(\boldsymbol{p})$ evaluated at the direction $\boldsymbol{d}$.

Define the epsilon P-norm ball as the set

$$\epsilon \mathcal{B}_P := \{\boldsymbol{p} \mid \, ||\boldsymbol{p}||_P \leq \epsilon\}, (1 \leq P \leq \infty).$$

.

Theorem 2.24 (P-norm ball support function) *The indicator of $\epsilon \mathcal{B}_P$ has conjugate*[3]

$$\delta^*_{\mathcal{C}}(\boldsymbol{p}^*) = \epsilon \, ||\boldsymbol{p}^*||_Q \,,$$

where $\frac{1}{Q} + \frac{1}{P} = 1$.

Proof Let Q be such that $1/P + 1/Q = 1$. Substitution and algebra yields the

[3] P and Q are used to avoid confusion with $\boldsymbol{p}$ and $\boldsymbol{q}$, which are vectors.

next three lines:

$$\begin{aligned}
\delta^*_{\epsilon\mathcal{B}_P}(\boldsymbol{p}^*) &= \sup_{\boldsymbol{p}\in\mathcal{X}}\{\, \langle \boldsymbol{p}, \boldsymbol{p}^* \rangle - \delta_{\epsilon\mathcal{B}_P}(\boldsymbol{p}) \,\} \\
&= \epsilon \sup_{\boldsymbol{p}\in\mathcal{X}}\{\, \left\langle \frac{\boldsymbol{p}}{||\boldsymbol{p}||_P}, \boldsymbol{p}^* \right\rangle \,\} \\
&= \epsilon \sup_{\boldsymbol{p}\in\mathcal{X}}\{\, \left\langle \frac{\boldsymbol{p}}{||\boldsymbol{p}||_P}, \frac{\boldsymbol{p}^*}{||\boldsymbol{p}^*||_Q} \right\rangle ||\boldsymbol{p}^*||_Q \,\} \\
&= \epsilon \left|\left| \frac{\boldsymbol{p}}{||\boldsymbol{p}||_P} \right|\right|_P \cdot \left|\left| \frac{\boldsymbol{p}^*}{||\boldsymbol{p}^*||_Q} \right|\right|_Q \cdot ||\boldsymbol{p}^*||_Q \\
&= \epsilon\, ||\boldsymbol{p}^*||_Q \quad .
\end{aligned}$$

On the fourth line the equality condition of the Hölder Inequality is used. ■

In finite dimensions we can also use the pairing of $P = 1$ and $Q = +\infty$ (corresponds to the 1-norm and sup-norm). These are dual to each other in finite dimensions, but not infinite dimensions. This shows that norms are the support function of a convex set, namely the epsilon-ball under the dual norm.

Additively Separable Functions

Some optimization problems have additively separable objective functions. In other words, they can be expressed as $F(\boldsymbol{p}) = \sum_n f_n(p_n)$, for $f_n : \mathbb{R} \to \bar{\mathbb{R}}$, sometimes with $f_n = f$. It will be useful to have a notation that represents element-wise application of the scalar function to a vector. We use significant square brackets, *e.g.* $f[\boldsymbol{p}]$, to represent the operation $\boldsymbol{p} \mapsto (f(p_1), \ldots, f(p_n))$. Then the expression $\mathbf{1}^{\mathrm{T}} f[\boldsymbol{p}]$ represents the sum ($\mathbf{1}$ is a conformable-length vector of ones). Many of the important properties of these vector valued functions are preserved under the sum, and it is often simpler to verify them for a single term. So when it is appropriate to write $F(\boldsymbol{p}) = \mathbf{1}^{\mathrm{T}} f[\boldsymbol{p}]$, then we can write $\nabla F\,(\boldsymbol{p}) = f'[\boldsymbol{p}]$ and $\nabla^2 F\,(\boldsymbol{p}) = \operatorname{diag} f''[\boldsymbol{p}]$.

Fact 2.25 (Conjugate of additively separable functions) *If* $f_n : \mathbb{R} \to \bar{\mathbb{R}}$ *and*

$$F(\boldsymbol{p}) := \sum_n f_n[\boldsymbol{p}],$$

then

$$F^*[\boldsymbol{p}^*] = \sum_n f_n^*[\boldsymbol{p}^*],$$

where f_n^* is the Fenchel conjugate of f_n.
Proof: (Hiriart-Urruty and Lemarechal, 2001, E.1.3.1(ix)).

2.2 Conjugate Calculus and Examples

Sometimes nothing is more helpful in promoting understanding than a few examples. In this section some helpful rules are summarized in 'cheat sheet' form.

Here are some examples of functions, their conjugates and their respective subdifferentials:

F	∂F	∂F^*	F^*
$\frac{1}{2}p^2$	p	p^*	$\frac{1}{2}(p^*)^2$
$p\log(p)$	$\log(p)+1$	$\exp(p^*-1)$	$\exp(p^*-1)$
$\lvert p\rvert$	$\begin{cases} -1, & p<0 \\ [-1,+1], & p=0 \\ +1, & p>0 \end{cases}$	$\begin{cases} \emptyset, & p^*<-1 \\ 0, & p^*\in[-1,1] \\ \emptyset, & p^*>1 \end{cases}$	$\delta_{[-1,+1]}$

Table 2.1: Some examples of conjugate functions and their subdifferentials. Set notation is dropped wherever a singleton set holds.

Some conjugation tasks can be accomplished using algebraic rules. In many case the rules are simple enough to be implemented in a calculator (see Bauschke and v. Mohrenschildt, 2006; Bauschke and von Mohrenschildt, 1999). Here are some of the more common ones:

$F(\boldsymbol{p})$	$F^*(\boldsymbol{p}^*)$	
$F(\boldsymbol{p})+c$	$F^*(\boldsymbol{p}^*)-c$	$c\in\mathbb{R}$
$F(\boldsymbol{p}+\boldsymbol{c})$	$F^*(\boldsymbol{p}^*)-\langle \boldsymbol{c},\boldsymbol{p}^*\rangle$	$\boldsymbol{c}\in\mathcal{X}$
$\lambda F(\boldsymbol{p})$	$\lambda F^*(\lambda^{-1}\boldsymbol{p}^*)$	$\lambda>0$
$F(\lambda\boldsymbol{p})$	$F^*(\lambda^{-1}\boldsymbol{p}^*)$	$\lambda\neq 0$

Table 2.2: Algebraic operations under conjugation.

Here is an elementary example which illustrates the use of some of the rules.

The object is to find f^* given that f is defined as follows:

$$\begin{aligned}
f(p) &= (p - \frac{1}{2})^2 + \frac{1}{2} = g(p) + \frac{1}{2}, \text{ where} \\
g(p) &= (x - \frac{1}{2})^2 = 2h(p), \text{ where} \\
h(p) &= \frac{1}{2}(x - \frac{1}{2})^2 = k(p - \frac{1}{2}), \text{ where} \\
k(p) &= \frac{1}{2}p^2.
\end{aligned}$$

All of the functions can be conjugated using the facts and rules mentioned so far.

$$\begin{aligned}
f^*(p^*) &= g^*(p^*) - \frac{1}{2} \\
g^*(p^*) &= 2h^*(p^*) \\
h^*(p^*) &= k^*(p^*) + \frac{1}{2}p^* \\
k^*(p^*) &= \frac{1}{2}p^{*2}.
\end{aligned}$$

Substitutions yields:

$$\begin{aligned}
f^*(p^*) &= g^*(p^*) - \frac{1}{2} \\
&= 2h^*(\frac{1}{2}p^*) - \frac{1}{2} \\
&= 2k^*(\frac{1}{2}p^* + \frac{1}{4}p^*) - \frac{1}{2} \\
&= 2\frac{1}{2}(\frac{1}{2}p^*)^2 + \frac{1}{2}p^*) - \frac{1}{2} \\
&= \frac{1}{4}p^{*2} + \frac{1}{2}p^* - \frac{1}{2}.
\end{aligned}$$

Correctness can be verified by checking the gradients of f and f^*.

2.3 Bregman Divergence

An important application of convex functions is to generate generalized distance measures.

Definition 2.26 (Bregman Divergence) *Suppose $F : \mathcal{X} \to (-\infty, \infty]$ is a closed convex function which is differentiable on the interior of its domain. The*

Bregman divergence $\Delta F : \mathcal{X} \times \text{int}(\text{dom}\, F) \to [0, \infty]$ *generated via* F *is*

$$\Delta F(\boldsymbol{p}, \boldsymbol{p}_0) := F(\boldsymbol{p}) - F(\boldsymbol{p}_0) - \langle \nabla F(\boldsymbol{p}_0), \boldsymbol{p} - \boldsymbol{p}_0 \rangle \,. \tag{2.8}$$

The concept of a Bregman divergence is illustrated in Figure 2.7.

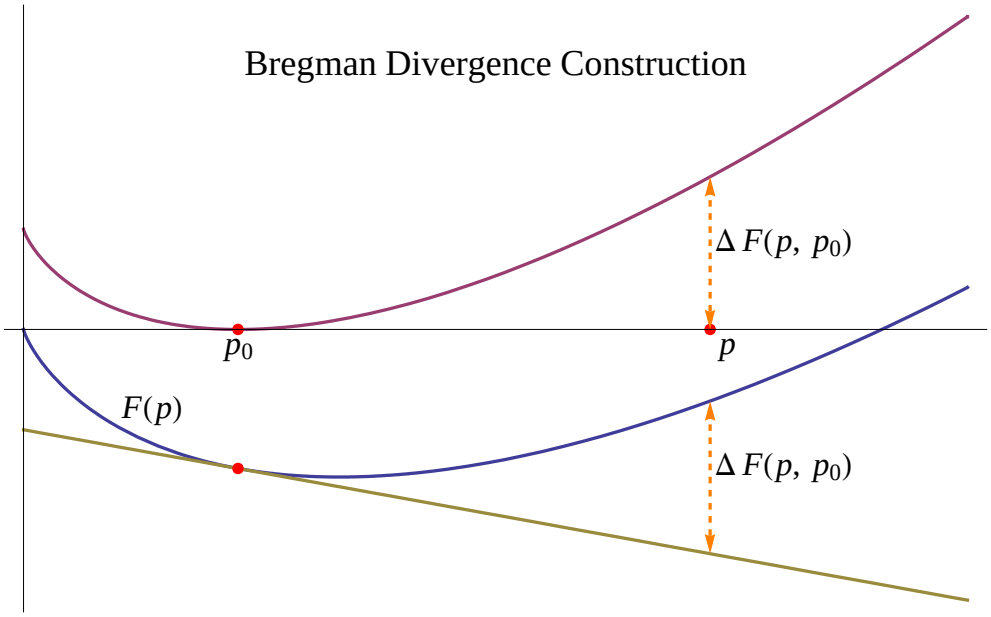


Figure 2.7: Bregman divergence

Bregman divergences are generalized relative entropy functions. To see this let $F(\boldsymbol{p}) = -\sum_n p_n \log(1/p_n)$. If $\boldsymbol{p}$ is an arbitrary probability distribution and $\boldsymbol{q}$ is the uniform distribution, *i.e.,* $q_n = 1/N$, we can calculate:

$$\begin{aligned}
\Delta F(\boldsymbol{p}, \boldsymbol{q}) &= \sum_{n=1}^{N} -p_n \log(1/p_n) + q_n \log(1/q_n) - (1 - \log(1/q_n))(p_n - q_n) \\
&= \sum_{n=1}^{N} -p_n \log(1/p_n) + \sum_{n=1}^{N} p_n \log(N) - \sum_{n=1}^{N} p_n + \sum_{n=1}^{N} \frac{1}{N} \\
&= S(\boldsymbol{p}) + \log(N).
\end{aligned} \tag{2.9}$$

Therefore, with the correct choice of F, SBG entropy is recoverable from Bregman divergence, up to an additive constant.

Bregman divergence is not symmetric in its arguments as illustrated in Figure 2.8. A summary of other relevant properties of Bregman divergences follows. For more details see *e.g.* Azoury and Warmuth (2001); Bauschke and Borwein (1997); Censor and Zenios (1997).

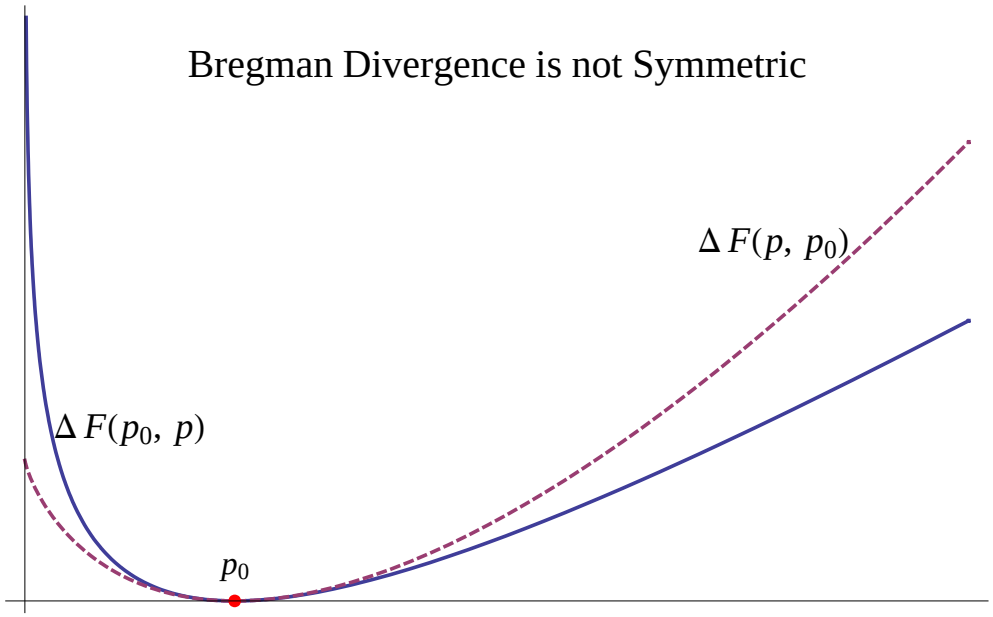


Figure 2.8: Bregman divergence is not symmetric *i.e.*, $\Delta F(p, p_0) \neq \Delta F(p_0, p)$

Lemma 2.27 (Pythagorean Relation) *Suppose* $F : \mathcal{X} \to (-\infty, \infty]$ *is a Legendre convex function,* $\boldsymbol{p}, \boldsymbol{q}, \boldsymbol{p}_0 \in \mathcal{X}$, $\boldsymbol{q}^* = \nabla F(\boldsymbol{q})$, *and* $\boldsymbol{p}_0^* = \nabla F(\boldsymbol{p}_0)$. *Then, the Bregman divergence* (2.8) *satisfies*

$$\Delta F(\boldsymbol{p}, \boldsymbol{p}_0) + \langle \boldsymbol{q}^* - \boldsymbol{p}_0^*, \boldsymbol{q} - \boldsymbol{p} \rangle = \Delta F(\boldsymbol{p}, \boldsymbol{q}) + \Delta F(\boldsymbol{q}, \boldsymbol{p}_0).$$

In particular if $\langle \boldsymbol{q}^* - \boldsymbol{p}_0^*, \boldsymbol{q} - \boldsymbol{p} \rangle = 0$

$$\Delta F(\boldsymbol{p}, \boldsymbol{p}_0) = \Delta F(\boldsymbol{p}, \boldsymbol{q}) + \Delta F(\boldsymbol{q}, \boldsymbol{p}_0). \tag{2.10}$$

Proof

$$\begin{aligned}
\Delta F(\boldsymbol{p}, \boldsymbol{q}) + \Delta F(\boldsymbol{q}, \boldsymbol{p}_0) = & \\
&= F(\boldsymbol{p}) - F(\boldsymbol{q}) + F(\boldsymbol{q}) - F(\boldsymbol{p}_0) - \langle \nabla F(\boldsymbol{q}), \boldsymbol{p} - \boldsymbol{q} \rangle + \langle \nabla F(\boldsymbol{p}_0), \boldsymbol{q} - \boldsymbol{p}_0 \rangle \\
&= \Delta F(\boldsymbol{p}, \boldsymbol{p}_0) + \langle \nabla F(\boldsymbol{p}_0), \boldsymbol{p} - \boldsymbol{p}_0 \rangle - \langle \nabla F(\boldsymbol{q}), \boldsymbol{p} - \boldsymbol{q} \rangle + \langle \nabla F(\boldsymbol{p}_0), \boldsymbol{q} - \boldsymbol{p}_0 \rangle \\
&= \Delta F(\boldsymbol{p}, \boldsymbol{p}_0) + \langle \nabla F(\boldsymbol{p}_0) - \nabla F(\boldsymbol{q}), \boldsymbol{p} - \boldsymbol{q} \rangle \\
&= \Delta F(\boldsymbol{p}, \boldsymbol{p}_0) + \langle \nabla F(\boldsymbol{p}_0) - \nabla F(\boldsymbol{q}), \boldsymbol{p} - \boldsymbol{q} \rangle \\
&= \Delta F(\boldsymbol{p}, \boldsymbol{p}_0) + \langle \boldsymbol{q}^* - \boldsymbol{p}_0^*, \boldsymbol{q} - \boldsymbol{p} \rangle
\end{aligned}$$

The first term is always non-negative, while the second term disappears if the special orthogonality condition of the statement is met. ■

Suppose $F : \mathcal{X} \to (-\infty, \infty]$ is a Legendre convex function. Then, the Bregman

divergence (2.8) generated via F is convex in its first argument—a simple exercise to show. Furthermore, suppose $\boldsymbol{p}, \boldsymbol{q} \in \operatorname{int}(\operatorname{dom} F)$, $\boldsymbol{p}^* = \nabla F(\boldsymbol{p})$, and $\boldsymbol{p}_0^* = \nabla F(\boldsymbol{p}_0)$, then the following facts hold.

Lemma 2.28 (Bregman gradient)

$$\nabla \Delta F(\boldsymbol{p}, \boldsymbol{p}_0) = \boldsymbol{p}^* - \boldsymbol{p}_0^*. \tag{2.11}$$

Proof Simply apply the definition of Bregman divergence, take gradients (whose existence is guaranteed by the Legendre property), and set $\boldsymbol{p}^* = \nabla F(\boldsymbol{p})$ and $\boldsymbol{p}_0^* = \nabla F(\boldsymbol{p}_0)$. ■

Lemma 2.29 (Bregman conjugate) *Suppose F is Legendre. Let $\boldsymbol{p}_0 \in \operatorname{int}(\operatorname{dom} F)$, and let $\boldsymbol{p}_0^* = \nabla F(\boldsymbol{p}_0)$. The (partial) Fenchel-Legendre conjugate of ΔF is:*

$$F^*(\boldsymbol{p}^* + \boldsymbol{p}_0^*) - F^*(\boldsymbol{p}_0^*). \tag{2.12}$$

Proof We will write $\Delta F(\boldsymbol{p}; \boldsymbol{p}_0)$ here with a semicolon, just to emphasize that $\boldsymbol{p}$ is the only argument under consideration at the moment. Using (2.4) and (2.8) we write

$$\begin{aligned}
\Delta F^*(\boldsymbol{p}^*; \boldsymbol{p}_0^*) &= \sup_{\boldsymbol{p} \in \mathcal{X}} \left\{ \langle \boldsymbol{p}, \boldsymbol{p}^* \rangle - F(\boldsymbol{p}) + F(\boldsymbol{p}_0) + \langle \boldsymbol{p} - \boldsymbol{p}_0, \boldsymbol{p}_0^* \rangle \right\} \\
&= \sup_{\boldsymbol{p} \in \mathcal{X}} \left\{ \langle \boldsymbol{p}, \boldsymbol{p}^* + \boldsymbol{p}_0^* \rangle - F(\boldsymbol{p}) \right\} + F(\boldsymbol{p}_0) - \langle \boldsymbol{p}_0, \boldsymbol{p}^* \rangle .
\end{aligned}$$

Since F is assumed to be Legendre, by Fact 2.19 and the definition of the Fenchel-Legendre conjugate (2.4), we can rewrite the above equation as

$$\Delta F^*(\boldsymbol{p}^*; \boldsymbol{p}_0^*) = F^*(\boldsymbol{p}^* + \boldsymbol{p}_0^*) - F^*(\boldsymbol{p}_0^*).$$

■

Note also that $\Delta F(\boldsymbol{p}, \boldsymbol{p}_0)$ is *not* always jointly convex in the argument $(\boldsymbol{p}, \boldsymbol{p}_0)$ or separately convex in $\boldsymbol{p}_0$ with $\boldsymbol{p}$ fixed. Exceptions to this general rule are discussed in Bauschke and Borwein (1997) and Bauschke and Borwein (2000). Our main interest will be in conjugating ΔF with respect to the first argument only, regarding the second argument as a fixed parameter. This function will be denoted as $(\Delta F)^*$. (Note that it is not the same as ΔF^*. This refers to the Bregman divergence generated by the conjugate of F).

Lemma 2.30 (Bregman Flip) *For $\boldsymbol{p}_0 \in \operatorname{int}(\operatorname{dom} F)$:*

$$\Delta F(\boldsymbol{p}, \boldsymbol{p}_0) = \Delta F^*(\boldsymbol{p}_0^*, \boldsymbol{p}^*). \tag{2.13}$$
$$\tag{2.14}$$

Proof Let $\boldsymbol{p}^* = \nabla F(\boldsymbol{p})$ and $\boldsymbol{p}_0^* = \nabla F(\boldsymbol{p}_0)$. The Legendre property ensures that we also have $\boldsymbol{p} = \nabla F^*(\boldsymbol{p}^*)$ and $\boldsymbol{p}_0 = \nabla F^*(\boldsymbol{p}_0^*)$. Two applications of the Fenchel-Young equality allows us to equate

$$F(\boldsymbol{p}) + F^*(\boldsymbol{p}^*) - \langle \boldsymbol{p}^*, \boldsymbol{p} \rangle = 0 = F(\boldsymbol{p}_0) + F^*(\boldsymbol{p}_0^*) - \langle \boldsymbol{p}_0^*, \boldsymbol{p}_0 \rangle$$

Algebraic manipulation yields:

$$\begin{aligned}
F(\boldsymbol{p}) - F(\boldsymbol{p}_0) + \langle \boldsymbol{p}_0^*, \boldsymbol{p}_0 \rangle &= F^*(\boldsymbol{p}_0^*) - F^*(\boldsymbol{p}^*) + \langle \boldsymbol{p}^*, \boldsymbol{p} \rangle \\
F(\boldsymbol{p}) - F(\boldsymbol{p}_0) + \langle \boldsymbol{p}_0^*, \boldsymbol{p}_0 \rangle - \langle \boldsymbol{p}_0^*, \boldsymbol{p} - \boldsymbol{p}_0 \rangle &= F^*(\boldsymbol{p}_0^*) - F^*(\boldsymbol{p}^*) + \langle \boldsymbol{p}^*, \boldsymbol{p} \rangle - \langle \boldsymbol{p}_0^*, \boldsymbol{p} - \boldsymbol{p}_0 \rangle \\
F(\boldsymbol{p}) - F(\boldsymbol{p}_0) + \langle \boldsymbol{p}_0^*, \boldsymbol{p}_0 \rangle - \langle \boldsymbol{p}_0^*, \boldsymbol{p} - \boldsymbol{p}_0 \rangle &= F^*(\boldsymbol{p}_0^*) - F^*(\boldsymbol{p}^*) + \langle \boldsymbol{p}^* - \boldsymbol{p}_0^*, \boldsymbol{p} \rangle + \langle \boldsymbol{p}_0^*, \boldsymbol{p}_0 \rangle \\
\Delta F^*(\boldsymbol{p}_0^*, \boldsymbol{p}^*) &= \Delta F(\boldsymbol{p}, \boldsymbol{p}_0),
\end{aligned}$$

with the last line using the substitutions mentioned at the beginning of the proof. ■

2.3.1 Support function of Bregman divergence sublevel set

We have already seen that the most common distance measure, a norm, is the support function of the unit ball (under the dual norm). The unit ball can be thought of as a sublevel set of the dual norm. As noted, Bregman divergence is one kind of generalized distance measure. The next results indicate what happens under conjugation if the distance measure defining a set is replaced by a Bregman divergence.

Fact 2.31 (Support function of a sublevel set) *If $F : \mathcal{Y} \to \bar{R}$ is ccp, and*

$$\mathcal{C} = \{\boldsymbol{y} \mid F(\boldsymbol{y}) \leq 0\} \text{ and } \inf F < 0,$$

then the conjugate of $\delta_{\mathcal{C}}$ is

$$\delta_{\mathcal{C}}^*(\boldsymbol{y}^*) = \inf_{\lambda > 0} \lambda F^*(\lambda^{-1} \boldsymbol{y}^*), \quad \forall \boldsymbol{y}^* \neq \boldsymbol{0}.$$

Proof: Rockafellar and Wets (1998, 11.6)

This fact is easily applied to the sublevel set of a Bregman divergence.

Theorem 2.32 (Support of Bregman sublevel set) *Let G be a Legendre convex function. If $G : \mathcal{Y} \to \bar{\mathbb{R}}$, $\epsilon > 0$, and $\boldsymbol{y} \in \operatorname{dom} G$ and $\boldsymbol{b} \in \operatorname{int}(\operatorname{dom} G)$, let*

$$\mathcal{C} = \{\boldsymbol{y} \mid \Delta G(\boldsymbol{y}, \boldsymbol{b}) \leq \epsilon\},$$

then the conjugate of $\delta_{\mathcal{C}}$ is given by

$$\delta^*_{\mathcal{C}}(\boldsymbol{y}^*) = \inf_{\lambda > 0} \lambda \left(G^*(\lambda^{-1}\boldsymbol{y}^* + \boldsymbol{b}^*) + \epsilon \right), \quad \forall \boldsymbol{y}^* \neq \boldsymbol{0}. \tag{2.15}$$

Proof To use Fact 2.31, let $F(\cdot) = \Delta G(\cdot, \boldsymbol{b}) - \epsilon$. Since $\epsilon > 0$, clearly $\inf F < 0$. Applying the conjugation formula for Bregman divergence (2.12), plus the conjugation rule for additive constants ($(F-c)^* = F^*+c$) and setting $\boldsymbol{b}^* = \nabla G(\boldsymbol{b})$ gives the result. ■

The requirement that G be Legendre is a sufficient condition to ensure that $\nabla G(\boldsymbol{b})$ can be calculated.

Extended Bregman Divergence

The definition of Bregman divergence given in Section 2.3 can be extended (Warmuth and Vishwanathan, 2006). Recall that the definition there only applies when $\boldsymbol{p}_0 \in \operatorname{int}(\operatorname{dom} F)$. Also the function F is required to be differentiable at $\boldsymbol{p}_0$. Both of these restrictions can be relaxed. Extending the definition to the boundary when subgradients exist there is a natural thing to do and will be useful in later chapters.

Definition 2.33 (Extended Bregman divergence) *Let $F : \mathcal{X} \to \mathbb{R}$ be convex and proper. Then the extended Bregman divergence ΔF is defined by*

$$\Delta F(\boldsymbol{p}, \boldsymbol{p}_0) = F(\boldsymbol{p}) - F(\boldsymbol{p}_0) - F'(\boldsymbol{p}, \boldsymbol{p} - \boldsymbol{p}_0). \tag{2.16}$$

This definition is a natural one if F is convex, but not necessarily differentiable. Recall from Fact 2.6 that

$$F'(\boldsymbol{p}, \boldsymbol{p} - \boldsymbol{p}_0) = \sup_{\boldsymbol{p}^* \in \partial F(\boldsymbol{p}_0)} \langle \boldsymbol{p}^*, \boldsymbol{p} - \boldsymbol{p}_0 \rangle,$$

so that F' is merely picking out the $\boldsymbol{p}^*$ that gives the smallest distance. An alternate expresion for the extended Bregman divergence is

$$\Delta F(\boldsymbol{p}, \boldsymbol{p}_0) = F(\boldsymbol{p}) - F(\boldsymbol{p}_0) - \sigma_{\partial F(\boldsymbol{p}_0)}(\boldsymbol{p} - \boldsymbol{p}_0), \tag{2.17}$$

where σ denotes a support function as usual.

2.3.2 Csiszar's divergence

The emphasis in this thesis will be on Bregman divergences, however there are other measures generalized distance that deserve mention.

It is easy to show that an alternative characterization for KL divergence is based on Csiszar's divergence (Csiszar, 1975). For a convex function $f : \mathcal{X} \to \mathbb{R}$, its Csiszar's divergence is $\Delta_f(\boldsymbol{p}, \boldsymbol{p}_0) = \boldsymbol{p}_0^{\mathrm{T}} f[\frac{\boldsymbol{p}}{\boldsymbol{p}_0}]$, with the division inside taken elementwise. This is just the sum of terms of the form $p_0 f(p/p_0)$. Using this as the basis of a maxent problem leads to consideration of the conjugate term $p_0 f^*(p^*)$ (see Table 2.2).

Still other convex distance-like measures have been proposed. For example Amari's α-divergence is formed as the weighted combination of two Bregamn divergences, with the arguments reversed. This is employed to restore symmetery. These other measures are not pursued further here.

2.4 Conclusion

In this chapter a streamlined version of convex analysis has been presented. I have included it mainly because I could have used it early in my studies! I have found the subject to be compelling because it provides a calculus for manipulating problems into other forms (dual problems), which enables them to be solved, typically on a computer. Although the subject can seem very abstract, its intended use is quite practical. The presentation is simpler than most convex analysis texts, but complete for the purpose at hand. The task is greatly simplified by restricting our attention to concepts appropriate for *convex optimization*. One of the things that makes the subject difficult for the uninitiated is that many concepts in convex analysis are defined in such a way as to make them extendible to more complicated settings. Extensions to non-convex optimization problems are one example and infinite dimensional spaces are another example. In a way this is unfortunate since one of the beauties of the subject is that it is an extension of basic calculus to non-smooth functions, albeit with the restriction of convexity. Extensions beyond this area are the main cause of complications.

As mentioned, many of the results in this chapter can be found in, or distilled from, one or more of the following excellent texts: Rockafellar (1970); Hiriart-Urruty and Lemarechal (2001); Borwein and Lewis (2000); Rockafellar and Wets (1998); Luenberger (1969). I have found all of them helpful in different ways. Most often, this presentation follows the path of Rockafellar and Wets (1998), because it is comprehensive and firmly focused on the finite dimensional case. On the other hand, their presentation is often nuanced and complicated by the authors' desire to extend many concepts to non-convex functions. Some other authors have been cited when their proofs seemed clearer. The presentation here

is meant to be very streamlined for the intended use, but hopefully intuitive as well.

Chapter 3

Duality Concepts for Generalized Maxent

In this chapter, the basic results of convex analysis are extended to cover optimization problems and notions of duality. The results are essentially theoretical and not specifically concerned with maxent problems. The maxent problem remains in the background for the moment, but there are a few good reasons to proceed this way. First the *variational specification* of a probability distribution presented later can be seen as a specialization of the results given here. Second it is easier to link directly to the numerical optimization literature, which is concerned with many other problems besides specifying probability distributions. Finally, it shows how much the definition of an entropy function can be broadened without sacrificing the hope for answers to a maxent problem.

This chapter depends on concepts defined previously. If terms such as gradient, subdifferential, level-bounded coercive, argmin, or the inversion rule for subgradients are not familiar, the reader may wish to return to the previous chapter which presents a concise summary.

3.1 Optimization and Duality Concepts

Duality concepts can be built up from simpler problems to more complex problems. This is done in three stages in this section. First, duality for the abstract minimization problem is established. This makes it clear that duality results call for the use of Fenchel conjugates. Second, duality for the sum of two convex functions is established. This is a classic result that is widely applied, since many optimization problems involve competing goals of one kind or another. Finally the case of linear constraints is examined. In the generalized maxent problem the choice of a measure is governed by the requirement that it satisfies, or nearly satisfies, a set of linear constraints. Duality results for this problem turn out to be a specialization of the previous case involving two convex functions.

3.2 Duality for Abstract Minimization

In the case of abstract minimization we are interested in obtaining the solution to the following problem:

$$\underset{\boldsymbol{p}}{\text{minimize}}\, F(\boldsymbol{p}). \tag{3.1}$$

Under the right conditions, properties of the problem captured by the function F^* are very helpful in characterizing the solution to (3.1).

Theorem 3.1 (Abstract minimization and duality) *If $F : \mathbb{R}^N \to \bar{\mathbb{R}}$ is ccp, then*

i. $\inf F = -F^*(\mathbf{0})$.

ii. $\operatorname{argmin} F = \partial F^*(0)$.

iii. $\operatorname{argmin} F = \{\bar{\boldsymbol{x}}\} \iff$ *F^* is differentiable at 0 and* $\nabla F^*(\mathbf{0}) = \bar{\boldsymbol{x}}$.

Proof

i. From the definition of conjugation we have

$$\begin{aligned} -F^*(\mathbf{0}) &= -\sup_{\boldsymbol{x}} \langle \mathbf{0}, \boldsymbol{x} \rangle - F(\boldsymbol{x}) \\ &= \inf_{\boldsymbol{x}} F(\boldsymbol{x}). \end{aligned}$$

ii. The definition of the subdifferential gives the first line:

$$\begin{aligned} \partial F^*(\mathbf{0}) &= \{\boldsymbol{p} \mid F^*(\boldsymbol{q}) \geq F^*(\mathbf{0}) + \langle \boldsymbol{p}, \boldsymbol{q}^* - \mathbf{0} \rangle, \forall \boldsymbol{q} \in \mathcal{X}^*\} \quad \text{(subgradient definition)} \\ &= \{\boldsymbol{p} \mid F^*(\boldsymbol{q}) - \langle \boldsymbol{p}, \boldsymbol{q}^* \rangle \geq F^*(\mathbf{0}), \forall \boldsymbol{q}^* \in \mathcal{X}^*\} \\ &= \{\boldsymbol{p} \mid F(\boldsymbol{p}) \leq -F^*(\mathbf{0})\} \\ &= \{\boldsymbol{p} \mid F(\boldsymbol{p}) \leq \inf F\} = \operatorname{argmin} F \end{aligned}$$

The third line follows from the definition of the conjugate of F^* and the fact that $F^{**} = F$ (Theorem 2.11). The fourth line follows from (i).

iii. With the previous point in mind the right-hand expression reduces to the fact that $\partial F^*(\mathbf{0})$ is a singleton iff F^* is differentiable.

■

3.3 Symmetric Duality for Perturbed Problems

This section introduces a pair of related optimization problems and discusses dual connections between them. Some of the more remarkable connections come from introducing a perturbed version of the original problem. The most important property of the perturbation scheme is that it respects convexity. This leads to the definition of a *perturbed problem* as one which has a convex objective function $F(\boldsymbol{p}, \boldsymbol{u})$, but where we are primarily interested in solving the primal problem

$$\underset{\boldsymbol{p}}{\text{minimize}}\, F(\boldsymbol{p}, \mathbf{0}). \tag{3.2}$$

This leads to the study of the optimal primal value

$$\inf_{\boldsymbol{p}} F(\boldsymbol{p}, \mathbf{0}).$$

and an associated *value function*:

$$\Pi(\boldsymbol{u}) := \inf_{\boldsymbol{p}} F(\boldsymbol{p}, \boldsymbol{u}).$$

It turns out that the perturbation scheme is mirrored in the problem

$$\underset{\boldsymbol{u}^*}{\text{maximize}} -F^*(\mathbf{0}, \boldsymbol{u}^*). \tag{3.3}$$

For this problem the value of interest is

$$\sup_{\boldsymbol{u}^*} -F^*(\mathbf{0}, \boldsymbol{u}^*) \left(= -\inf_{\boldsymbol{u}^*} F^*(\mathbf{0}, \boldsymbol{u}^*) \right),$$

along with the value function

$$\Delta(\boldsymbol{p}^*) := \inf_{\boldsymbol{u}^*} F^*(\boldsymbol{p}^*, \boldsymbol{u}^*).$$

The optimal value for the primal problem is clearly $\Pi(\mathbf{0})$, while that of the dual problem is -$\Delta(\mathbf{0})$. Note that when F is linear in $\boldsymbol{u}$, for example when $F(\boldsymbol{p}) = G(\boldsymbol{p}) + \langle \boldsymbol{c}, \boldsymbol{u} \rangle$, what we are talking about here is simply the more commonly seen form of Lagrangian duality and Δ^* is called the Lagrangian dual function (see e.g. Section 5.1.6, Boyd and Vandenberghe (2004)). For more details on the connection between Fenchel and Lagrangian duality in this more generalized setting, see Section I, Chapter 11, Rockafellar and Wets (1998).

Using the definitions above the following theorem can be stated:

Theorem 3.2 (Duality for Perturbed Optimization) *Assuming* $F : \mathcal{X} \times \mathcal{M} \to \bar{\mathbb{R}}$ *is ccp, then*

i. $\Pi(0) < +\infty$ *(primal feasibility) iff* $\mathbf{0} \in \operatorname{dom}\Pi$

ii. $-\Delta(\mathbf{0}) > -\infty$ *(dual feasibility) iff* $\mathbf{0} \in \operatorname{dom}\Delta$

iii. The primal problem (3.2) *and dual problem* (3.3) *obey the weak inequality:*

$$\sup_{\boldsymbol{u}^*} -F^*(\mathbf{0}, \boldsymbol{u}^*) \leq \inf_{\boldsymbol{p}} F(\boldsymbol{p}, \mathbf{0}) \tag{3.4}$$

iv. If <u>*either*</u> $\mathbf{0} \in \operatorname{int}(\operatorname{dom}\Pi)$ *or* $\mathbf{0} \in \operatorname{int}(\operatorname{dom}\Delta)$ *then* (3.4) *holds with equality (strong duality).*

v. If $\mathbf{0} \in \operatorname{int}(\operatorname{dom}\Pi)$ *and* $\Delta(\mathbf{0}) > -\infty$ *then* $\operatorname{argmax}_{\boldsymbol{u}} -F^*(\mathbf{0}, \boldsymbol{u}^*) = \partial\Pi(\mathbf{0})$

vi. If $\mathbf{0} \in \operatorname{int}(\operatorname{dom}\Delta)$ *and* $\Pi(\mathbf{0}) < +\infty$ *then* $\operatorname{argmax}_{\boldsymbol{u}} -F^*(\mathbf{0}, \boldsymbol{u}^*) = \partial\Pi(\mathbf{0})$

vii. Optimal solutions are given (in any case) by

$$(\bar{\boldsymbol{p}}, \mathbf{0}) \in \partial F^*(\mathbf{0}, \bar{\boldsymbol{u}}^*) \text{ and } (\mathbf{0}, \bar{\boldsymbol{u}}^*) \in \partial F(\bar{\boldsymbol{p}}, \mathbf{0}) \tag{3.5}$$

Proof The function Π, is an inf-projection, and therefore it is convex and proper when F is ccp (Fact 2.23). The same is true for Δ. This establishes (i) and (ii)

Note that $\operatorname{dom}\Pi = \{\boldsymbol{u} \mid \exists\, \boldsymbol{p} \text{ such that } (\boldsymbol{p}, \boldsymbol{u}) \in \operatorname{dom} F\}$. On this set, F is bounded below. Its infimum may or may not be attained. If attained, it may or may not be unique.

Also by Fact 2.23, we have $\Pi^*(\boldsymbol{u}^*) = F^*(\mathbf{0}, \boldsymbol{u}^*)$. This requires no additional assumptions on F. Conjugating Π^* in turn yields,

$$\begin{aligned} \Pi^{**}(\boldsymbol{u}) &= \sup_{\boldsymbol{u}^*}\{\langle \boldsymbol{u}, \boldsymbol{u}^*\rangle - F^*(\mathbf{0}, \boldsymbol{u}^*)\}, \text{ which implies} \\ \Pi^{**}(\mathbf{0}) &= \sup_{\boldsymbol{u}^*} -F^*(\mathbf{0}, \boldsymbol{u}^*). \end{aligned}$$

Note that we always have

$$\Pi^{**}(\mathbf{0}) \leq \Pi(\mathbf{0}), \text{ and} \tag{3.6}$$

$$-\inf_{\boldsymbol{u}^*} F^*(\mathbf{0}, \boldsymbol{u}^*) \leq \inf_{\boldsymbol{p}} F(\boldsymbol{p}, \mathbf{0}). \tag{3.7}$$

Thus we have established the weak duality result (i). Strong duality ($\Pi^{**}(\mathbf{0}) = \Pi(\mathbf{0})$) holds when Π is subdifferentiable at $\mathbf{0}$, *i.e.,* Π is actually equal to one of its supporting hyperplanes there. A sufficient condition for this is

$$\mathbf{0} \in \operatorname{int}(\operatorname{dom}\Pi) = \{\boldsymbol{u} \mid \exists\, \boldsymbol{p} \text{ such that } (\boldsymbol{p}, \boldsymbol{u}) \in \operatorname{dom} F\} \tag{3.8}$$

This condition is provided in Fact 2.23. Next, consider the dual problem (3.3) along with the value function $\Delta(\boldsymbol{p}^*)$. We are interested in the value

$$-\Delta(\mathbf{0}) = -\inf_{\boldsymbol{u}^*} F^*(\mathbf{0}, \boldsymbol{u}^*) = \sup_{\boldsymbol{u}^*} -F^*(\mathbf{0}, \boldsymbol{u}^*),$$

since this value corresponds to the solution to the problem. The function Δ has[1]

$$\operatorname{dom}\Delta = \{\boldsymbol{p}^* \mid \exists\, \boldsymbol{u}^* \text{ such that } (\boldsymbol{p}^*, \boldsymbol{u}^*) \in \operatorname{dom} F^*\}.$$

Likewise, Δ is a convex, proper function and $\operatorname{dom}\Delta$ is a convex set.

Using the fact $F = F^{**}$, the conjugate of Δ can be derived as follows:

$$\begin{aligned}
\Delta^*(\boldsymbol{p}) &= \sup_{\boldsymbol{p}^*}\{\, \langle \boldsymbol{p}, \boldsymbol{p}^* \rangle - \inf_{\boldsymbol{u}^*}\{F^*(\boldsymbol{p}^*, \boldsymbol{u}^*)\}\,\} \\
\Delta^*(\boldsymbol{p}) &= \sup_{\boldsymbol{p}^*}\{\, \langle \boldsymbol{p}, \boldsymbol{p}^* \rangle + \sup_{\boldsymbol{u}^*}\{-F^*(\boldsymbol{p}^*, \boldsymbol{u}^*)\}\,\} \\
&= \sup_{(\boldsymbol{p}^*, \boldsymbol{u}^*)}\{\, \langle (\boldsymbol{p}, \mathbf{0}), (\boldsymbol{p}^*, \boldsymbol{u}^*) \rangle - F^*(\boldsymbol{p}^*, \boldsymbol{u}^*)\,\} \\
&= F^{**}(\boldsymbol{p}, \mathbf{0}) \\
&= F(\boldsymbol{p}, \mathbf{0}).
\end{aligned}$$

Conjugating one more time gives[2]:

$$\begin{aligned}
\Delta^{**}(\boldsymbol{p}^*) &= \sup_{\boldsymbol{p}}\{\, \langle \boldsymbol{p}^*, \boldsymbol{p} \rangle - F(\boldsymbol{p}, \mathbf{0})\,\}, \text{ which implies} \\
\Delta^{**}(\mathbf{0}) &= \sup_{\boldsymbol{p}} -F(\boldsymbol{p}, \mathbf{0}).
\end{aligned}$$

This time weak duality corresponds to

$$\begin{aligned}
\Delta(\mathbf{0}) &\geq \Delta^{**}(\mathbf{0}) \\
-\sup_{\boldsymbol{u}^*} -F^*(\mathbf{0}, \boldsymbol{u}^*) &\geq \sup_{\boldsymbol{p}^*} -F(\boldsymbol{p}, \mathbf{0}) \\
-\inf_{\boldsymbol{u}^*} F^*(\mathbf{0}, \boldsymbol{u}^*) &\leq \inf_{\boldsymbol{p}} F(\boldsymbol{p}, \mathbf{0}),
\end{aligned} \tag{3.9}$$

with strong duality ($\Delta^{**}(\mathbf{0}) = \Delta(\mathbf{0})$) holding when (again by Fact 2.11)

$$\mathbf{0} \in \operatorname{dom}\Delta = \{\, \boldsymbol{p}^* \mid \exists (\boldsymbol{p}^*, \boldsymbol{u}^*) \in \operatorname{dom} F^*\,\}. \tag{3.10}$$

Note that the duality statements (3.6) and (3.9) are identical. The difference lies in the sufficient conditions for strong duality, (3.8) and (3.10). Either one

[1] Δ has no connection to Bregman divergence here.

[2] Δ^* takes $\boldsymbol{p}$ as an argument while Δ and Δ^{**} take $\boldsymbol{p}^*$ since the dual problem started in the dual space. This represents a rare departure from the usual alignment of stars.

of them is therefore sufficient for equality of optimal primal and dual objective functions. This establishes (ii).

If $\mathbf{0} \in \operatorname{int}(\operatorname{dom}\Delta)$, then it is subdifferentiable at $\mathbf{0}$ (Fact 2.6). Combined with Theorem 3.1(ii) we have

$$\partial\Delta(0) = \operatorname{argmin}\Delta^* = \operatorname{argmin} F(\boldsymbol{p}, \mathbf{0}) = \bar{\boldsymbol{p}}$$

This proves (v). Starting with $\mathbf{0} \in \operatorname{int}(\operatorname{dom}\Pi)$, using the same argument yields

$$\partial\Pi(0) = \operatorname{argmin}\Pi^* = \operatorname{argmin} F^*(\mathbf{0}, \boldsymbol{u}^*) = \bar{\boldsymbol{u}}^*,$$

which establishes (vi).

Suppose now there is a pair of optimal solutions $\bar{\boldsymbol{p}}$ and $\bar{\boldsymbol{u}}^*$. F must be subdifferentiable at $(\bar{\boldsymbol{p}}, \mathbf{0})$ and F^* must be subdifferentiable at $(\mathbf{0}, \bar{\boldsymbol{u}}^*)$. Application of the Fenchel-Young equality (Theorem 2.13) yields

$$\begin{aligned} F(\bar{\boldsymbol{p}}, \mathbf{0}) + F^*(\mathbf{0}, \bar{\boldsymbol{u}}^*) &= \langle(\bar{\boldsymbol{p}}, \mathbf{0}), (\mathbf{0}, \bar{\boldsymbol{u}}^*)\rangle \\ F(\bar{\boldsymbol{p}}, \mathbf{0}) &= -F^*(\mathbf{0}, \bar{\boldsymbol{u}}^*). \end{aligned}$$

This implies that strong duality exists, and also via Fact 2.6 and Fact 2.12 that

$$\begin{aligned} (\bar{\boldsymbol{p}}, \mathbf{0}) &\in \partial F^*(\mathbf{0}, \bar{\boldsymbol{u}}^*) \text{ and} \\ (\mathbf{0}, \bar{\boldsymbol{u}}^*) &\in \partial F(\bar{\boldsymbol{p}}, \mathbf{0}), \end{aligned}$$

which establishes (vii). ■

See also Rockafellar and Wets (1998, Thm. 11.39).

3.4 Duality with Linear Constraints

We have reached the high watermark of abstraction. The next theorem uses Theorem 3.2 in a more specialized setting. It also will be used as the point of departure for generalized maxent problems. After this result most other problems we will look at are specialized further in some way.

Theorem 3.3 (Parameterized Fenchel Duality) *Consider the dual optimization problems*

$$\underset{\boldsymbol{p}\in\mathcal{X}}{\text{minimize}}\, K(\boldsymbol{p}, \mathbf{0}) \qquad \textit{(primal problem)} \tag{3.11}$$

$$\underset{\boldsymbol{u}^*\in\mathcal{M}^*}{\text{maximize}}\, -K^*(\mathbf{0}, \boldsymbol{u}^*) \qquad \textit{(dual problem).} \tag{3.12}$$

where $K : \mathcal{X} \times \mathcal{M} \to (-\infty, \infty]$ *is the perturbed primal objective function defined by*

$$K(\boldsymbol{p}, \boldsymbol{u}) := \langle \boldsymbol{p}, \boldsymbol{c}^* \rangle + H(\boldsymbol{p}) + G(\boldsymbol{u} + \boldsymbol{b} - \boldsymbol{A}\boldsymbol{p}), \tag{3.13}$$

and where $H : \mathcal{X} \to (-\infty, \infty]$ *and* $G : \mathcal{M} \to (-\infty, \infty]$ *are closed and convex, with* $\boldsymbol{A} : \mathcal{X} \to \mathcal{M}$ *linear,* $\boldsymbol{u}, \boldsymbol{b} \in \mathcal{M}$*,* $\boldsymbol{p} \in \mathcal{X}$*, and* $\boldsymbol{c}^* \in \mathcal{X}^*$*. Consider also the following constraint qualification (CQ):*

$$\boldsymbol{b} \in \operatorname{int}(\boldsymbol{A} \operatorname{dom} H + \operatorname{dom} G) \text{ and } \boldsymbol{c}^* \in \operatorname{int}(\boldsymbol{A}^* \operatorname{dom} G^* - \operatorname{dom} H^*). \tag{3.14}$$

The following hold:

i. The dual objective $K^* : \mathcal{X}^* \times \mathcal{M}^* \to (-\infty, \infty]$*, with* $\boldsymbol{p}^* \in \mathcal{X}^*$ *and* $\boldsymbol{u}^* \in \mathcal{M}^*$*, is*

$$K^*(\boldsymbol{0}, \boldsymbol{u}^*) = -\langle \boldsymbol{b}, \boldsymbol{u}^* \rangle + G^*(\boldsymbol{u}^*) + H^*(\boldsymbol{A}^*\boldsymbol{u}^*). \tag{3.15}$$

ii. If CQ (3.14) *holds then strong duality holds:* $\inf_{\boldsymbol{p}} K(\boldsymbol{p}, \boldsymbol{0}) = \sup_{\boldsymbol{u}^*} K(\boldsymbol{0}, \boldsymbol{u}^*)$*.*

iii. Optimal solutions $\bar{\boldsymbol{p}}$ *and* $\bar{\boldsymbol{u}}^*$*, if they exist, satisfy the following relations:*

$$\bar{\boldsymbol{u}}^* \in \partial G(\boldsymbol{b} - \boldsymbol{A}\bar{\boldsymbol{p}}) \tag{3.16}$$

$$\boldsymbol{b} - \boldsymbol{A}\bar{\boldsymbol{p}} \in \partial G^*(\bar{\boldsymbol{u}}^*) \tag{3.17}$$

$$\boldsymbol{A}^*\bar{\boldsymbol{u}}^* - \boldsymbol{c}^* \in \partial H(\bar{\boldsymbol{p}}) \tag{3.18}$$

$$\bar{\boldsymbol{p}} \in \partial H^*(\boldsymbol{A}^*\bar{\boldsymbol{u}}^* - \boldsymbol{c}^*). \tag{3.19}$$

Proof To show (i) conjugate K directly, using $\boldsymbol{v} = \boldsymbol{u} + \boldsymbol{b} - \boldsymbol{A}\boldsymbol{p}$:

$$\begin{aligned}
K^*(\boldsymbol{p}^*, \boldsymbol{u}^*) &= \sup_{(\boldsymbol{p},\boldsymbol{u})} \{ \langle (\boldsymbol{p}, \boldsymbol{u}), (\boldsymbol{p}^*, \boldsymbol{u}^*) \rangle - (\langle \boldsymbol{p}, \boldsymbol{c}^* \rangle + H(\boldsymbol{p}) + G(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{p} + \boldsymbol{u})) \} \\
&= \sup_{(\boldsymbol{p},\boldsymbol{v})} \{ \langle (\boldsymbol{p}, \boldsymbol{v} - \boldsymbol{b} + \boldsymbol{A}\boldsymbol{p}), (\boldsymbol{p}^*, \boldsymbol{u}^*) \rangle - \langle \boldsymbol{p}, \boldsymbol{c}^* \rangle - H(\boldsymbol{p}) - G(\boldsymbol{v}) \} \\
&= \sup_{(\boldsymbol{p},\boldsymbol{v})} \{ \langle \boldsymbol{p}, \boldsymbol{p}^* \rangle + \langle \boldsymbol{v}, \boldsymbol{u}^* \rangle + \langle \boldsymbol{p}, \boldsymbol{A}^*\boldsymbol{u}^* \rangle - \langle \boldsymbol{b}, \boldsymbol{u}^* \rangle - \langle \boldsymbol{p}, \boldsymbol{c}^* \rangle - H(\boldsymbol{p}) - G(\boldsymbol{v}) \} \\
&= \sup_{\boldsymbol{p}} \{ \langle \boldsymbol{p}, \boldsymbol{p}^* + \boldsymbol{A}^*\boldsymbol{u}^* - \boldsymbol{c}^* \rangle - H(\boldsymbol{p}) \} + \sup_{\boldsymbol{v}} \{ \langle \boldsymbol{v}, \boldsymbol{u}^* \rangle - G(\boldsymbol{v}) \} - \langle \boldsymbol{b}, \boldsymbol{u}^* \rangle \\
&= H^*(\boldsymbol{p}^* + \boldsymbol{A}^*\boldsymbol{u}^* - \boldsymbol{c}^*) + G^*(\boldsymbol{u}^*) - \langle \boldsymbol{b}^*, \boldsymbol{u}^* \rangle .
\end{aligned} \tag{3.20}$$

The first equality uses the definition of Fenchel conjugation. The second substitutes for $\boldsymbol{u}$. The third follows from inner product rules, including the fact $\langle \boldsymbol{A}\boldsymbol{p}, \boldsymbol{u}^* \rangle = \langle \boldsymbol{p}, \boldsymbol{A}^*\boldsymbol{u}^* \rangle$. The fourth uses the definition of Fenchel conjugates again and the fifth replaces $\boldsymbol{v}$. Now setting $\boldsymbol{c}^* = \boldsymbol{0}$ yields the dual problem's

objective function (3.15). This demonstrates (i). (It is interesting to note the symmetry here. This last step corresponds to an unperturbed problem analogous to setting $\boldsymbol{u}$ to zero. However the perturbation does not occur in the constraint space, instead it occurs in the dual of the primal space.)

To apply Theorem 3.2 it is necessary to identify the domain of the primal value function:

$$\begin{aligned}\operatorname{dom}\Pi &= \{\, \boldsymbol{u} \mid \exists \boldsymbol{p} \text{ such that } (\boldsymbol{p}, \boldsymbol{u}) \in \operatorname{dom} K \,\} \\ &= \{\, \boldsymbol{u} \mid \exists \boldsymbol{p} \text{ such that } K(\boldsymbol{p}, \boldsymbol{u}) < +\infty \,\}.\end{aligned}$$

The first term of K (3.13) is finite, therefore the condition $\mathbf{0} \in \operatorname{dom}\Pi$ equates to

$$\exists \boldsymbol{p} \in \operatorname{dom} H \text{ such that } \boldsymbol{b} - \boldsymbol{A}\boldsymbol{p} \in \operatorname{dom} G \Leftrightarrow \qquad \boldsymbol{b} \in \boldsymbol{A} \operatorname{dom} H + \operatorname{dom} G.$$

Likewise, using (3.20), the condition for the dual problem, $\mathbf{0} \in \operatorname{dom}\Delta$, boils down to

$$\exists \boldsymbol{u}^* \text{ such that } H^*(\boldsymbol{A}^*\boldsymbol{u}^* - \boldsymbol{c}^*) + G^*(\boldsymbol{u}^*) < +\infty \Leftrightarrow \boldsymbol{c}^* \in \boldsymbol{A}^* \operatorname{dom} H^* + \operatorname{dom} G^*.$$

Free placement of the int sign in front follows from Fact 2.1 and this establishes the sufficient condition (ii).

Recall for are problem of interest, the constraints are unperturbed. This corresponds to $\boldsymbol{u} = \mathbf{0}$. Now suppose there are optimal solutions $\bar{\boldsymbol{p}}$ and $\bar{\boldsymbol{u}}^*$. They therefore obey the optimality subgradient relation (3.5) in Theorem 3.2. Calculating with the subgradient chain rule[3] (Fact 2.8),

$$\begin{aligned}\partial K(\bar{\boldsymbol{p}}, \mathbf{0}) &= \partial \{\langle (\bar{\boldsymbol{p}}, \mathbf{0}), (\boldsymbol{c}^*, \mathbf{0}) \rangle + H(\bar{\boldsymbol{p}}) + G(\boldsymbol{b} - \boldsymbol{A}\bar{\boldsymbol{p}})\} \\ &= (\boldsymbol{c}^*, \mathbf{0}) + \partial H(\bar{\boldsymbol{p}}) \times \{\mathbf{0}\} + (-\boldsymbol{A}^* \partial G(\boldsymbol{b} - \boldsymbol{A}\bar{\boldsymbol{p}})) \times \partial G(\boldsymbol{b} - \boldsymbol{A}\bar{\boldsymbol{p}}).\end{aligned}$$

The optimality condition $(\mathbf{0}, \bar{\boldsymbol{u}}^*) \in \partial K(\bar{\boldsymbol{p}}, \mathbf{0})$ then becomes the pair of relations:

$$\begin{aligned}\mathbf{0} &\in \boldsymbol{c}^* + \partial H(\bar{\boldsymbol{p}}) - \boldsymbol{A}^* \partial G(\boldsymbol{b} - \boldsymbol{A}\bar{\boldsymbol{p}}) \\ \bar{\boldsymbol{u}}^* &\in \partial G(\boldsymbol{b} - \boldsymbol{A}\bar{\boldsymbol{p}}),\end{aligned}$$

which are equivalent to

$$\begin{aligned}\boldsymbol{A}^*\bar{\boldsymbol{u}}^* - \boldsymbol{c}^* &\in \partial H(\bar{\boldsymbol{p}}) \\ \bar{\boldsymbol{u}}^* &\in \partial G(\boldsymbol{b} - \boldsymbol{A}\bar{\boldsymbol{p}}),\end{aligned}$$

[3]Recall + and × denotes Minkowski sum and Cartesian product here!

as required. Subgradient inversion yields

$$\begin{aligned} \bar{\boldsymbol{p}} &\in \partial H^*(\boldsymbol{A}^*\bar{\boldsymbol{u}}^* - \boldsymbol{c}^*) \\ \boldsymbol{b} - \boldsymbol{A}\bar{\boldsymbol{p}} &\in \partial G^*(\bar{\boldsymbol{u}}^*). \end{aligned}$$

This completes the proof of (iii). ■

3.5 The Classic Maxent Problem

Now we turn to look at the classic maxent problem and some generalizations. Most of the results are obtained by specializing the results of the previous section. This allows us to see the maxent problem(s) in the broader context of specifiying a probability distribution using variational methods.

The tools of convex analysis are especially useful for characterizing distributions without first making strong assumptions about the entropy function employed. For example we do not assume the Legendre property up front. Rather we see how it specializes the optimality conditions to be analytically nice. This 'niceness' is neither required nor desirable in all applications, so it is helpful to see it in the broader context.

From this point we consider only Bregman divergences generated by F with $\operatorname{dom} F = \mathbb{R}^N_+$, setting F equal to $+\infty$ outside the non-negative cone. In this way we can drop explicit non-negativity constraints. Bregman divergences constructed with this requirement are suitable for probability distributions, in contrast to some measures studied elsewhere (*e.g.* Banerjee et al., 2005), which are suitable for other objects, such as sufficient statistics.

Maximum Entropy (maxent) is a well studied problem (Jaynes, 1957b, 1982b). The classic version in finite dimensional Euclidean space corresponds to the problem:

$$\begin{aligned} &\text{minimize} && S(\boldsymbol{p}) \\ &\text{subject to} && \mathbf{1}^{\mathrm{T}}\boldsymbol{p} = 1 \text{ and} \\ &&& \boldsymbol{A}\boldsymbol{p} = \boldsymbol{b} \\ &&& p_n \geq 0 \text{ for } n = 1, \ldots, N, \text{ with } \boldsymbol{p} := (p_1, \ldots, p_N). \end{aligned} \tag{3.21}$$

Here $S : \mathbb{R}^N \to \mathbb{R}$ is the negative Shannon-Boltzmann-Gibbs (SBG) entropy

$$S(\boldsymbol{p}) := -\sum_{n=1}^{N} p_n \log\left(\frac{1}{p_n}\right),$$

$\mathbf{1}$ denotes the vector of all ones, matrix $\boldsymbol{A} \in \mathbb{R}^{M\times N}$ and vector $\boldsymbol{b} \in \mathbb{R}^M$, with $M+1 \leq N$. The vector $\boldsymbol{b}$ is often referred to as the *mean vector* or the *data*, which is suggestive of its role in applications. When convenient, we will think of normalization as just another constraint and write

$$\boldsymbol{B} = \begin{pmatrix} \mathbf{1}^{\mathrm{T}} \\ \boldsymbol{A} \end{pmatrix} \text{ and } \boldsymbol{b_1} = \begin{pmatrix} 1 \\ \boldsymbol{b} \end{pmatrix}, \tag{3.22}$$

so that the combined constraint becomes $\boldsymbol{Bp} = \boldsymbol{b_1}$.

We also assume that the problem is feasible in the following sense: there exists some probability distribution $\boldsymbol{p_b}$ such that $\boldsymbol{Ap_b} = \boldsymbol{b}$. This ensures that the objective function is finite on at least one point in the feasible set and hence the optimization problem is non-trivial.

A number of variations of the problem (3.21) have appeared over time, sometimes in a form where the classic problem can be recovered as a special case. In this thesis we study the following generalization

$$\min_{\boldsymbol{p}\in\mathbb{R}^N} \Delta F(\boldsymbol{p}, \boldsymbol{p}_0) \text{ subject to } \boldsymbol{Ap} = \boldsymbol{b} \quad \mathbf{1}^{\mathrm{T}}\boldsymbol{p} = 1 \text{ and } p_n \geq 0 \text{ for } n = 1, \ldots, N. \tag{3.23}$$

Here $F : \mathbb{R}^N \to (-\infty, \infty]$ is a convex function, and the Bregman divergence $\Delta F(\boldsymbol{p}, \boldsymbol{p}_0) := F(\boldsymbol{p}) - F(\boldsymbol{p}_0) - \langle \nabla F(\boldsymbol{p}_0), \boldsymbol{p} - \boldsymbol{p}_0 \rangle$ behaves like a negative relative entropy with respect to the initial guess $\boldsymbol{p}_0$ (see the next section).

Our setting is not the most general maxent problem. The focus work with finite dimensional Euclidean spaces, although much of the analysis carries over to the case where $\boldsymbol{p}$ is an element of a Banach space (*e.g.* Borwein and Zhu (2005); Altun and Smola (2006)). Furthermore, we do not generalize the form of the constraints. This is an active line of research in learning theory. For example, Dudík and Schapire (2006), and Altun and Smola (2006) investigate norm bounded relaxations of the constraints to address over-fitting encountered in statistical settings. In Altun and Smola (2006), a theoretical treatment of both problems is pursued in infinite dimensional Banach spaces. These papers address performance guarantees useful in a statistical setting. Here, our aim is to establish connections to research in other fields. In the cases at hand, exact constraint matching is all that is needed to point out existing connections. In this way we provide a starting point for extensions to relaxed constraints and infinite-dimensional state spaces.

3.6 A Generalization of Maxent

Convex analysis permits the description of a constraint set using an indicator function. By combining the objective function with such an indicator function, we arrive at a formal statement of the problem that is 'unconstrained'. Fenchel duality is an elegant tool used to derive a dual version of this problem. The dual problem which is constructed out of Fenchel-Legendre conjugates of the objective and constraint function, may be easier to solve, or at least aid in finding a solution.

Theorem 3.3 is broad enough to cover many problems, besides generalized maxent problems. In our setting it specializes as follows (See also Bauschke and Borwein (1997, Theorem 3.12) or Borwein and Lewis (2000, Corollary 3.4)):

Corollary 3.4 *Let $F : \mathcal{X} \to (-\infty, \infty]$ be Legendre, $\boldsymbol{p}_0 \in \operatorname{int}(\operatorname{dom} F)$, $\boldsymbol{A} : \mathcal{X} \to \mathcal{M}$ be a full rank linear map such that $\boldsymbol{A}\boldsymbol{p} = \boldsymbol{b}$ for $\boldsymbol{b} \in \mathcal{M}$ and some $\boldsymbol{p} \in \operatorname{int}(\operatorname{dom} F)$. Then the problem*

$$\min_{\boldsymbol{p}\in\mathcal{X}} \Delta F(\boldsymbol{p}, \boldsymbol{p}_0) \text{ subject to } \boldsymbol{A}\boldsymbol{p} = \boldsymbol{b} \tag{3.24}$$

has a dual formulation

$$\max_{\boldsymbol{u}^*\in\mathcal{M}^*} \langle \boldsymbol{b}, \boldsymbol{u}^* \rangle - F^*(\boldsymbol{A}^*\boldsymbol{u}^* + \boldsymbol{p}_0^*) + F^*(\boldsymbol{p}_0^*), \tag{3.25}$$

where $\boldsymbol{p}_0^* = \nabla F(\boldsymbol{p}_0) \in \operatorname{int}(\operatorname{dom} F^*)$. Given a solution to (3.25), $\bar{\boldsymbol{u}}^*$, we can recover the primal solution, $\bar{\boldsymbol{p}}$, via

$$\bar{\boldsymbol{p}} = \nabla F^*(\boldsymbol{A}^*\bar{\boldsymbol{u}}^* + \boldsymbol{p}_0^*). \tag{3.26}$$

Furthermore, $\bar{\boldsymbol{p}}$ and $\bar{\boldsymbol{u}}^*$ are unique.

Proof In order to apply Theorem 3.3 we first need to identify the primal objective function (3.24) with (3.13). This is easily achieved by setting $\boldsymbol{c}^* = \boldsymbol{0}$, $H(\cdot) = \Delta F(\cdot, \boldsymbol{p}_0)$, and G to indicate the set which matches the (perturbed) constraints, $G(\cdot) = \delta_{\boldsymbol{0}}(\cdot)$. Next, we need to ensure that the constraint qualifications (3.14) are met. To this end, observe that since $\operatorname{dom} G = \{\boldsymbol{0}\}$, then $\operatorname{int}(\boldsymbol{A}\operatorname{dom} H + \operatorname{dom} G) = \operatorname{int}(\boldsymbol{A}\operatorname{dom} F)$. Also, since $\boldsymbol{A}$ is full rank, the map preserves interiors, so we can rewrite $\operatorname{int}(\boldsymbol{A}\operatorname{dom} F)$ as $\boldsymbol{A}\operatorname{int}(\operatorname{dom} F)$. In other words,

$$\boldsymbol{b} \in \operatorname{int}(\boldsymbol{A}\operatorname{dom} H + \operatorname{dom} G) \iff \exists\, \boldsymbol{p} \in \operatorname{int}(\operatorname{dom} F) \text{ such that } \boldsymbol{A}\boldsymbol{p} = \boldsymbol{b},$$

which holds by assumption.

To write the second constraint qualification we need expressions for G^* and H^*. First note that $G^*(\cdot) = 0$, and therefore $\operatorname{dom} G^* = \mathcal{M}^*$, and $\boldsymbol{A}^* \operatorname{dom} G^* =$

range $\boldsymbol{A}^*$. Next, using (2.12) we can write

$$H^*(\cdot) = F^*(\cdot + \boldsymbol{p}_0^*) - F^*(\boldsymbol{p}_0^*), \tag{3.27}$$

which implies that $\operatorname{dom} H^* = \operatorname{dom} F^* - \boldsymbol{p}_0^*$. Therefore, the constraint qualification boils down to

$$\mathbf{0} \in \operatorname{int}(\operatorname{range} \boldsymbol{A}^* + \operatorname{dom} F^* - \boldsymbol{p}_0^*).$$

Since $\boldsymbol{p}_0 \in \operatorname{int}(\operatorname{dom} F)$ by assumption, then $\boldsymbol{p}_0^* \in \operatorname{int}(\operatorname{dom} F^*)$ by Fact 2.19. This implies $\mathbf{0} \in \operatorname{int}(\operatorname{dom} F^* - \boldsymbol{p}_0^*)$. We also have $\mathbf{0} \in \operatorname{range} \boldsymbol{A}^*$, which implies the relation $\operatorname{dom} F^* - \boldsymbol{p}_0^* \subseteq \operatorname{range} \boldsymbol{A}^* + \operatorname{dom} F^* - \boldsymbol{p}_0^*$. Since int is monotone with respect to $\subseteq$, it is true that $\operatorname{int}(\operatorname{dom} F^* - \boldsymbol{p}_0^*) \subseteq \operatorname{int}(\operatorname{range} \boldsymbol{A}^* + \operatorname{dom} F^* - \boldsymbol{p}_0^*)$. But we already saw that $\mathbf{0}$ is in the left-hand set, so it must also be in the right hand set, and therefore the second constraint qualification always holds.

By substituting $G^* = 0$, H^*, from (3.27), and $\boldsymbol{c}^* = \mathbf{0}$ into (3.15), we can write the dual objective function as

$$\Psi(\boldsymbol{u}^*) = -k^*(\mathbf{0}, \boldsymbol{u}^*) = \langle \boldsymbol{b}, \boldsymbol{u}^* \rangle - F^*(\boldsymbol{A}^* \boldsymbol{u}^* + \boldsymbol{p}_0^*) + F^*(\boldsymbol{p}_0^*),$$

which verifies the first part of the corollary.

Since F is assumed Legendre, ∂F is single-valued when it is non-empty. Using (2.11) we can rewrite (3.18) as

$$\boldsymbol{A}^* \bar{\boldsymbol{u}}^* = \nabla F(\bar{\boldsymbol{p}}) - \boldsymbol{p}_0^*,$$

which in turn can be rearranged with the help of Fact 2.19 to yield (3.26). ■

In this section, both Theorem 3.3 and Corollary 3.4 were stated in terms of the constraint $\boldsymbol{Ap} = \boldsymbol{b}$ rather than $\boldsymbol{Bp} = \boldsymbol{b_1}$, from (3.22), which explicitly distinguishes the normalization constraint. Clearly both results apply in either case. The only material consequence of the change would be that if $\boldsymbol{B}$ is used, the full rank requirement of the corollary applies also to $\boldsymbol{A}$, by implication. This means that a normalization constraint must not already be implicit in the optimization problem before it is imposed.

3.7 The Consequences of Normalization

In Chapter 1, normalization was characterized as a distinguished constraint that forces the solution to be a probability distribution, instead of simply a non-negative function of the state i. We can find a solution for $\bar{\boldsymbol{p}}$ in (3.26), which carries an explicit parameter for the normalization constraint. In textbooks,

statistical distributions carry no explicit reference to the normalization constraint – it is implicit. Next we examine how to make normalization of the solution for the maxent problem implicit as well. One property we will see is that a solution to the maxent problem (3.23) is normalized in a way that is closely related to the entropy function that was used in its derivation.

First we look at this process for Bregman divergences and later for phi-exponential families. The scale function (or normalizing function) we will describe retains some but not all of the familiar properties of the log-partition function associated with exponential family distributions. Generalization of the properties of this function leads to an additional concept, the escort distribution. The need for the concept of an escort arises directly out of a desire to make normalization implicit in the description of the solution $\bar{\boldsymbol{p}}$. Note, in previous sections we repeatedly emphasized dual connections through the use of the '*' notation. Now we will emphasize the connection to statistical models and switch notation for expressions using '*', like $\boldsymbol{u}^*$, to ones with Greek letters, like $\boldsymbol{\mu}$. Nothing in the setting has really changed.

First we show the existence of a normalizing function T, and some of its properties.

Lemma 3.5 (Scale Function) *Let $F : \mathcal{X} \to (-\infty, \infty]$ be Legendre and differentiable three times*[4]*, $\boldsymbol{p}_0$ be a point in $\operatorname{int}(\operatorname{dom} F)$, $\boldsymbol{A} : \mathcal{X} \to \mathcal{M}$ be a full rank linear map such that $\boldsymbol{A}\boldsymbol{p} = \boldsymbol{b}$ and $\mathbf{1}^{\mathrm{T}}\boldsymbol{p} = 1$ for some $\boldsymbol{b} \in \mathcal{M}$, and $\boldsymbol{p} \in \operatorname{int}(\operatorname{dom} F)$. Furthermore let $\boldsymbol{B}$ as defined in* (3.22) *also be full rank. Then the problem*

$$\min_{\boldsymbol{p}\in\mathcal{X}} \Delta F(\boldsymbol{p}, \boldsymbol{p}_0) \text{ subject to } \boldsymbol{A}\boldsymbol{p} = \boldsymbol{b} \text{ and } \mathbf{1}^{\mathrm{T}}\boldsymbol{p} = 1 \tag{3.28}$$

has a dual formulation

$$\max_{(\boldsymbol{\mu_1},\boldsymbol{\mu})\in\mathbb{R}\times\mathcal{M}^*} \langle \boldsymbol{b}, \boldsymbol{\mu}\rangle + \boldsymbol{\mu_1} - F^*(\boldsymbol{A}^*\boldsymbol{\mu} + \boldsymbol{p}_0^* + \mathbf{1}\mu_1), \tag{3.29}$$

where $\boldsymbol{p}_0^* = \nabla F(\boldsymbol{p}_0)$. Given a solution to (3.29)*, $\bar{\boldsymbol{\mu}}$ and $\bar{\mu}_{\mathbf{1}}$, we can recover the primal solution, $\bar{\boldsymbol{p}}$, via*

$$\bar{\boldsymbol{p}} = \nabla F^* \left(\boldsymbol{A}^*\bar{\boldsymbol{\mu}} + \boldsymbol{p}_0^* + \mathbf{1}\bar{\mu}_{\mathbf{1}}\right). \tag{3.30}$$

Furthermore, there is a differentiable (scalar-valued) function $T : \mathcal{M}^* \to \mathbb{R}$ such that

$$\bar{\boldsymbol{p}} = \nabla F^* \left(\boldsymbol{A}^*\bar{\boldsymbol{\mu}} + \boldsymbol{p}_0^* - \mathbf{1}\, T(\bar{\boldsymbol{\mu}})\right). \tag{3.31}$$

[4]This differentiability condition could potentially be relaxed. We only need to ensure that ∇F^* is continuously differentiable.

Proof The dual formulation (3.29) and the form of the solution (3.30) follow directly from Corollary 3.4 by writing the primal problem as

$$\min_{\boldsymbol{p}\in\mathcal{X}} \Delta F(\boldsymbol{p}, \boldsymbol{p}_0) \text{ subject to } \boldsymbol{B}\boldsymbol{p} = \boldsymbol{b}_{\mathbf{1}}, \tag{3.32}$$

where $\boldsymbol{B} : \mathcal{X} \to \mathbb{R} \times \mathcal{M}$ and $\boldsymbol{b}_{\mathbf{1}} \in \mathbb{R} \times \mathcal{M}$ are defined in (3.22).

For each $\bar{\boldsymbol{\mu}}$, there is a unique $\bar{\mu}_{\mathbf{1}}$ (otherwise the solution to the dual problem is not unique). Existence and uniqueness of T therefore follows from existence and uniqueness of the dual solution. We now only need to show that T is differentiable. Using (3.31) and the sum-to-one constraint we can define T implicitly on the open set $\operatorname{int}(\operatorname{dom} T)$ via

$$g(T(\bar{\boldsymbol{\mu}}), \bar{\boldsymbol{\mu}}) := \mathbf{1}^{\mathrm{T}} \nabla F^* \left(\boldsymbol{A}^* \bar{\boldsymbol{\mu}} + \boldsymbol{p}_0^* - \mathbf{1} T(\bar{\boldsymbol{\mu}})\right) - 1 = 0,$$

where $g : \mathbb{R} \times \mathcal{M}^* \to \mathbb{R}$.

F is assumed to be Legendre and thrice differentiable. Therefore ∇F^* is twice differentiable, and the function $g(T(\bar{\boldsymbol{\mu}}), \bar{\boldsymbol{\mu}}) := \mathbf{1}^{\mathrm{T}} \nabla F^* \left(\boldsymbol{A}^* \bar{\boldsymbol{\mu}} + \boldsymbol{p}_0^* + \mathbf{1}\, T(\bar{\boldsymbol{\mu}})\right) - 1$ is also twice differentiable. Since $T(\bar{\boldsymbol{\mu}})$ is unique, the Implicit Function Theorem applies and (see e.g. Spivak, 1965, Theorem 2-12) guarantees that T must be differentiable. ■

The function T is also termed a partition function.[5] In the special case where the function F corresponds to SBG entropy (and $\nabla F^* = \exp$), it is more commonly pulled outside and identified as the log-partition function. When F is something other than SBG entropy, two things happen: T can no longer be pulled outside in the usual fashion and the the gradient of T is no longer the expectation of the features, as represented by the rows of $\boldsymbol{A}$. Instead we have:

Lemma 3.6 (Escort measure) *Assume that all the assumptions of the above lemma hold. Then the gradient of the function T takes the form*

$$\nabla T\left(\boldsymbol{\mu}\right) = \boldsymbol{A}\boldsymbol{q}, \tag{3.33}$$

for some $\boldsymbol{q}$ that satisfies the sum-to-one constraint $\mathbf{1}^{\mathrm{T}}\boldsymbol{q} = 1$. This $\boldsymbol{q}$ is called the escort measure.

Proof Since T is differentiable, we can use implicit differentiation to write its gradient as

$$0 = (\boldsymbol{A} - \nabla T\left(\boldsymbol{\mu}\right) \mathbf{1}^{\mathrm{T}}) \nabla^2 F^* \left(\boldsymbol{A}^* \boldsymbol{\mu} + \boldsymbol{p}_0^* - \mathbf{1} T(\boldsymbol{\mu})\right) \mathbf{1}.$$

[5] An alternative to the proof strategy used would be to use a global implicit function theorem (Sandberg, 1981).

Since F is Legendre (and hence strictly convex), F^* is strictly convex, and we can rearrange the above expression to yield

$$\nabla T(\boldsymbol{\mu}) = \boldsymbol{A}\frac{\nabla^2 F^*(\boldsymbol{A}^*\boldsymbol{\mu} + \boldsymbol{p}_0^* - \mathbf{1}T(\boldsymbol{\mu}))\,\mathbf{1}}{\mathbf{1}^{\mathrm{T}}\nabla^2 F^*(\boldsymbol{A}^*\boldsymbol{\mu} + \boldsymbol{p}_0^* - \mathbf{1}T(\boldsymbol{\mu}))\,\mathbf{1}}.$$

Setting

$$\boldsymbol{q} = \frac{\nabla^2 F^*(\boldsymbol{A}^*\boldsymbol{\mu} + \boldsymbol{p}_0^* - \mathbf{1}\,T(\boldsymbol{\mu}))\,\mathbf{1}}{\mathbf{1}^{\mathrm{T}}\nabla^2 F^*(\boldsymbol{A}^*\boldsymbol{\mu} + \boldsymbol{p}_0^* - \mathbf{1}\,T(\boldsymbol{\mu}))\,\mathbf{1}}, \tag{3.34}$$

it is easy to observe that $\mathbf{1}^{\mathrm{T}}\boldsymbol{q} = 1$. ■

The reader can observe that $\boldsymbol{q}$ looks 'almost' a probability distribution insofar as it is normalized. However, it may have negative entries. We can go further and guarantee that $\boldsymbol{q}$ is in fact a probability distribution if we impose just one more assumption: additive separability of F.

Lemma 3.7 (Escort probability) *Assume that all the assumptions of the above lemma hold. Furthermore assume that F is additively separable. Then, the gradient of T takes the form*

$$\nabla T(\boldsymbol{\mu}) = \boldsymbol{A}\boldsymbol{q}, \tag{3.35}$$

where $\boldsymbol{q}$ is a probability distribution, that is, $q_n > 0$ and it satisfies the sum-to-one constraint $\mathbf{1}^{\mathrm{T}}\boldsymbol{q} = 1$. This $\boldsymbol{q}$ is called the escort distribution.

Proof In order to show that $\boldsymbol{q}$ is a valid probability distribution, it is enough to show that $q_n \geq 0$ for all i. Since F is Legendre (and hence strictly convex), its Hessian is positive definite. Therefore, the denominator of (3.34) is positive (Horn and Johnson, 1985). On the other hand, since F is additively separable, the Hessian is a diagonal matrix with positive entries along the diagonal. Therefore, the numerator of (3.34) is a positive vector. Together they imply that $q_n > 0$ for all i. ■

Now we have seen that the gradient property: $\nabla T(\boldsymbol{\mu}) = \boldsymbol{A}\,\boldsymbol{p}(\boldsymbol{\mu})$, well-known from the exponential family, is altered to $\nabla T(\boldsymbol{\mu}) = \boldsymbol{A}\,\boldsymbol{q}(\boldsymbol{\mu})$ in this setting. Additive separability of F, along with the other assumptions ensures that $\boldsymbol{q}$ is actually a probability distribution. In this way, (3.34) establishes the relation between $\boldsymbol{q}$ and $\boldsymbol{p}$ known as escorting.

Next, we show that additive separability also turns out to be helpful in establishing another expected property of T: convexity.

Lemma 3.8 *Assume that all the assumptions of the above lemma hold. If each of the componentwise gradients ∇f_i^* is convex, then T is convex.*

Proof Pick $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ in $\operatorname{dom} F^*$ and $\alpha \in (0,1)$, and let $\boldsymbol{\mu}_\alpha := \alpha\boldsymbol{\mu}_1 + (1-\alpha)\boldsymbol{\mu}_2$. Clearly,

$$\mathbf{1}^{\mathrm{T}} \nabla f_i^* [\boldsymbol{A}^* \boldsymbol{\mu}_\alpha + \boldsymbol{p}_0^* - \mathbf{1} T(\boldsymbol{\mu}_\alpha)] = 1. \tag{3.36}$$

Furthermore, since each ∇f_i^* is assumed convex, one can also write

$$\begin{aligned} \mathbf{1}^{\mathrm{T}} \nabla f_i^* [\boldsymbol{A}^* \boldsymbol{\mu}_\alpha - \mathbf{1}(\alpha T(\boldsymbol{\mu}_1) + (1-\alpha) T(\boldsymbol{\mu}_2)) + \boldsymbol{p}_0^*] &\leq \alpha \mathbf{1}^{\mathrm{T}} \nabla f_i^* [\boldsymbol{A}^* \boldsymbol{\mu}_1 - \mathbf{1} T(\boldsymbol{\mu}_1) + \boldsymbol{p}_0^*] \\ &\quad + (1-\alpha) \mathbf{1}^{\mathrm{T}} \nabla f_i^* [\boldsymbol{A}^* \boldsymbol{\mu}_2 - \mathbf{1} T(\boldsymbol{\mu}_2) + \boldsymbol{p}_0^*] \\ &= 1. \end{aligned} \tag{3.37}$$

F^* is Legendre and additively separable, and hence each f_i^* is strictly convex. This in turn implies that ∇f_i^* is an increasing function (Theorem 24.1 Rockafellar, 1970). Using (3.36) and (3.37) one can thus conclude

$$T(\boldsymbol{\mu}_\alpha) \leq \alpha T(\boldsymbol{\mu}_1) + (1-\alpha) T(\boldsymbol{\mu}_2),$$

which shows that T is convex. ■

Whether there are other conditions sufficient to ensure that $\boldsymbol{q}$ is a distribution and T is convex is an open question as far as we know.

To sum up, from the perspective of convex optimization, T arises out of the desire to include a normalization constraint–and make that constraint implicit. Its properties depend essentially on those of F. In particular, T is convex and differentiable given our assumptions on F. The gradient of T determines the escorting relation between $\boldsymbol{q}$ and $\bar{\boldsymbol{p}}$, which collapses in the case of the exponential family. In statistical physics the appearance of escorts has sparked some debate, even to the point of generating debate over whether physical models ought to lead to optimization problems with constraints that are linear in $\boldsymbol{q}$ instead of $\boldsymbol{p}$. We do not pursue that direction further here. Instead, in the next chapter, we look at a class of entropy functions which exhibit escort distributions in a very concrete way and, in later chapters, seek a role for these entropy functions in machine learning.

Chapter 4

Distributions From Deformed Logarithms and Exponentials

In this chapter, we take a different tack in looking at Bregman divergences. Here we will focus on constructing parameterized families of divergences which satisfy certain desirable properties. Our point of departure is the observation that the SBG entropy is defined as $S(\boldsymbol{p}) = \mathbf{1}^{\mathrm{T}} f[\boldsymbol{p}]$ where $f(p_n) = -p_n \log(1/p_n)$. We study entropies and associated divergences generated by replacing log with another well-chosen function, which will be called a *deformed logarithm*. A variety of non-SBG entropies widely studied in statistical physics, such as the Tsallis entropy (*e.g.* Tsallis, 1988; Gell-Mann and Tsallis, 2004), fit into this framework. See Naudts (2004a,b, 2002) for an extensive discussion. The families of distributions associated with these entropies are known as phi-exponential families. Within the framework developed earlier, they are easily seen to be minimizers of a Bregman divergence, recovering one of the main results of (Naudts, 2004b).

We will look at some examples of deformed logarithms and develop a test to check if the entropy associated with them has the Legendre property. The test, which is simple, is also novel to the best of my knowledge. It shows that many parameterized families of entropies are not in fact Legendre. Even so the non-Legendre entropies still yield convex optimization problems, with some different and potentially important properties of their own. We then examine some of the properties of the phi-exponential family which generalize those of the exponential family in an elegant way.

4.1 Phi-logarithm and generalized entropy

First, we define the deformed logarithm, $\log_\phi$, as follows:

Definition 4.1 (Deformed logarithm) *Let* $\phi : [0, \infty) \to [0, \infty)$ *be strictly*

positive and non-decreasing on $(0, \infty)$. *Define* $\log_\phi$ *via*

$$\log_\phi(p) := \int_1^p \frac{1}{\phi(y)}\, dy \tag{4.1}$$

If this integral converges for all finite $p > 0$, *then* $\log_\phi$ *is called a* deformed logarithm.

To see that this definition generalizes log simply set $\phi(p) = p$, which recovers log. It is also easy to see that $\log_\phi$ is a concave increasing function, negative on $(0, 1)$, positive on $(1, +\infty)$, with $\log_\phi(1) = 0$. Of course the integral may diverge at $p = 0$. All these are properties of the familiar log function. An example is

Example 4.1 (q-logarithm) *Let* $\phi(p) = p^q$, $q > 0$. *Then*

$$\log_q(p) := \log_\phi(p) = \begin{cases} \log(p) & \text{if } q = 1 \\ \frac{p^{1-q}-1}{1-q} & \text{otherwise} \end{cases}. \tag{4.2}$$

The function $\log_q$ turns out to be useful in its own right and will reappear in examples to follow (see Figure 4.1(c)).

As an aside, it is worth noting that one important property of log that is generally *not* preserved by $\log_\phi$ is the identity $\log(p) = -\log(1/p)$. Instead this expression can be used to define a *dual logarithm*[1] denoted $\log_\phi^\bullet$:

$$\log_\phi^\bullet(p) := -\log_\phi(1/p). \tag{4.3}$$

In general, this expression may not lead to a properly deformed logarithm conforming to Definition 4.1. It is not hard to show this requirement amounts to defining a new function $\phi^\bullet(p) = p^2\phi(1/p)$, and checking the definition based on this new function. We mention dual logarithms because in cases where they do exist they lead to alternative parameterizations of generalized entropies.

Example 4.2 *Continuing with* $\phi(p) = p^q$, *assuming* $0 < q < 2$, *we still have* $\log_\phi(p) = \log_q(p)$. *Now we also have* $\log_\phi^\bullet(p) = \log_{2-q}(p)$, *since* $p^2\phi(1/p) = p^{2-q}$. *Furthermore,* $\log_\psi(p) = \log_{2-q}(p)/(2-q)$. *In this case* $\log_\psi$ *turns out to be a scaled dual version of* $\log_\phi$.

In the Tsallis literature, sometimes the restriction $0 < q < 2$ is observed, even though $\log_q$ is still concave for higher values of q. This corresponds to the range where $\log_q$ has a dual.

Another property not guaranteed so far, but worth preserving, is the smooth derivative of the logarithm. The following theorem shows how to achieve this

[1] Not meant to be taken in the same sense as Fenchel duality.

by constructing a closely related function ψ and basing a deformed log upon it instead.

Theorem 4.2 *Let ϕ generate a deformed logarithm. Assume also that it satisfies the following regularity condition, which defines a constant, k_ϕ*

$$k_\phi := \int_1^0 \log_\phi(y)\ dy < \infty. \tag{4.4}$$

Now define a function, ψ, via

$$\psi(p) := \left[\int_0^{1/p} \frac{y}{\phi(y)} dy\right]^{-1}. \tag{4.5}$$

This function is well defined and also generates a deformed logarithm.

Proof First note that

$$k_\phi = \int_1^0 \log_\phi(u) du = \int_1^0 \int_1^u \frac{1}{\phi(y)} dy\, du = \int_0^1 \int_u^1 \frac{1}{\phi(y)} dy\, du = \int_0^1 \int_0^y \frac{1}{\phi(y)} du\, dy$$

$$= \int_0^1 \frac{y}{\phi(y)} dy.$$

The integral used in defining $\psi(p)$ can be rewritten as follows:

$$\begin{aligned}\int_0^{1/p} \frac{y}{\phi(y)} dy &= \int_0^1 \frac{y}{\phi(y)} dy + \int_1^{1/p} \frac{y}{\phi(y)} dy \\ &= k_\phi + \int_1^{1/p} \frac{y}{\phi(y)} dy.\end{aligned}$$

The first term is a finite constant by assumption. For the second term fix $p > 0$. Then the integral converges, since $1/\phi(y)$ is integrable over the same interval. Also the second term is strictly greater than -k_ϕ. This implies $1/\psi(p) > 0$, so the same is true of $\psi(p)$. Now if we let p vary it is easy to see that ψ is increasing in p, as required.

Next, the defining equation for $\log_\psi$ (4.1) can be rewritten as

$$\begin{aligned}\log_\psi(p) &= \int_1^p (\psi(y))^{-1}\, dy = \int_1^p \left(k_\phi + \int_1^{1/y} \frac{u}{\phi(u)} du\right) dy \\ &= (p-1)k_\phi + \int_1^p \int_1^{1/y} \frac{u}{\phi(u)} du\, dy.\end{aligned}$$

Again fix $p > 0$. The first term of the sum is clearly finite. In the second term, the inner integral converges, again since $1/\phi(u)$ converges on the same interval. Its absolute value is upper bounded at one end of the range [1,p], which, when multiplied by $p - 1$ serves to bound the entire integral. This shows that $\log_\psi$ is well defined for $p > 0$. ■

Now we can define the generalized entropy using the function $\log_\psi$ as follows:

$$s_\phi(p) := -p \log_\psi \left(\frac{1}{p}\right), \tag{4.6}$$

and verify that the gradient of s_ϕ has the following analytically tractable form

Lemma 4.3

$$\frac{d}{dp} s_\phi(p) = \frac{d}{dp}\left(-p \log_\psi(1/p)\right) = \log_\phi(p) + k_\phi. \tag{4.7}$$

Proof For completeness we include a proof based on Naudts (2004b), Lemma 7.1. in the Appendix. ■

In Figure 4.1 the process of constructing s_ϕ is depicted, with $\phi(p) = p^q$, with q set to each of the three alternatives: 0.5, 1, and 1.5. The setting $q = 1$, shown for comparison, simply recovers SBG entropy.

4.1.1 Phi-exponentials

The inverse of $\log_\phi$ is the *phi-exponential* function, denoted $\exp_\phi$. When $\log_\phi$ takes on a finite value this is well defined. But, unlike log, there is no guarantee that $\log_\phi$ takes on all values in $\mathbb{R}$. Therefore, define $\exp_\phi(v) = 0$ if v is less than every element of range $\log_\phi$ and $\exp_\phi(v) = +\infty$ if v is too large. Figure 4.2 shows some phi-exponential functions. They are convex, increasing functions. The key differences are apparent in the graph: The domain may have an upper bound and the function may reach zero on the left-hand side. The domain is clearly open.

Other properties of $\exp_\phi$,such as convexity, mirror those of $\log_\phi$ (see Naudts, 2004a, 2002). A key difference involves the fact that exp is the only non-trivial function which is its own derivative. However $\exp_\phi$ has the almost-as-useful property:

$$\exp'_\phi(p) = \phi(\exp_\phi(p)). \tag{4.8}$$

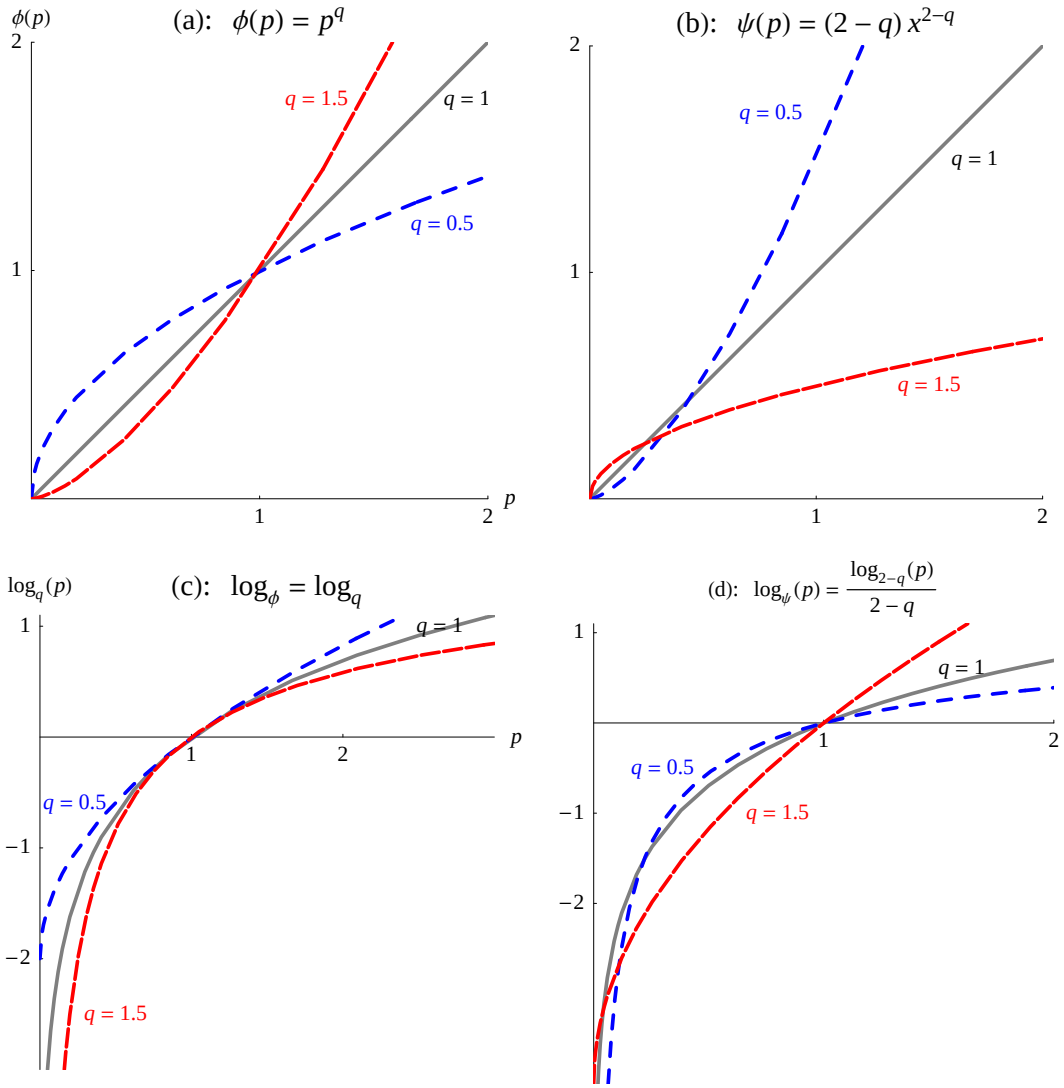


Figure 4.1: Constructing s_ϕ. (a): Select ϕ (b): Calculate ψ using (4.5) (c): Calculate $\log_\phi$ using (4.1) (d) Calculate $\log_\psi$ using (4.1) again, using ψ. (e): (continued on next page)

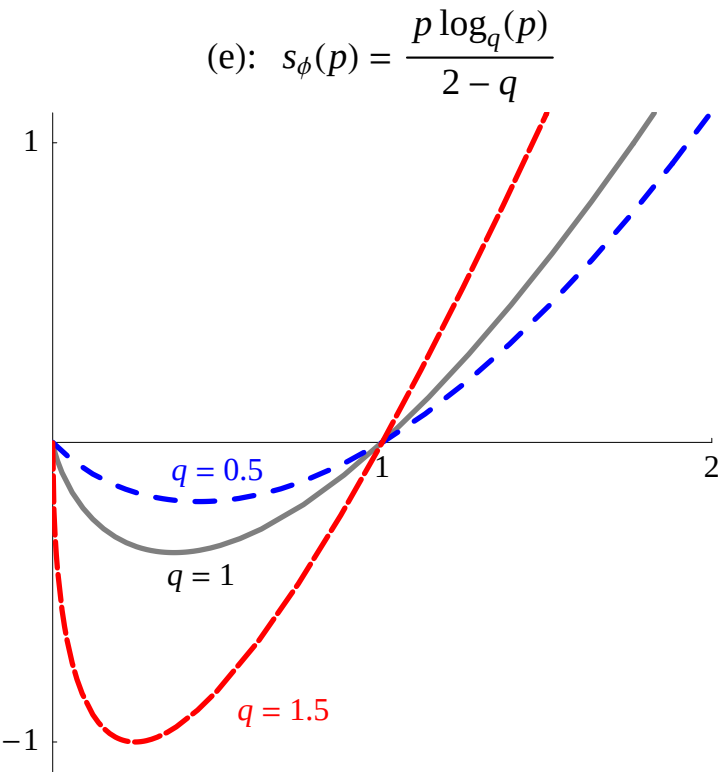


Figure 4.1 *(continued).* (e) Form s_ϕ using (4.6) (continued)

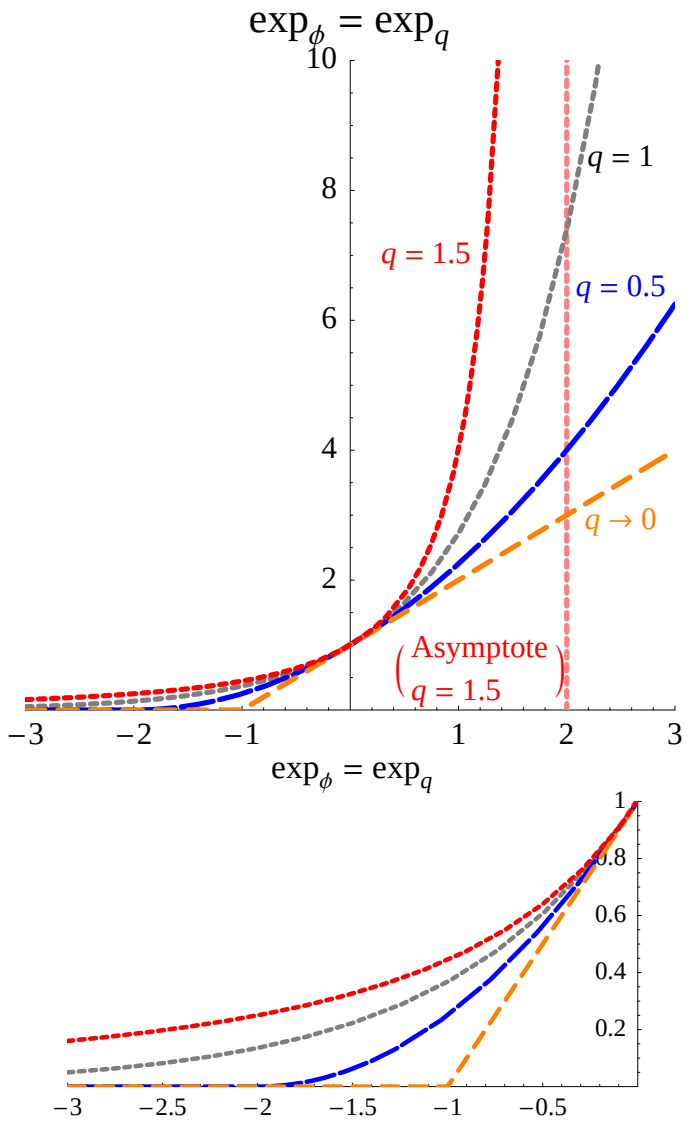


Figure 4.2: The top figure depicts $\exp_\phi$ for $\phi(p) = p^q$ for the various values of q indicated. The lower figure zooms in to better depict when $\exp_\phi$ can achieve the value zero.

Example 4.3 *Define the notation* $[v]_+$ *to mean* v *if the* $v > 0$ *and* 0 *otherwise. Continuing with* $\phi(p) = p^q$, $0 < q < 2$ *we have* $\exp_q(v) = [1 + (1-q)v]_+^{\frac{1}{1-q}}$

Example 4.4 *Notice that the limit as* $q \to 0$ *in Example 4.3 corresponds to the function* $\phi(p) = 1, p > 0$ *(with* $\phi(0) = 0$*). The corresponding* $\exp_\phi$ *is the 'hockey stick' depicted in Figure 4.2.*

With the definitions of $\log_\phi$, $\log_\psi$, and $\exp_\phi$ in hand we can proceed our next task.

4.2 Computing Conjugates

We consider Bregman divergences generated by $S_\phi : \mathbb{R}_+^N \to \mathbb{R}$, where

$$S_\phi(\boldsymbol{p}) := \sum_{n=1}^{N} s_\phi(p_n) \text{ and } s_\phi(p_n) = -p_n \log_\psi(1/p_n). \tag{4.9}$$

First, since S_ϕ is an additively separable function, its conjugate can be expressed in terms of the single term s_ϕ^*. Second, if s_ϕ is Legendre, it is often easier to derive s_ϕ^*. Unfortunately, s_ϕ is not Legendre for all ϕ but there is a simple test for this.

Lemma 4.4 (Legendre Test) *Assume* ϕ *satisfies the conditions of Definition 4.1 and Theorem 4.2. Furthermore if* $\log_\phi(p) \to -\infty$ *as* $p \to 0$ *then* S_ϕ *is Legendre.*

Proof In order to show that s_ϕ is Legendre we need to ensure s_ϕ is differentiable on the interior of its domain, and not even subdifferentiable outside that set. The first part is simple since it is defined using a convergent integral and therefore is differentiable. To verify the condition at the boundary, suppose $\boldsymbol{p}$ is a point on the and $\boldsymbol{x} = \boldsymbol{p} + \boldsymbol{h}$ is a point on the interior. The condition means that $\nexists \boldsymbol{x}^*$ such that $S_\phi(\boldsymbol{p}) + \langle \boldsymbol{x}^*, \boldsymbol{x} - \boldsymbol{p} \rangle \leq S_\phi(\boldsymbol{x}) \forall \boldsymbol{x}$. (For entropy functions the boundary is known *i.e.,* $\boldsymbol{p} = 0$.) Given the additive separability of S_ϕ this is equivalent to the coordinate-wise condition:

$$\begin{aligned} &\nexists x^* \text{ such that } s_\phi(p) + x^* h \leq s_\phi(p+h) \ \forall h \\ &\Leftrightarrow \frac{s_\phi(p+h) - s_\phi(p)}{h} \leq x^* \quad \forall x^*, h \end{aligned}$$

This implies the (one-sided only) derivative $s'_\phi(0) = -\infty$, which occurs iff $\log_\phi(0) = -\infty$. Thus the Legendre test for s_ϕ is to check that $\log_\phi$ diverges to $-\infty$ at zero.

■

Example 4.5 *The q-logarithm (Example (4.2)) passes the test only when $q > 1$, since it has a finite limit $1/(q-1)$ for $0 < q < 1$.*

Lemma 4.5 *Under the same conditions as Theorem 4.4*

1. *If s_ϕ is Legendre the conjugate of s_ϕ can be written as*

$$s^*_\phi(v) = \exp_\phi(v - k_\phi)\left(v + \log_\psi\left(\frac{1}{\exp_\phi(v-k_\phi)}\right)\right). \tag{4.10}$$

2. *If s_ϕ is not Legendre, there is a cut-off c, above which the formula holds, and below which $s^*_\phi = 0$.*

Proof Recall that $s'_\phi(p) = \log_\phi(p) + k_\phi$. In the Legendre case we can write $(s'_\phi)^{-1}(v) = \exp_\phi(v - k_\phi)$. Now use the conjugate formula from Fact 2.19 to get

$$\begin{aligned} s^*_\phi(v) &= v(s'_\phi)^{-1}(v) + (s'_\phi)^{-1}(v)\log_\psi\left(\frac{1}{(s'_\phi)^{-1}(v)}\right) \\ &= \exp_\phi(v-k_\phi)\left(v + \log_\psi\left(\frac{1}{\exp_\phi(v-k_\phi)}\right)\right). \end{aligned} \tag{4.11}$$

In the non-Legendre case, the reasoning is slightly altered. Let $c = \log_\phi(0) + k_\phi$, which is a finite value. If $v \leq c$ then $\operatorname{argmax}_p(vp - s_\phi(p)) \in (\partial s)^{-1}(v) = 0$. Substitution, in the conjugate definition yields the result $s^*_\phi(v) = 0$. For $v > c$, the subdifferential of s_ϕ is still a singleton so the same procedure yields (4.11) again. ■

Example 4.6 *In the SBG case ($\phi(p) = p$), the expression* (4.10) *collapses to* $\exp(v-1)$*, which can be seen by noting that* $\exp_\phi = \exp^\bullet_\phi = \exp$*,* $\log_\psi = \log_\phi = \log$*, and* $k_\phi = 1$*.*

4.3 Phi-exponential Families as Solutions to Generalized Maxent

The central theorem of Naudts (2004b, Theorem 7.2) states that maximizing generalized entropy $\Delta S_\phi(\boldsymbol{p}, \boldsymbol{p}_0)$ under an expectation constraint on the estimators leads to phi-exponential families. By casting the optimization as a generalized maxent problem (3.23), the following lemma is natural:

Lemma 4.6 *Suppose S_ϕ is Legendre and $\boldsymbol{b} \in \mathrm{ri}(\boldsymbol{A}\,\mathrm{dom}\,S_\phi)$. Then, every maximizer of*

$$\max_{\boldsymbol{p}\in\mathbb{R}^N} \Delta S_\phi(\boldsymbol{p},\boldsymbol{p}_0) \text{ subject to } \boldsymbol{A}\boldsymbol{p}=\boldsymbol{b} \quad \mathbf{1}^{\mathrm{T}}\boldsymbol{p}=1 \text{ and } p_n \geq 0. \tag{4.12}$$

is a phi-exponential distribution.

Proof Identify (4.12) with (3.28), from which it follows that the optimal solution $\bar{\boldsymbol{p}}$ satisfies (3.31). The lemma follows by recalling that

$$\bar{\boldsymbol{p}} = \nabla S^*_\phi(\boldsymbol{A}^*\bar{\boldsymbol{\mu}} + \boldsymbol{p}^*_0 - \mathbf{1}\,T(\bar{\boldsymbol{\mu}})) = \exp_\phi[\boldsymbol{A}^*\bar{\boldsymbol{\mu}} + \boldsymbol{p}^*_0 - \mathbf{1}\,T(\bar{\boldsymbol{\mu}}) - \mathbf{1}k_\phi]. \tag{4.13}$$

■

Some observations are worth carrying forward from the more general setting of the previous section. First, we can compute $\boldsymbol{p}^*_0 = \nabla S_\phi\,(\boldsymbol{p}_0)$ as

$$\boldsymbol{p}^*_0 = s'_\phi[\boldsymbol{p}_0] = \log_\phi[\boldsymbol{p}_0] + \mathbf{1}k_\phi.$$

Using this in (4.13) we get

$$\bar{\boldsymbol{p}} = \nabla S^*_\phi\,(\boldsymbol{A}^*\bar{\boldsymbol{\mu}} + \boldsymbol{p}^*_0 - \mathbf{1}\,T(\bar{\boldsymbol{\mu}})) = \exp_\phi[\boldsymbol{A}^*\bar{\boldsymbol{\mu}} + \log_\phi[\boldsymbol{p}_0] - \mathbf{1}\,T(\bar{\boldsymbol{\mu}})].$$

Only in the exponential family case can the term involving $\boldsymbol{p}_0$ be pulled outside cleanly. Next, plugging $\exp'_\phi(v) = \phi(\exp_\phi(v))$ into (3.35) allows us to write the escort distribution (defined in section 3.7), $\boldsymbol{q}$, as

$$\boldsymbol{q} = \frac{\phi \circ \exp_\phi[\boldsymbol{A}^*\bar{\boldsymbol{\mu}} + \boldsymbol{p}^*_0 - \mathbf{1}\,T(\bar{\boldsymbol{\mu}}) - k_\phi]}{\mathbf{1}^{\mathrm{T}}\left(\phi \circ \exp_\phi[\boldsymbol{A}^*\bar{\boldsymbol{\mu}} + \boldsymbol{p}^*_0 - \mathbf{1}\,T(\bar{\boldsymbol{\mu}}) - k_\phi]\right)} = \frac{\phi(\bar{\boldsymbol{p}})}{\mathbf{1}^{\mathrm{T}}\phi(\bar{\boldsymbol{p}})}, \tag{4.14}$$

which recovers Proposition 5.2 of Naudts (2004b). Lastly, the expression for $\boldsymbol{q}$ also provides some perspective on escorting in the phi-exponential family. Recall from Example 4.6, that ϕ is the identity function. So $\boldsymbol{q}$ is simply $\bar{\boldsymbol{p}}$. In the more general case (3.34), we saw that the escorting relation depends on $\nabla^2 F^*$, a function with few restrictions on its form. Comparing (4.14) and (3.34) we see the phi-exponential family case falls in the middle, analytically speaking, since the escorting relation is succinctly determined by ϕ.

4.3.1 Examples

Table 4.1, provides some examples of phi-logarithms and related functions. The 'SBG' column is included as a point of reference and also highlights how the generalized setting can collapse to this familiar case. See also Example 4.6. The column headed 'Naudts' summarizes our running example which also parallels

	SBG	Naudts	Naudts ($q \to 0$)	Tsallis	η-type
$\phi(p)$	p	p^q	1	$\frac{1}{q}p^{2-q}$	$p - \eta$
$\psi(p)$	p	$(2-q)p^{2-q}$	$2p^2$	p^q	$(\frac{1}{p} + \eta \log(\frac{\eta p - 1}{\eta p}))^{-1}$
$\log_\phi(p)$	$\log(p)$	$\log_q(p)$	$p - 1$	$q \log_{2-q}(p)$	$\log(\frac{p-\eta}{1-\eta})$
$\exp_\phi(p)$	$\exp(p)$	$\exp_q(p)$	$(p+1)_+$	$\exp_{2-q}\left(\frac{p}{q}\right)$	$(1-\eta)\exp(p) + \eta$
$\log_\psi(p)$	$\log(p)$	$\frac{1}{2-q}\log_{2-q}(p)$	$\frac{1}{2}(1 - \frac{1}{p})$	$\log_q(p)$	not closed form
k_ϕ	1	$1/(2-q)$	$1/2$	1	$\log(1-\eta)$
$s_\phi(p)$	$p\log(p)$	$\frac{1}{2-q}p\log_q(p)$	$\frac{1}{2}(p^2 - p)$	$-p\log_q\left(\frac{1}{p}\right)$	$-p\log_\psi(1/p)$
$s^*_\phi(v)$	$\exp(v-1)$	$\exp_q(v - k_\phi)^{\frac{(1-q)v+k_\phi}{2-q}}$	$(v - \frac{1}{2})_+ \cdot (\frac{v}{2} + \frac{1}{4})$	$\exp_{2-q}(v-1)^{\frac{(q-1)v+1}{q}}$	$(1-\eta)\exp(v) + \eta v$
Parameter	none	$0 < q < 2$	none	$0 < q < 2$	$\eta < 0$

Table 4.1: Examples of deformed logarithms and related functions. See Section 4.3.1 for discussion.

Naudts (2004b, Ex. 2). The column headed '$q \to 0$' examines the limiting case based on $\phi(x) = x^q$. The column marked 'Tsallis' is closely related to our running example, and many of the expressions, especially those given for s_ϕ appear in the Tsallis literature. The expression for $\exp_\phi$ is suggestive of the form of the distribution resulting from use of these entropy functions.

Other Examples. The final column of Table 4.1 marked 'η-type' is taken from Murata et al. (2004) and is developed as far as possible in this framework. Another example from that paper, which corresponds the Naudts case, with the restriction that $q > 1$ ('beta-type' with $\beta > 0$).

Another example from the physics literature can be found in Kaniadakis et al. (2004, 2005). They study deformed logarithms of the form

$$\log_{(\kappa,\rho)}(p) := p^\rho \frac{p^\kappa - p^{-\kappa}}{2\kappa},$$

where κ and ρ are parameters which satisfy the following set of relations:

$$\rho + |\kappa| \geq 0 \ \wedge \ \rho - |\kappa| \leq 0 \ \wedge \ \rho + |\kappa| < 1 \ \wedge \ \rho - |\kappa| > -1.$$

Over the years physicists have suggested a variety of functions that may be thought of as an 'entropy'. See Masi (2005) for a summary of some interesting relations between some of them. Not all of these functions are convex. In some cases such as Renyi entropy, there is a monotonic transformation that renders it convex. However we do not pursue these examples further here.

4.4 Appendix: Proof of Lemma 4.3.

Proof

$$\begin{aligned}
\frac{d}{dp}\left(-p \log_\psi(1/p)\right) &= -\log_\psi(1/p) + \frac{1}{p\,\psi(1/p)} \\
&= -\int_1^{1/p} \frac{1}{\psi(u)} du + \frac{1}{p}\int_0^p \frac{y}{\phi(y)} dy \\
&= -\int_1^{1/p}\int_0^{1/u} \frac{y}{\phi(y)} dy\, du + \frac{1}{p}\int_0^p \frac{y}{\phi(y)} dy \\
&= -\int_1^{p}\int_0^{z} \frac{-1}{z^2}\frac{y}{\phi(y)} dy\, dz + \frac{1}{p}\int_0^p \frac{y}{\phi(y)} dy
\end{aligned}$$

Concentrating on the first integral for a moment yields

$$\begin{aligned}
-\int_1^p \int_0^z \frac{-1}{z^2}\frac{y}{\phi(y)}dy\,dz &= -\int_1^p \int_0^1 \frac{-1}{z^2}\frac{y}{\phi(y)}dy\,dz - \int_1^p \int_1^z \frac{-1}{z^2}\frac{y}{\phi(y)}dy\,dz \\
&= -\int_0^1 \int_1^p \frac{-1}{z^2}\frac{y}{\phi(y)}dz\,dy - \int_1^p \int_y^p \frac{-1}{z^2}\frac{y}{\phi(y)}dz\,dy \\
&= -\int_0^1 \frac{y}{\phi(y)}\left(\frac{1}{p}-1\right)dy - \int_1^p \frac{y}{\phi(y)}\left(\frac{1}{p}-\frac{1}{y}\right)dy \\
&= -\frac{1}{p}\int_0^p \frac{y}{\phi(y)}dy + \int_0^1 \frac{y}{\phi(y)}dy + \int_1^p \frac{1}{\phi(y)}dy
\end{aligned}$$

Adding back the dropped term cancels the first term in the last line, giving

$$\int_0^1 \frac{y}{\phi(y)}dy + \int_1^p \frac{1}{\phi(y)}dy = k_\phi + \log_\phi(p),$$

which is the desired result. ■

Chapter 5

Applications of the Framework

5.1 Introduction

In this chapter we show how various existing models fit into the framework developed in Chapters 2 to 4. Topics covered include information geometry and regression models inspired by the l1-regularization literature. The range of topics highlights the usefulness of the framework and the process of translating these topics into the notation of this thesis hopefully clarifies how to use it in applications.

5.2 Information Geometry and Bregman Projections

In this section, we apply the duality theorems from the previous chapter to derive two well-known results. The first is from Pietra et al. (2002b) and Bauschke (2003), and characterizes the properties of a Bregman projection. The second is from Murata et al. (2004), and characterizes projections onto m-flat and U-flat manifolds. Even though different techniques were originally used to prove these results, when translated to the language of convex analysis, both are essentially equivalent.

5.2.1 Bregman Projections

Suppose $F : \mathcal{X} \to (-\infty, \infty]$ is a Legendre convex function, $\mathcal{C}$ is a closed convex set in $\mathcal{X}$ with $\mathcal{C} \cap \operatorname{int}(\operatorname{dom} F) \neq \emptyset$, and $\boldsymbol{p}_0 \in \operatorname{int}(\operatorname{dom} F)$. Then, the Bregman projection of $\boldsymbol{p}_0$ onto $\mathcal{C}$ is defined as (Bauschke, 2003)

$$P_{\mathcal{C}}^{F}(\boldsymbol{p}_0) = \operatorname*{argmin}_{\boldsymbol{p} \in \mathcal{C}} \Delta F(\boldsymbol{p}, \boldsymbol{p}_0). \tag{5.1}$$

Since F is assumed Legendre, the Bregman projection is a unique point in $\text{int}(\text{dom}\, F)$ (Bauschke, 2003). It is easy to see that the generalized maxent problem (3.23) is a Bregman projection onto the set

$$\{\boldsymbol{p} \in \mathbb{R}^N : \boldsymbol{B}\boldsymbol{p} = \boldsymbol{b_1} \text{ and } p_n \geq 0\}.$$

In this section the presence or absence of the normalization constraint does not affect the results, so with out loss of generality we can consider the set

$$\mathcal{C} = \{\boldsymbol{p} \in \mathbb{R}^N : \boldsymbol{A}\boldsymbol{p} = \boldsymbol{b} \text{ and } p_n \geq 0\},$$

where $\boldsymbol{A}$ may or may not include a normalization constraint as one of its rows.

We now characterize the Bregman projection onto a linear subspace (see Proposition 3.2 of Pietra et al. (2002a) as well as Corollary 4.3 of Bauschke (2003)).

Theorem 5.1 *Suppose $F : \mathcal{X} \to (-\infty, \infty]$ is a Legendre convex function, and $\boldsymbol{A} : \mathcal{X} \to \mathcal{M}$ is a linear map. Furthermore, assume that F is co-finite,* i.e., $\text{dom}\, F^* = \mathcal{X}^*$. *Define*

$$\mathcal{T} := \{\boldsymbol{p} \in \text{dom}\, F : \boldsymbol{A}\boldsymbol{p} = \boldsymbol{A}\boldsymbol{p_b}\} \tag{5.2}$$

for some $\boldsymbol{p_b} \in \text{int}(\text{dom}\, F)$. Define

$$\mathcal{Q} := \nabla F^* \left(\boldsymbol{p}_0^* + \text{range}\, \boldsymbol{A}^*\right), \tag{5.3}$$

and let $\boldsymbol{p}_0 \in \text{int}(\text{dom}\, F)$. Then each of the following conditions for a point $\bar{\boldsymbol{p}} \in \mathcal{X}$ characterizes the Bregman projection $P_{\mathcal{T}}^F(\boldsymbol{p}_0)$:

i. $\bar{\boldsymbol{p}} = \text{argmin}_{\boldsymbol{p} \in \mathcal{T}}\, \Delta F(\boldsymbol{p}, \boldsymbol{p}_0)$.

ii. $\bar{\boldsymbol{p}} = \text{argmin}_{\boldsymbol{q} \in \mathcal{Q}}\, \Delta F(\boldsymbol{p_b}, \boldsymbol{q})$.

iii. $\mathcal{T} \cap \mathcal{Q} = \{\bar{\boldsymbol{p}}\}$.

iv. $\forall \boldsymbol{p} \in \mathcal{T}$ *and* $\forall \boldsymbol{q} \in \mathcal{Q}$ *we have* $\Delta F(\boldsymbol{p}, \boldsymbol{q}) = \Delta F(\boldsymbol{p}, \bar{\boldsymbol{p}}) + \Delta F(\bar{\boldsymbol{p}}, \boldsymbol{q})$.

Proof The first part is simply the definition of $P_{\mathcal{T}}^F(\boldsymbol{p}_0)$. Identify $\min_{\boldsymbol{p} \in \mathcal{T}} \Delta F(\boldsymbol{p}, \boldsymbol{p}_0)$ with (3.24), by setting $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{p_b}$. After some algebra, the dual objective (3.25) can be written as

$$\langle \boldsymbol{b}, \boldsymbol{u}^* \rangle - F^*(\boldsymbol{A}^*\boldsymbol{u}^* + \boldsymbol{p}_0^*) + F^*(\boldsymbol{p}_0^*) = \Delta F^*(\boldsymbol{p}_0^*, \boldsymbol{p}_{\boldsymbol{b}}^*) - \Delta F^*(\boldsymbol{A}^*\boldsymbol{u}^* + \boldsymbol{p}_0^*, \boldsymbol{p}_{\boldsymbol{b}}^*),$$

where $\boldsymbol{p}_{\boldsymbol{b}}^* := \nabla F^* (\boldsymbol{p_b})$. Since $\Delta F^*(\boldsymbol{p}_0^*, \boldsymbol{p}_{\boldsymbol{b}}^*)$ does not depend on $\boldsymbol{u}^*$, maximizing the dual is equivalent to minimizing $\Delta F^*(\boldsymbol{A}^*\boldsymbol{u}^* + \boldsymbol{p}_0^*, \boldsymbol{p}_{\boldsymbol{b}}^*)$, which in turn can

be rewritten using (2.13) as $\Delta F(\boldsymbol{p_b}, \nabla F^*\left(\boldsymbol{A}^*\boldsymbol{u}^* + \boldsymbol{p}_0^*\right))$. All in all, we want to find a $\boldsymbol{q} \in \nabla F^*\left((\boldsymbol{p}_0^* + \operatorname{range}\boldsymbol{A}^*) \cap \operatorname{dom} F^*\right)$ which minimizes $\Delta F(\boldsymbol{p_b}, \boldsymbol{q})$. But since F is assumed co-finite, $\operatorname{dom} F^* = \mathcal{X}^*$, and we can equivalently find a $\boldsymbol{q}$ in $\nabla F^*\left(\boldsymbol{p}_0^* + \operatorname{range}\boldsymbol{A}^*\right) = \mathcal{Q}$. This proves the second part. From the first and second parts and the fact that $\bar{\boldsymbol{p}}$ is unique it follows that $\mathcal{T} \cap \mathcal{Q} = \{\bar{\boldsymbol{p}}\}$. This proves the third part.

For every $\boldsymbol{q} \in \mathcal{Q}$ there exists a $\boldsymbol{u}^*$ such that $\nabla F\left(\boldsymbol{q}\right) = \boldsymbol{A}^*\boldsymbol{u}^* + \boldsymbol{p}_0^*$. Since $\bar{\boldsymbol{p}} \in \mathcal{Q}$, let $\bar{\boldsymbol{u}}^*$ be associated with $\bar{\boldsymbol{p}}$. Then $\nabla F\left(\bar{\boldsymbol{p}}\right) - \nabla F\left(\boldsymbol{q}\right) = \boldsymbol{A}^*\bar{\boldsymbol{u}}^* - \boldsymbol{A}^*\boldsymbol{u}^*$. Given $\boldsymbol{p} \in \mathcal{T}$ we can write

$$\langle \boldsymbol{p} - \bar{\boldsymbol{p}}, \nabla F(\bar{\boldsymbol{p}}) - \nabla F(\boldsymbol{q})\rangle = \langle \boldsymbol{p} - \bar{\boldsymbol{p}}, \boldsymbol{A}^*(\bar{\boldsymbol{u}}^* - \boldsymbol{u}^*)\rangle = \langle \boldsymbol{A}(\boldsymbol{p} - \bar{\boldsymbol{p}}), \bar{\boldsymbol{u}}^* - \boldsymbol{u}^*\rangle\,.$$

Since both $\boldsymbol{p}$ and $\bar{\boldsymbol{p}}$ are in $\mathcal{T}$ it follows that $\boldsymbol{A}(\boldsymbol{p} - \bar{\boldsymbol{p}}) = 0$, and hence

$$\langle \boldsymbol{p} - \bar{\boldsymbol{p}}, \nabla F(\bar{\boldsymbol{p}}) - \nabla F(\boldsymbol{q})\rangle = 0 \quad . \tag{5.4}$$

The last part now follows from Fact 2.27. ■

Figure 5.1 serves to illustrate each of the four characterizations of optimality from Theorem 5.1, ordered left to right, top to bottom. All four plots picture the same solution. The first depicts $\bar{\boldsymbol{p}}$ as the solution to the Bregman projection problem (5.1). The second depicts the solution as the intersection of two manifolds. Under the conditions given in this section, that intersection is a singleton, as depicted. The third plot depicts the reverse distance problem.

The orthogonality condition 5.4 also serves as an equivalent characterization of the triangle equality statement in the fourth part of the theorem. Formally it is the orthogonality condition between $\boldsymbol{p} - \bar{\boldsymbol{p}}$ and $\bar{\boldsymbol{p}}^* - \boldsymbol{q}^*$, where $\bar{\boldsymbol{p}} = \nabla F\left(\bar{\boldsymbol{p}}\right)$ and $\boldsymbol{q}^* = \nabla F\left(\boldsymbol{q}\right)$. Informally, we refer to it as pseudo-orthogonality of the angle formed by $\boldsymbol{p}$, $\bar{\boldsymbol{p}}$, and $\boldsymbol{q}$. This is depicted in the fourth plot of Figure 5.1.

A final comment is in order about the co-finiteness of F, used in this section, which is not required elsewhere in the thesis. It is employed here for two reasons: First, it ensures that we can restrict our attention to the set $\mathcal{Q} := \nabla F^*\left(\boldsymbol{p}_0^* + \operatorname{range}\boldsymbol{A}^*\right)$ instead of $\nabla F^*\left((\boldsymbol{p}_0^* + \operatorname{range}\boldsymbol{A}^*) \cap \operatorname{dom} F^*\right)$. Second, since $\operatorname{dom} F^* = \mathcal{X}^*$, the set

$$\nabla F^*\left((\boldsymbol{p}_0^* + \operatorname{range}\boldsymbol{A}^*) \cap \operatorname{dom} F^*\right) = \mathcal{Q}$$

is a closed convex set. This ensures that the Bregman projection onto $\mathcal{Q}$ is unique.

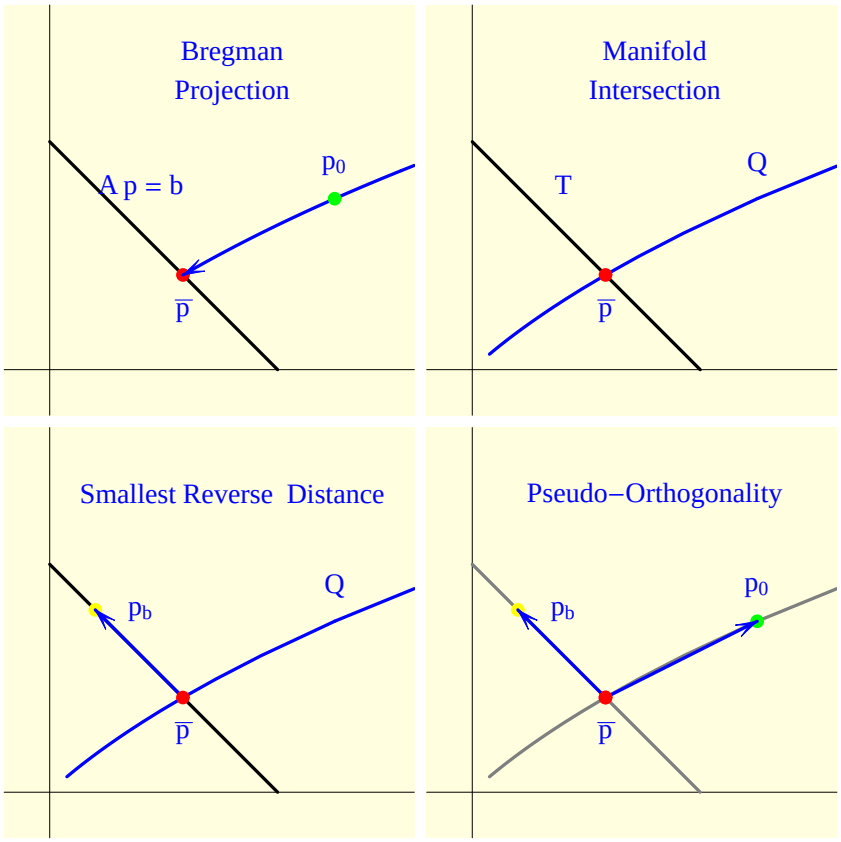


Figure 5.1: Four views of optimality. Illustrates the four characterizations of optimality provided by Theorem 5.1, ordered left to right, top to bottom.

5.2.2 U-Divergences

In the information geometry literature, statistical manifolds are often discussed in terms of their differential geometry (Amari and Nagaoka, 1993). In this section we look at manifolds in terms of the duality theory discussed so far. In this way we are able to motivate them from the perspective of optimization problems and their duals.

One type of statistical manifold emerges from the analysis of so-called U-divergences (Eguchi, 2005; Murata et al., 2004). These turn out to be equivalent to Bregman divergences, and are defined as follows.

Definition 5.2 (*e.g.* Eguchi (2005)) *Suppose $F : \mathcal{X} \to (-\infty, \infty]$ is a Legendre convex function, and F^* its Fenchel-Legendre conjugate. Then, the U-divergence $D_{F^*} : \operatorname{int}(\operatorname{dom} F) \times \operatorname{int}(\operatorname{dom} F) \to [0, \infty)$ is defined as*

$$D_{F^*}(\boldsymbol{p}, \boldsymbol{p}_0) = F^*(\nabla F(\boldsymbol{p}_0)) - F^*(\nabla F(\boldsymbol{p})) - \langle \boldsymbol{p}, \nabla F(\boldsymbol{p}_0) - \nabla F(\boldsymbol{p}) \rangle . \qquad (5.5)$$

Using (2.8) and (2.13) it is easy to see that for all $\boldsymbol{p}, \boldsymbol{p}_0 \in \operatorname{int}(\operatorname{dom} F)$

$$D_{F^*}(\boldsymbol{p}, \boldsymbol{p}_0) = \Delta F^*(\boldsymbol{p}_0^*, \boldsymbol{p}^*) = \Delta F(\boldsymbol{p}, \boldsymbol{p}_0),$$

where $\boldsymbol{p}^* = \nabla F(\boldsymbol{p})$ and $\boldsymbol{p}_0^* = \nabla F(\boldsymbol{p}_0)$. In other words, except at the boundary points, U-divergences are identical to Bregman divergences. Recall that, for a Legendre convex function F the Bregman projection is a unique point in $\operatorname{int}(\operatorname{dom} F)$. Hence the following lemma is immediate.

Lemma 5.3 *Let F and F^* as above, $\mathcal{C}$ a closed convex set in $\mathcal{X}$ with $\mathcal{C} \cap \operatorname{int}(\operatorname{dom} F) \neq \emptyset$, and $\boldsymbol{p}_0 \in \operatorname{int}(\operatorname{dom} F)$. Then*

$$P_{\mathcal{C}}^F(\boldsymbol{p}_0) = \operatorname*{argmin}_{\boldsymbol{p} \in \mathcal{C}} \Delta F(\boldsymbol{p}, \boldsymbol{p}_0) = \operatorname*{argmin}_{\boldsymbol{p} \in \mathcal{C}} D_{F^*}(\boldsymbol{p}, \boldsymbol{p}_0). \tag{5.6}$$

Two types of subspaces, defined below, are often associated with U-divergences.

Definition 5.4 (U-flat subspace) *Let F, F^* and $\boldsymbol{p}_0$ as above. Let $\boldsymbol{A} : \mathcal{X} \to \mathcal{M}$ be a linear mapping and $\boldsymbol{p}_0^*$ be defined as $\boldsymbol{p}_0^* := \nabla F(\boldsymbol{p}_0)$. Then, the set $\mathcal{Q} \subseteq \mathcal{X}$ of the form*

$$\mathcal{Q}(\boldsymbol{p}_0) = \{\boldsymbol{p} \in \mathcal{X} : \boldsymbol{p} = \nabla F^*(\boldsymbol{A}^* \boldsymbol{u}^* + \boldsymbol{p}_0^*) \text{ for some } \boldsymbol{u}^* \in \mathcal{M}^*\} \tag{5.7}$$

is called a U-flat subspace.

Definition 5.5 (m-flat subspace) *A affine subspace in $\mathcal{X}$ that passes through a point $\boldsymbol{p_b} \in \operatorname{int}(\operatorname{dom} F)$ and orthogonal to $\mathcal{Q}$ defined by*

$$\mathcal{T}(\boldsymbol{p_b}) := \{\boldsymbol{p} \in \operatorname{dom} F : \boldsymbol{A}(\boldsymbol{p} - \boldsymbol{p_b}) = 0\}. \tag{5.8}$$

is called an m-flat subspace.

It is easy to see that the set $\mathcal{Q}$ defined in (5.3) is a U-flat subspace. Similarly, the set $\mathcal{T}$ defined in (5.2) is a m-flat subspace. The following is a key theorem of (Murata et al., 2004).

Theorem 5.6 (Theorem 2 Murata et al. (2004)) *The two optimizations problems*

$$\operatorname*{argmin}_{\boldsymbol{p}} D_{F^*}(\boldsymbol{p}, \boldsymbol{p}_0) \text{ s.t. } \boldsymbol{p} \in \mathcal{T}(\boldsymbol{p_b}) \text{ for a fixed } \boldsymbol{p}_0 \tag{5.9}$$

and

$$\min_{\boldsymbol{q}} D_{F^*}(\boldsymbol{p_b}, \boldsymbol{q}) \text{ s.t. } \boldsymbol{q} \in \mathcal{Q}(\boldsymbol{p}_0) \text{ for a fixed } \boldsymbol{p_b} \tag{5.10}$$

have the same solution $\bar{\boldsymbol{p}} = \mathcal{Q}(\boldsymbol{p}_0) \cap \mathcal{T}(\boldsymbol{p_b})$.

The theorem follows immediately from Lemma 5.3 and Theorem 5.1. For completeness we repeat the proof of Murata et al. (2004) which does not use any concepts from convex analysis. Although their proof is simpler, it is not clear if it can be extended to other constraint sets. On the other hand, as shown by Bauschke (2003), Theorem 5.1 can be extended in a straightforward manner to deal with affine subspaces and convex cones.

Proof Using the dimension of $\mathcal{Q}$ and co-dimension of $\mathcal{T}$ it is easy to see that $\mathcal{T} \cap \mathcal{Q}$ consists of a singleton $\bar{\boldsymbol{p}}$.

Choose $\boldsymbol{q} \in \mathcal{Q}$. By the definition of $\mathcal{Q}$ we have $\nabla F(\boldsymbol{q}) - \nabla F(\boldsymbol{p}_0) \in \operatorname{range} \boldsymbol{A}^*$. Similarly, since $\bar{\boldsymbol{p}} \in \mathcal{Q}$, we have $\nabla F(\bar{\boldsymbol{p}}) - \nabla F(\boldsymbol{p}_0) \in \operatorname{range} \boldsymbol{A}^*$. Putting both facts together $\nabla F(\boldsymbol{q}) - \nabla F(\bar{\boldsymbol{p}}) \in \operatorname{range} \boldsymbol{A}^*$. But $\bar{\boldsymbol{p}}$ also belongs to $\mathcal{T}$, hence $\boldsymbol{A}(\boldsymbol{p} - \bar{\boldsymbol{p}}) = 0$ for all $\boldsymbol{p} \in \mathcal{T}$. Altogether $\langle \boldsymbol{p} - \bar{\boldsymbol{p}}, \nabla F(\boldsymbol{q}) - \nabla F(\bar{\boldsymbol{p}}) \rangle = 0$. Using the Pythagorean relation (2.10) from Fact 2.27

$$\Delta F(\boldsymbol{p}, \boldsymbol{q}) = \Delta F(\boldsymbol{p}, \bar{\boldsymbol{p}}) + \Delta F(\bar{\boldsymbol{p}}, \boldsymbol{q}).$$

In particular

$$\Delta F(\boldsymbol{p_b}, \boldsymbol{q}) = \Delta F(\boldsymbol{p_b}, \bar{\boldsymbol{p}}) + \Delta F(\bar{\boldsymbol{p}}, \boldsymbol{q}), \text{ for any } \boldsymbol{q} \in \mathcal{Q}.$$

Since $\Delta F(\bar{\boldsymbol{p}}, \boldsymbol{q}) \geq 0$ we have

$$\Delta F(\boldsymbol{p_b}, \boldsymbol{q}) \geq \Delta F(\boldsymbol{p_b}, \bar{\boldsymbol{p}}), \text{ for any } \boldsymbol{q} \in \mathcal{Q},$$

which shows that $\bar{\boldsymbol{p}} = \operatorname{argmin}_{\boldsymbol{q} \in \mathcal{Q}} \Delta F(\boldsymbol{p_b}, \boldsymbol{q})$. Similarly, we observe that

$$\Delta F(\boldsymbol{p}, \boldsymbol{p}_0) \geq \Delta F(\bar{\boldsymbol{p}}, \boldsymbol{p}_0), \text{ for any } \boldsymbol{p} \in \mathcal{T},$$

which shows that $\bar{\boldsymbol{p}} = \operatorname{argmin}_{\boldsymbol{p} \in \mathcal{T}} \Delta F(\boldsymbol{p}, \boldsymbol{p}_0)$. ■

Throughout this section, none of the results depended on the normalization constraint in any way. In particular $\boldsymbol{p}$ and $\boldsymbol{p}_0$ are not required to be normalized, although this could be imposed if desired. However, this does not mean that normalization is a trivial constraint, particularly when we look at the way it is addressed in statistics, as the next section shows.

5.3 Boosting

In this section we show how the *boosting problem* (Schapire, 2001) fits into our framework. We distinguish between the boosting problem and boosting algo-

rithms. Our focus is the boosting problem, and in a later chapter we will discuss algorithms for solving it. Originally the two concepts were introduced together and were intertwined. This turns out to be unnecessary. The boosting problem can be described and analyzed without reference to a specific algorithm.

Our development closely follows that of Lebanon and Lafferty (2001), with necessary notation changes to make the picture clear in our setting. They showed that AdaBoost (Collins et al., 2002) and exponential family maximum likelihood estimation could be seen as identical, except for a special normalization constraint utilized in the latter. Further specialization of both models leads to a correspondence between logistic regression and binary Adabost. Here we continue the connection-making, fitting these models into the even larger picture developed in this thesis.

5.3.1 Notation

The general problem in boosting involves an input-output pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and a number of deterministic functions $f_j : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. Often, each f_j is a classifier, termed a weak learner. The goal is to find the best set of weights to use to combine the outputs of the weak learners to form a better classifier, hence the use of the term boosting. The outputs of f_j's are known on the data set, and are available to classify new examples. In other respects they are black boxes and the results of this section do not require the interpretation just given in order to hold.

In the problem setup of Lebanon and Lafferty (2001), the data sample is denoted $((x_n, y_n))_{n=1}^N$. We will need several facts about the sample. The empirical distribution is

$$\tilde{p}_{X,Y}(x, y) = \frac{1}{N} \sum_{n=1}^{N} \delta(x_n, x)\delta(y_n, y).$$

This function mainly has the value zero on points not in the sample and (usually) takes the value $1/N$ on points in the sample. The need for the redundant looking subscript on $\tilde{p}_{X,Y}(x, y)$ will become apparent in a moment. The empirical marginal and empirical conditional distributions are defined analogously as

$$\tilde{p}_X(x, y) = \frac{1}{N} \sum_{n=1}^{N} \delta(x_n, x), \text{ and}$$

$$\tilde{p}_{Y|X}(x, y) = \begin{cases} \frac{\tilde{p}_{X,Y}(x,y)}{\tilde{p}_X(x,y)} & \text{if} \tilde{p}_X(x, y) \neq 0 \\ 0 & \text{if } \tilde{p}_X(x, y) = 0. \end{cases}$$

Note that $\tilde{p}_X(\cdot, y)$ is a probability distribution–the same one at each y. Also,

$\tilde{p}_{Y|X}(x,\cdot)$ is a probability distribution–a different one at each x. This notation makes it easier to switch to vector notation. Consider the pair (x,y) as the result of a one-to-one finite map from a generic index set $\mathcal{I} = \{1, ..., |\mathcal{X}|\,|\mathcal{Y}|\}$ to the sample space $\mathcal{X} \times \mathcal{Y}$. Thus for $i \in I$ it makes sense to write (x,y) = (x,y)(i), with the (x, y) on the right as the name of the map in one direction. We can switch back with the map i = $(x, y)^{(-1)}$, so that we write $i = i(x, y)$. Naturally the index i is more conducive to defining vectors elementwise. In a mathematical sense this could be considered trivial, but, from the viewpoint of understanding, it is critical. The change makes it clear that if all indexes and are taken over (x, y) rather than separately for x and y, it becomes easy to take advantage of the vector/matrix notation. In the standard statistical style of presentation, this is not so common.

Based on the foregoing discussion the vectors $\widetilde{\boldsymbol{p}}_{X,Y}$, $\widetilde{\boldsymbol{p}}_X$, $\widetilde{\boldsymbol{p}}_{Y|X}$ all have obvious definitions based on the definitions of $\tilde{p}_{X,Y}(x, y)$, $\tilde{p}_X(x, y)$ and $\tilde{p}_{Y|X}(x, y)$.

The objective function is defined as

$$\sum_{x\in\mathcal{X}} \tilde{p}_X(x) \sum_{y\in\mathcal{Y}} \Delta s(p(y|x), q(y|x)).$$

In Lebanon and Lafferty (2001), s is SBG entropy ($s(p) = p\log(p)$) and $q(y|x)$ is an initial guess, an input variable. That means Δs is the same as the unnormalized KL-divergence (theorem ref). Also, the notation used in Lebanon and Lafferty (2001) contains some syntactic sugar meant to convey the role of certain terms. The term $\tilde{p}_X(x)$ could be written as $\tilde{p}_X(x, y)$. The index $(y|x)$ should be regarded as just a different way to express the index (x, y), while $p(y|x)$ and $q(y|x)$ are actually meant to be just positive measures. With all that in mind, the objective is equivalent to

$$\begin{aligned}\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \tilde{p}_X(x, y)\Delta s(p(x, y), q(x, y) &= \sum_{i\in\mathcal{I}} \tilde{p}_X(i)\Delta s(p(i), q(i)) \\ &= \langle \widetilde{\boldsymbol{p}}_X, \Delta s[\boldsymbol{p}, \boldsymbol{q}]\rangle\,.\end{aligned}$$

Now the connection of the objective function to our setting is clear. The function is a weighted version of unnormalized KL-divergence. The weights are dependent on the data sample (but not the y-values). Otherwise it perfectly fits the format of (3.23). Next, the feature constraints (Equation 1, Lebanon and Lafferty (2001)). are given as:

$$\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \tilde{p}_X(x)p(y|x)(f_j(x, y) - \mathbb{E}_{\tilde{p}}(f_j|x)) = 0 \text{ for } j = 1 \ldots J. \tag{5.11}$$

We can rewrite the inner expectation term, showing how to express it as a function

of (x,y)

$$\tilde{f}_j(x,y) := \tilde{f}_j(x) := \mathbb{E}_{\tilde{p}}(f_j|x) = \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \tilde{p}_{Y|X}(x,y) f_j(x,y).$$

Now we can introduce the term, $\tilde{\boldsymbol{f}}_j$ using the usual definition, and use it to construct the matrix

$$\widetilde{\boldsymbol{F}} = \left(\tilde{\boldsymbol{f}}_1, \ldots, \tilde{\boldsymbol{f}}_J\right)^T,$$

which can be used to write the feature constraint (5.11) as

$$(\boldsymbol{F} - \widetilde{\boldsymbol{F}})\ (\operatorname{diag} \widetilde{\boldsymbol{p}}_X)\ \boldsymbol{p} = \mathbf{0}. \tag{5.12}$$

In this formulation, $\widetilde{\boldsymbol{F}}$ reflects an interaction between the features and the sample data[1]. This differs from the unconditional formulation typified by the classic maxent problem. There the goal is to satisfy a constraint of the form $\boldsymbol{A}\boldsymbol{p} = \boldsymbol{b}$, where the data only enters the problem through $\boldsymbol{b}$. Here $\boldsymbol{F}$ plays the role of $\boldsymbol{A}$, while $\boldsymbol{p}$ plays the same role as before, and the middle term is purely data dependent. The first term $(\boldsymbol{F} - \widetilde{\boldsymbol{F}})$, in effect, represents a data-dependent feature matrix.

The normalization constraint is also affected by the conditional setup. Instead of the unconditional normalization corresponding to $\mathbf{1}^{\mathrm{T}}\boldsymbol{p} = 1$, we have

$$\sum_{y\in\mathcal{Y}} p(y|x) = 1 \quad \forall x \in \mathcal{X}.$$

Following a similar line of reasoning, and relying on the natural index map (where y values vary fastest) we can express this as

$$(\boldsymbol{I}_{|\mathcal{X}|} \otimes \mathbf{1}^{\mathrm{T}}_{|\mathcal{Y}|})\ \boldsymbol{p} = \mathbf{1}.$$

In the boosting problem the single normalization constraint may be replaced with many sparse constraints, one for each x. The choice of whether or not to impose this constraint leads to the distinction between AdaBoost (unnormalized), and maximum likelihood for exponential models. Lebanon and Lafferty point out that further model choice distinctions can be made based on choices of $\mathcal{Y}$ and f_j. For example letting $\mathcal{Y} = \{-1, 1\}$ leads to logistic regression in the normalized case and binary AdaBoost otherwise.

[1]Under the *consistent data assumption* each value of x_n is associated with only one value of y. In this case the expectation above is concentrated on a single value equal to $f(x_n, y_n)$.

5.4 Regression Models and Generalized Maxent

The usual aim of linear regression is to use data to construct an estimate of the form $\widehat{y} = \langle \boldsymbol{\beta}, \boldsymbol{x} \rangle$ where $\boldsymbol{x}$ is a vector of observations. This is done with the aid of historical observations denoted $\boldsymbol{y}$ and a data matrix $\boldsymbol{X}$, consisting of an $M \times R$ matrix of M observations of R regressors.

The traditional approach to regression is based on the method of least squares. In undergraduate econometrics, students learn about the assumptions that give this method a maximum likelihood interpretation: iid observations, Gaussian noise, and so on. Later they learn methods adapted to other statistical assumption. Often the methods result in changing the likelihood, regularizer, or both. In recent years a number of approaches to regression emphasize l1-regularization. These include the Dantzig selector, the LASSO algorithm ridge regression, and other variants. See Hastie et al. (2001) for a survey. One of the advantages seen by this type of regularization is that it promotes sparsity in the resulting predictor. This is seen as potentially useful in reducing the number of regressors in an era of ever-larger data sets. This form of regularization is also popular in the related area of compressed sensing. Recent surveys of the field include Candes (2006); DeVore (2007); Donoho (2006); Tsaig and Donoho (2006).

This section presents a short review of selected regression techniques, some old, and a new one, the Dantzig Selector, proposed in Candes and Tao (2007). A pattern of construction will become clear. Following that, a variation based on generalized entropy is introduced. The Dantzig selector can be seen as a limiting case of the new model. [Experiments and regularization paths should be mentioned here.]

In order to fit regression into an 'entropic' setting a little work is required. Normally the vector $\boldsymbol{\beta}$ is not restricted to the non-negative orthant. In order to utilize the generalized maxent framework, which for most generalized entropies typically produces a non-negative vector we can employ a '$\pm$' trick described in Kivinen and Warmuth (1997) and other places. This is simply a linear map that distinguishes the positive and negative coefficients in $\boldsymbol{\beta}$, where $\boldsymbol{\beta}^+$ and $\boldsymbol{\beta}^-$ denote the positive and negative parts of $\boldsymbol{\beta}$. Let

$$\boldsymbol{p} = \begin{pmatrix} \boldsymbol{\beta}^+ \\ \boldsymbol{\beta}^- \end{pmatrix}.$$

$\boldsymbol{\beta}^+$ and $\boldsymbol{\beta}^-$ denote the positive and negative parts of $\boldsymbol{\beta}$.

Then we can write a fitted values as

$$\widehat{\boldsymbol{y}} \approx \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}_{\pm}\, \boldsymbol{p},$$

where $\boldsymbol{X}_{\pm} := (\boldsymbol{X} \mid -\boldsymbol{X})$, whichever form is more convenient. The cost of this

trick is a doubling of the length of the parameter vector. Whether this fact is significant or trivial depends on the application.

5.4.1 Review of Least Squares

Least squares: The best known linear predictor is the least squares estimator, which will be denoted $\boldsymbol{\beta}_{ls}$. The least squares estimator gets its name from the fact that it solves the following problem:

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \frac{1}{2} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 .$$

Assuming $\boldsymbol{X}$ is 'tall' ($M > R$) and full (column) rank the solution is given by:

$$\boldsymbol{\beta}_{ls} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}.$$

Fitted value are given by:

$$\boldsymbol{y}_{ls} = \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{b}.$$

The least squares predictor provides a familiar and important point of comparison for almost any other predictor one may wish to analyze.

An important property of $\boldsymbol{\beta}_{ls}$ is the fact:

Lemma 5.7 *The following holds for all $\boldsymbol{\beta}$:*

$$||\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_{ls})||_2^2 \geq ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 - ||\boldsymbol{y} - \boldsymbol{y}_{ls}||_2^2$$

Proof

$$\begin{aligned} ||\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_{ls})||_2^2 &= ||\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y} + \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_{ls}||_2^2 \\ &= \langle \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y} + \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_{ls}, \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y} + \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_{ls} \rangle \\ &= \langle \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}, \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y} \rangle + \langle \boldsymbol{y} - \boldsymbol{y}_{ls}, \boldsymbol{y} - \boldsymbol{y}_{ls} \rangle - 2 \langle \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{y} - \boldsymbol{y}_{ls} \rangle \end{aligned}$$

The last term has

$$\begin{aligned} \langle \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{y} - \boldsymbol{y}_{ls} \rangle &\leq ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2 \, ||\boldsymbol{y} - \boldsymbol{y}_{ls}|| \text{ (Cauchy-Schwarz)} \\ &\leq ||\boldsymbol{y} - \boldsymbol{y}_{ls}||_2^2 \text{ (least squares property)} \end{aligned}$$

Substitution in the previous line gives the result. ■

5.4.2 Three Regressions

Many other regression techniques have been proposed. In this section we briefly cover three, enough to highlight a pattern of construction.

In *Ridge regression* the following problem is solved

$$\underset{\boldsymbol{\beta}}{\text{minimize}}\, \frac{1}{2}\,||\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}||_2^2+\lambda\,||\boldsymbol{\beta}||_2^2\,.$$

The *Lasso* uses a one-norm regularizer instead:

$$\underset{\boldsymbol{\beta}}{\text{minimize}}\, \frac{1}{2}\,||\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}||_2^2+\lambda\,||\boldsymbol{\beta}||_1\,.$$

Both of these methods are covered in Hastie et al. (2001). The next is newer. The *Dantzig Selector* (Candes and Tao, 2007) is solution to:

$$\min_{\beta}||\boldsymbol{\beta}||_1 \text{ subject to } \left|\left|\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right|\right|_\infty<\epsilon.$$

The regularizer is an l1-norm, but the statistical loss is merely a constraint on the largest error, a rotated one at that. Two of the unique aspects of this method are the use of the sup-norm, and the fact that its inspiration comes from results in compressed sensing. One feature of that literature is that there is always something that corresponds to the 'true' model. Also Candace and Tao show that under special conditions of low noise and other conditions on the covariance of the regressors, the method can recover the exact true set of regressors with high probability. This result is interesting, but our concern at the moment is to see how the procedure might or might not relate to generalized maxent.

In the table below, the three objective functions are summarized. In addition the statistical loss ($\mathcal{L}(\boldsymbol{\beta},\boldsymbol{y})$) is restated in terms of the deviation from the least squares predictor ($\mathcal{L}(\boldsymbol{\beta},\boldsymbol{\beta}_{ls})$), which is obviously a data-dependent quantity.

The table summarizes what we have so far[2]:

5.5 q-version of Dantzig

The Dantzig selector can be approached in the limit via a model based on generalized maxent. Let $\boldsymbol{A}=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}_{\pm}$, $\boldsymbol{b}=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$. Let $Q>0$, and $0<q<1$. Then the Dantzig selector can be obtained as a limiting case of models of the form:

$$\Delta S_q(\boldsymbol{p},\boldsymbol{p}_0)+\delta_\epsilon(\boldsymbol{A}-\boldsymbol{b}).$$

[2]The reformulated Lasso problem using the loss given in the table is equivalent because of Lemma 5.7. The original loss has been replaced with an upper bound and the constant term dropped.

Name	$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{y})$	$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\beta}_{ls})$	$\mathcal{R}(\beta)$
Ridge regression	$\frac{1}{2}\left\|\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right\|\|_2^2$	$\frac{1}{2}\left\|\|\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_{ls})\right\|\|_2^2$	$\lambda \left\|\|\boldsymbol{\beta}\right\|\|_2^2$
Lasso	$\frac{1}{2}\left\|\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right\|\|_2^2$	$\left\|\|\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_{ls})\right\|\|_2^2$	$\lambda \left\|\|\boldsymbol{\beta}\right\|\|_1$
Dantzig selector	$\delta_{\epsilon\mathcal{B}_\infty}(\, \boldsymbol{X}^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\,)$	$\delta_{\epsilon\mathcal{B}_\infty}(\, \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_{ls})\,)$	$\left\|\|\boldsymbol{\beta}\right\|\|_1$

This is nothing more than the problem discussed in Chapter 3. Letting $q \to 0$, and $Q- > \infty$, recovers the Dantzig selector problem.

Candace and Tao found the Dantzig selector to be optimal for certain special situations. Could it be 'good' for situations outside of the ones they consider? We hope to investigate this experimentally in the future.

The next chapter presents another application of the framework, and follows it through he experimental setting.

Chapter 6

Non-Negative Matrix Factorization

Reduction of dimensionality is a common strategy in machine learning. In recent years a complementary strategy of achieving *sparse* models has received attention, especially in the compressed sensing and l1-regularization literature (see e.g. Cohen et al. (2007)). The non-negative matrix factorization problem provides an opportunity to employ both of these strategies simultaneously. This chapter shows that the entropy functions described in earlier chapters can play a role in constructing and interpreting the models developed here.

In the non-negative matrix factorization problem, the goal is to approximate a data matrix $\boldsymbol{V}$ that has elements that are all non-negative. If $\boldsymbol{V}$ is $N \times T$, the hope is that it can be approximated by a product:

$$\boldsymbol{V} \approx \boldsymbol{W}\boldsymbol{H},$$

where $\boldsymbol{W}$ is an $N \times M$ matrix and $\boldsymbol{H}$ is a $M \times T$ matrix. The columns of matrix $\boldsymbol{W}$ are often thought of as basis vectors, while the columns of $\boldsymbol{H}$ are vectors of weights. There are many ways to devise such approximate decompositions. For example one could start with the problem:

$$\underset{\boldsymbol{W}\in\mathbb{R}^{N\times M},\boldsymbol{H}\in\mathbb{R}^{M\times T}}{\text{minimize}} \left|\left|\boldsymbol{V}-\boldsymbol{W}\boldsymbol{H}\right|\right|^2. \tag{6.1}$$

By further constraining the values of $\boldsymbol{W}$ and $\boldsymbol{H}$ to be non-negative, we arrive at the original NMF problem (Paatero and Tapper, 1994; Lee and Seung, 1999). The originators suggested that the resulting solution was more amenable to the interpretation of 'construction by parts' in contrast to what might be called the 'superposition of signals' yielded by methods such as vector quantization (VQ), principal component analysis (PCA) and independent component analysis (ICA) and other methods (Hyvarinen et al. (2001); Plumbley (2003); Srebro and Jaakkola (2003)). In other respects the goal of all these methods is similar in that we hope to extract something from the data matrix that is simple (low rank)

yet useful.

Many variations on the NMF problem have appeared (e.g. Paatero, 1997; Tropp, 2006; Plumbley, 2003; Pascual-Montano et al., 2006; Tropp et al., 2006; Dhillon and Sra, 2005; Cichocki et al., 2006; Hoyer, 2004). Applications include image analysis, test processing, polyphonic music transcription and others (Feng et al., 2002; Lee and Seung, 1999; Jakulin and Buntine, 2004; Griffiths and Steyvers, 2003; Shahnaz et al., 2006). In contrast there are only a few papers on theory (Donoho and Stodden, 2003), and a few others on the convergence properties of existing algorithms (e.g. Heiler and Schnörr, 2006; Lee and Seung, 2001; Dhillon and Sra, 2005).

In the NMF literature, research has focused on three main threads. One is the suggestion of new distance measures. For example the norm in (6.1) can be replaced by other measures, such as (unnormalized) KL divergence (Cichocki et al., 2006; Dhillon and Sra, 2005; Hoyer, 2004). These can be motivated by various means, but an important observation is that norms do not incorporate the sign of the data, while entropy based distance measures do. A second thread involves regularization methods for $\boldsymbol{W}$ and $\boldsymbol{H}$, or alternatively constraints of some kind. For example it seems clear the 'construction by parts' interpretation might be more easily discerned if the resulting solutions for $\boldsymbol{W}$ and $\boldsymbol{H}$ are also sparse, *i.e.,* containing a significant proportion of zeros. In Hoyer (2004), a constraint designed to promote and control the sparsity of the resulting solution was introduced. A third thread involves solving the optimization problem itself. The objective function in (6.1) is non-convex, regardless of any constraints or regularization we may impose. However it is easy to see that holding $\boldsymbol{W}$ or $\boldsymbol{H}$ fixed, results in a convex function. All the known algorithms therefore concentrate on some variation of alternating minimization with respect to $\boldsymbol{W}$ and $\boldsymbol{H}$. Depending on the exact formulation it can be shown that this converges to a local minimum. However the resulting solution is heavily subject to the initial conditions and exact procedure of the algorithm.

6.0.1 The Model

The algorithm proposed in this chapter finds a local minimum of the following objective function:

$$\Delta S_L(\boldsymbol{W}\boldsymbol{H}, \boldsymbol{V}) + \lambda_W \Delta S_H(\boldsymbol{H}, \boldsymbol{H}_0) + \lambda_H \Delta S_W(\boldsymbol{W}, \boldsymbol{W}_0). \tag{6.2}$$

This objective function is composed of three Bregman divergences, each based on a different version of the q-entropy discussed in Chapter 4. Here the entropy is applied element-wise to the matrix valued arguments. The parameter q_L selects the distance measure used for measuring deviations from the data (statistical loss). The focus here is on values of q_L that are less than one since these can

accommodate elements equal to zero. Algorithms which employ KL-divergence must adjust the data to accommodate the objective function. The values of q_W and q_H are also in the interval (0,1). The values of λ_W and λ_H must be 'tuned' in real applications. In general they should depend inversely on the size of M if they are to have any influence over the solution.

The objective function is used in a 'ping-pong' style procedure which alternatively solves for $\boldsymbol{H}$ and $\boldsymbol{W}$, described in Figure 6.1. At each step the algorithm requires consideration of one of the following two problems:

$$\begin{aligned}
&\underset{\boldsymbol{W}}{\text{minimize}}\, \Delta S_{q_L}(\boldsymbol{W}\boldsymbol{H}, \boldsymbol{V})) + \lambda_W \Delta S_{q_W}(\boldsymbol{W}, \boldsymbol{W}_0)) \quad \text{(Problem } \boldsymbol{W}\text{)} \\
&\underset{\boldsymbol{H}}{\text{minimize}}\, \Delta S_{q_L}(\boldsymbol{H}\boldsymbol{W}, \boldsymbol{V})) + \lambda_H \Delta S_{q_H}(\boldsymbol{H}, \boldsymbol{H}_0)) \quad \text{(Problem } \boldsymbol{H}\text{)}
\end{aligned}$$

These result from ignoring the parts of (6.2) held fixed. One noteworthy feature of the algorithm is that upon returning to a W- or H-step, the initial guess is replaced by the previous intermediate value. This tends to promote convergence in practice.

6.0.2 Dual Characterization

Problems H and W of the previous section are 'nice' in the sense of Chapters 2 and 3. Although the actual algorithm will employ the primal method, it is useful to characterize the solution by looking at the dual problem. Since the problems are almost entirely symmetric in notation, we will concentrate on Problem W.

Let $\bar{\boldsymbol{V}}^*$ denote the solution to dual problem, $\boldsymbol{W}_0^* = s'_{q_L}[\boldsymbol{W}_0^*]$, with the scalar s'_{q_L} function applied element-wise to the matrix. It can be shown (details of the derivation are given in Appendix 6.4) that the primal solution can be characterized as

$$\bar{\boldsymbol{W}} = s^{*'}_{q_W}[\boldsymbol{V}^*\boldsymbol{H}^{\mathrm{T}} + \boldsymbol{W}_0]. \tag{6.3}$$

Thus at each step, $\boldsymbol{W}$ is the image of a q-exponential function. As we saw in Chapter 4, low values of q_W increase the potential for zero values to appear in the solution. Additional intuition about why this occurs can be gained by reviewing Figure 4.1. Since the gradient of the entropy function is finite at the origin, the origin is never infinitely far away from the initial guess $\boldsymbol{W}_0$. If the dual optimal values are negative enough, then the $\boldsymbol{W}$ will be set to zero via (6.3).

This approach can be compared with that of Hoyer (2004). In that paper, sparsity is achieved by enforcing an equality constraint on non-linear sparsity measures for $\boldsymbol{W}$ and $\boldsymbol{H}$, which uses their 1-norm and 2-norm as arguments. In experiments, the approach taken here is clearly influenced by the parameter q_W, but control of the exact amount of sparsity achieved is subject to interaction with

input : $\boldsymbol{V}, \boldsymbol{W}_{-1}, q_L, q_W, q_H, \lambda_W, \lambda_H, \mathsf{itmax}, \mathsf{tol}$
$\mathsf{t} \leftarrow 0$
while $t < itmax$ **do**
  #W-step:
  $\boldsymbol{W}_0 \leftarrow \boldsymbol{W}_{t-1}$
  $\boldsymbol{H}_t \leftarrow \boldsymbol{H}_{t-1}$
  $\boldsymbol{W}_t \leftarrow$ argmin Problem W with parameters: $\boldsymbol{H}_t, \boldsymbol{W}_0, q_L, q_W, \lambda_W$.
  $\mathsf{loss}_t \leftarrow \Delta S_L(\boldsymbol{W}_t \boldsymbol{H}_t, \boldsymbol{V})$
  $\mathsf{regW}_t \leftarrow \Delta S_W(\boldsymbol{W}_t, \boldsymbol{W}_0)$
  $\mathsf{regH}_t \leftarrow \mathsf{regH}_{t-1}$
  $\mathsf{improvement} \leftarrow \mathsf{loss}_t - \mathsf{loss}_{t-1} + \lambda_W(\mathsf{regW}_t - \mathsf{regW}_{t-1})$
  if $|improvement| < tol$ **then** break
  $\mathsf{t} \leftarrow \mathsf{t} + 1$
  #H-step:
  $\boldsymbol{H}_0 \leftarrow \boldsymbol{H}_{t-1}$
  $\boldsymbol{W}_t \leftarrow \boldsymbol{W}_{t-1}$
  $\boldsymbol{H}_t \leftarrow$ argmin Problem H with parameters: $\boldsymbol{W}_t, \boldsymbol{H}_0, q_L, q_H, \lambda_H$.
  $\mathsf{loss}_t \leftarrow \Delta S_L(\boldsymbol{W}_t \boldsymbol{H}_t, \boldsymbol{V})$
  $\mathsf{regH}_t \leftarrow \Delta S_W(\boldsymbol{H}_t, \boldsymbol{H}_0)$
  $\mathsf{regW}_t \leftarrow \mathsf{regW}_{t-1}$
  $\mathsf{improvement} \leftarrow L_t - L_{t-1} + \lambda_H(\mathsf{regH}_t - \mathsf{regH}_{t-1})$
  if $|improvement| < tol$ **then** break
  $\mathsf{t} \leftarrow \mathsf{t} + 1$
end
$\mathsf{objval} \leftarrow \mathsf{loss}_t + \lambda_w \mathsf{regW}_t + \lambda_H \mathsf{regH}_t$
$\bar{\boldsymbol{W}} \leftarrow \boldsymbol{W}_t$
$\bar{\boldsymbol{H}} \leftarrow \boldsymbol{H}_t$
return $\bar{\boldsymbol{W}}, \bar{\boldsymbol{H}}, objval$

Figure 6.1: Ping-pong algorithm

the other parameters of the problem.

6.0.3 Experimental Setup

To examine the performance of the algorithm and the characteristics of its output, experiments on several data sets were performed. Each of these data sets will be introduced in turn. Code for all the experiments was written in Python and C, and made use of several libraries.

Elefant (Efficient Learning, Large-scale Inference, and Optimization Toolkit), currently under development by the Statistical Machine Learning Group at NICTA, is an open source library for machine learning licensed under the Mozilla Public License (MPL). It contains a wrapped version of two numeric libraries PETSc

and TAO which were then used to solve Problems H and W as described below. Further information is available at elefant.developer.nicta.com.au.

PETSc, (pronounced petsee) is a suite of data structures and routines for the scalable, parallel solution of scientific applications (Balay et al. (1997, 2004, 2001)). The main focus of PETSc to provide data types and operations for sparse matrix/vector calculations used by its suite of specialized solvers. Other packages are built on top of the operations provided by PETSc. The work here makes use of one of these: the Toolkit for Advanced Optimization (TAO).

TAO is aimed at the solution of large-scale optimization problems on high-performance architectures. One of the attractions of TAO is its use of PETSc data types to support optimization on both single-processor and parallel architectures. Most importantly for this work TAO has very good implementations for unconstrained and bound-constrained optimization. For the work in this chapter the TAO routine blmvm (bounded, limited memory variable metric method) was used. This algorithm is a quasi-Newton method with bound constraints, quite similar to B-LBFGS Benson et al. (2007). The default line search method (Moré and Thuente, 1994) was used.

The results were generated by high-level code in Python to loop through experiments and save data structures between calls to the blmvm routine in TAO. TAO relies on PETSc data structures which were initialized and managed by the author's code with the assistance of some PETSC and TAO wrappers from Elefant. The objective function and callback routines were implemented by the author in the form of a dynamically loadable callback function library written in C [1] Most of the post-processing of the results was accomplished using Python, with the aid of user-written and outside modules, most notably scipy and pylab.

6.1 ORL Face Data Set

One application of the NMF problem is in the area of decomposition of images. This task provides a useful first case, because the visualization of the results is very natural.

The ORL face data set (www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html) consists of facial images of 40 individuals in 10 different poses, making a total of 400 images. Figure 6.2 shows a sample image of each subject, providing a sense of the variation among the individuals, along with the full sequence from the first three individuals.

Initialization of NMF algorithms has been identified as an important issue Langville et al. (2006). To investigate this aspect of the problem, several different

[1] The code also contains hooks for direct calls from Python. These were used for calculating some values before and after the optimization routines were invoked. Also, Python could have been used for the entire task, but would likely have been slower to execute.

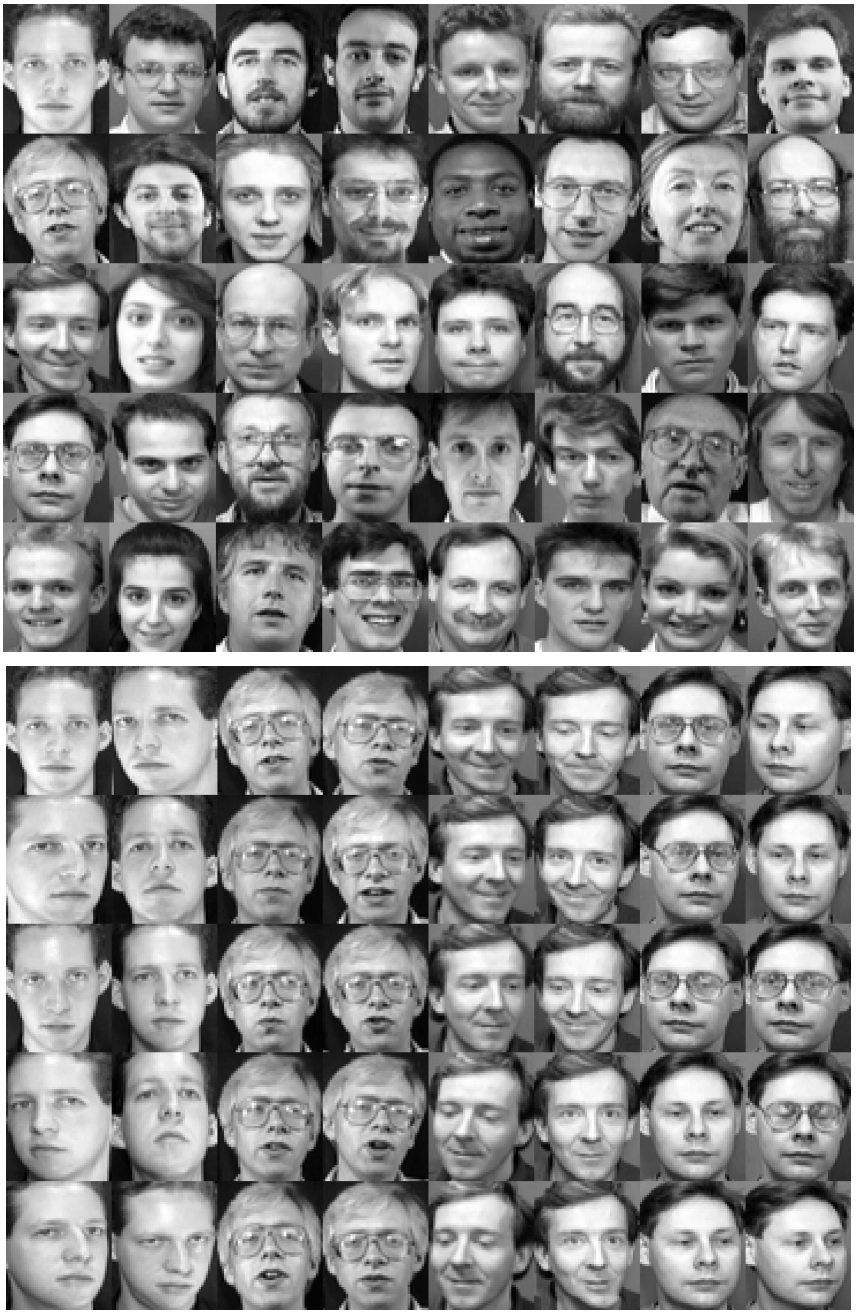


Figure 6.2: The ORL face data set contains 400 image of 40 subjects in various poses. This upper image depicts one photograph of each subject, while the lower one shows a complete sequence of images for the first three subjects

methods of initializing $\boldsymbol{W}_0$ were examined. These are illustrated in Figure 6.3. These consist of (1) patches from the first face in the data set, (2) entire random faces, (3) random pixel values, (4) white patches against a field of black (5) black patches against a field of white, and (6) patches of random values against a field of black. All random values were generated a iid uniform $[0, 1]$. The elements of the initial weighting matrix, $\boldsymbol{H}$, were uniformly set to one. In preliminary runs only the choices 1, 2, and 4 generated sensible results, and therefore only these choices were investigated further.

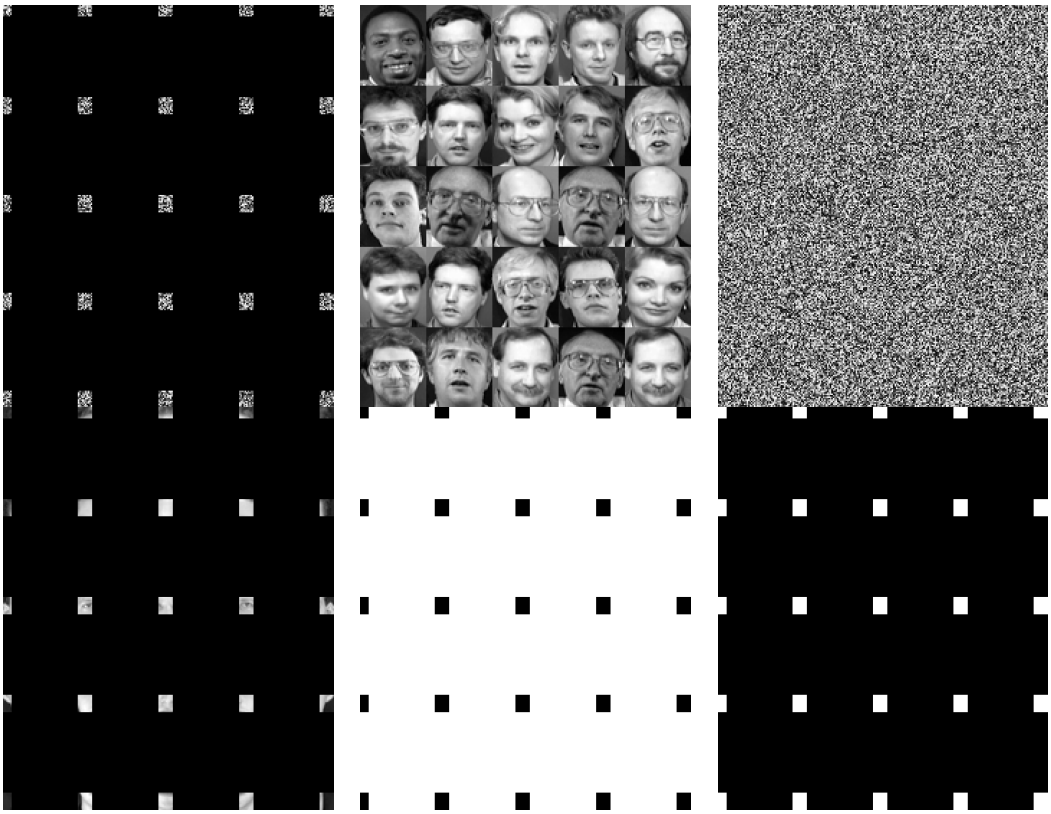


Figure 6.3: Several alternatives for initializing the algorithm were employed. Each image contains 25 subimages which were used as an initial basis. Clockwise from upper left: (1) patches from the first face in the data set, (2) entire random faces, (3) random pixel values, (4) white patches against a field of black (5) black patches against a field of white, (6) patches of random values against a field of black. All random values are uniform $[0, 1]$.

A typical experiment summary is illustrated in Figure 6.5. This single run helps to illustrate a number of general points. Eventually hundreds of such summaries were assembled in a single document, enabling them to be quickly evaluated for inclusion in this discussion. In the upper left panel, the total objective function is shown at each iteration as well as the cumulative number of func-

tion/gradient evaluations. For the most part this latter series increases linearly. Adequate rates of convergence were obtained by limiting the number of blmvm iterations TAO was allowed to take by setting the parameter `tao_max_its`. Additional evaluations for line searches, and projections associated with the non-negativity constraint were infrequently taken.

Overall, the objective function tended to be dominated by the statistical loss function. In the middle left panel the regularization contributions are shown. The jagged appearance of these series is to be expected as the alternate in the influence on the overall objective. Additionally the contribution from the basis matrix $\boldsymbol{W}$ was generally small even after increasing its weight. In the ping-pong algorithm employed, the initial guess $\boldsymbol{W}_0$ is replaced with the result from the previous step. The small values on the left-hand axis are typical, with the implication that the basis matrix is only adjusted slightly after the first iteration. Although there is some movement, it was difficult to visually distinguish between the first and final choice of $\boldsymbol{W}$ by the algorithm in any of the runs.

The lower left panel reports the input variables, along with some output statistics. The output statistics include the 2-norm, and 10-norm deviation between the fitted values and the data. The latter was included because it emphasizes large deviations.

In the upper left panel, the fitted faces are shown. In the middle right panel the final basis is depicted. In this case M was set to 25, so the image contains 25 subimages, one for each basis vector. These appear as dark images. Since the weightings for each fit are non-negative, the fitted images are brighter. Additional detail of the basis vectors is discernible when they are shown in the negative. This is represented in the lower right panel. In this run the algorithm found features that are sparse (fraction .53 zeros) and localized spatially.

Nine additional sample runs (Figures 6.6 - 6.13) were included to illustrate some additional observations. Experiments 1-3 allow a comparison of the initialization methods described above. Using patches from the first face creates a set of basis vectors that have highly localized non-zero values. This aspect is inherited by the solution. By contrast, using entire faces (Experiment 2) leads to globalized features, which are not as sparse. They are however far more sparse (.23) than the input values which are essentially all non-zero. The fitted faces appear to be superior and have substantially lower 2- and 10- norm deviations. In Experiment 3 the input basis is essentially a less informative version of Experiment 1, and contains noticeable artifacts of the starting basis vectors. Notably the 2 and 10-norm deviations are *lower* than Experiment 1, casting some doubt on the value of these distance measures in this setting.

A series of experiments with M set to 49 was also run. Experiments 4-5 correspond to Experiments 1 and 2 with only this change. Overall, the fitted faces contain more detail and the resulting basis images are slightly more sparse.

Interestingly, Experiment 3 shows higher 2- and 10-norm values, despite a slight, albeit subjective improvement in fit. Experiment 4 shows highly localized sparse features. In Experiment 5 the features are not as sparse, being somewhat face-like.

Experiments 7-10, are included to exhibit some of the better fits and they also display the sensitivity to the input values of q_W, q_H and q_L. Better fits were obtained with high values of q_L (closer to 1.0) , and low values of q_H and q_W ($< .2$). The high values of q_L effectively penalize large deviations from pixel values near zero (black). Also it was important to keep the values of q_W lower than q_H. In effect this made $\boldsymbol{W}$ the favored choice for achieving sparsity.

In sum, sparse values of $\boldsymbol{W}$ can be achieved, in just a few iterations with a low q_W and the right initialization. Patches of faces and random faces provide 'acceptable' results, visually speaking, but the patches produce greater sparsity in the final basis. Sparsity seems to traded off against statistical error. However the situation regarding locality is not clear cut. The face-like global features are fairly sparse but not localized in Cartesian coordinates. Closer examination of Experiment 2 (middle right panel), or Experiment 10, on the preceding page, reveals that many of the faces are rather grotesque, with missing or distorted facial parts with other parts left somewhat intact.

As final thought on this data set, the basis images are reminiscent of the work of the 19th century Spanish print maker Goya (Figure 6.4) who was noted for his illustration of the human condition via depictions of the grotesque. Perhaps some other 'face space' is the appropriate one in which to judge locality and the artist understood such spaces intuitively.

Figure 6.4: Closeups of two faces by Goya. (www.worldprintmakers.com)

6.1.1 ORL Face Data Experiments

This section contains the experiment summary sheets referred to in the previous section, beginning on the next page.

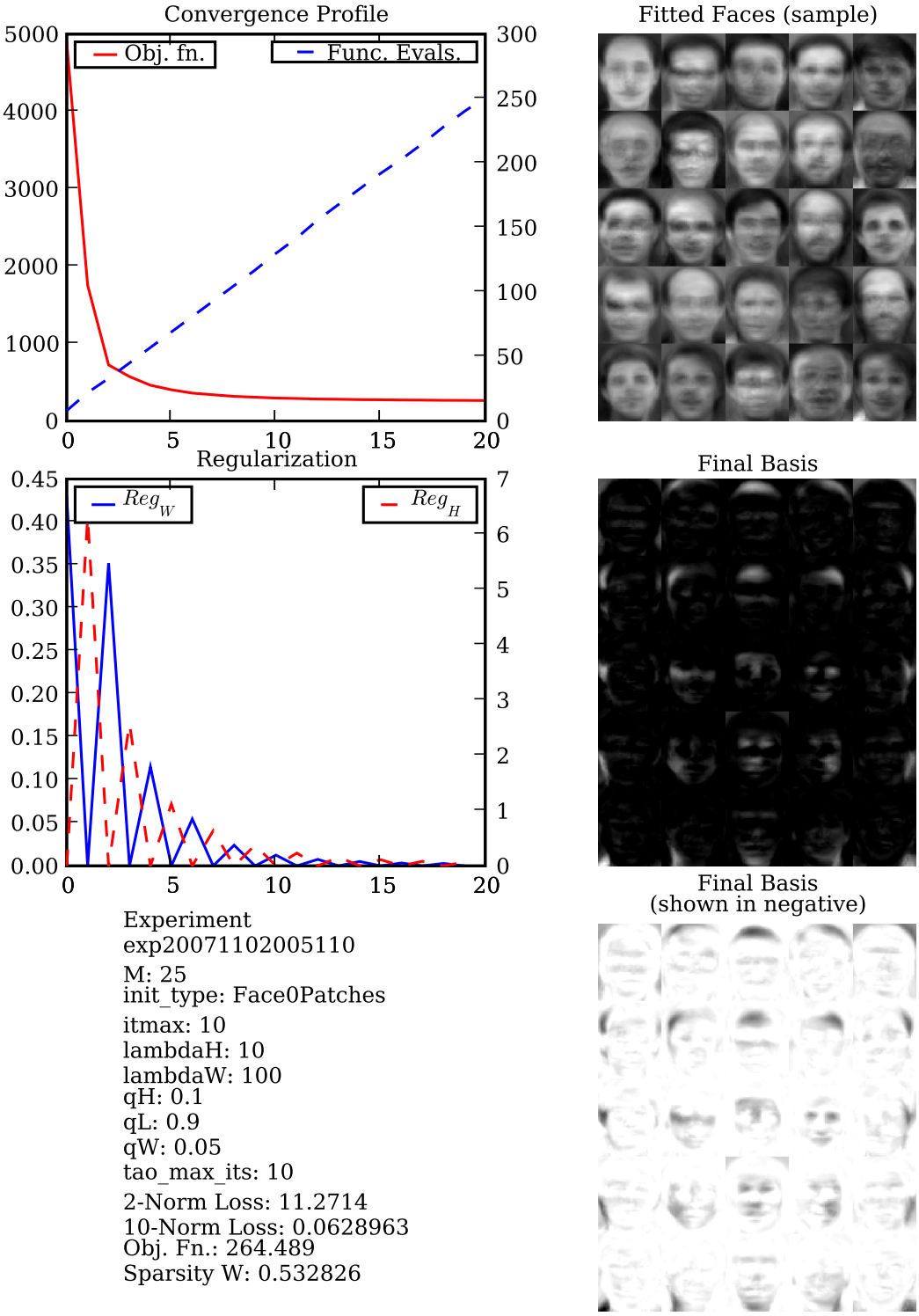


Figure 6.5: Sample Experiment 1

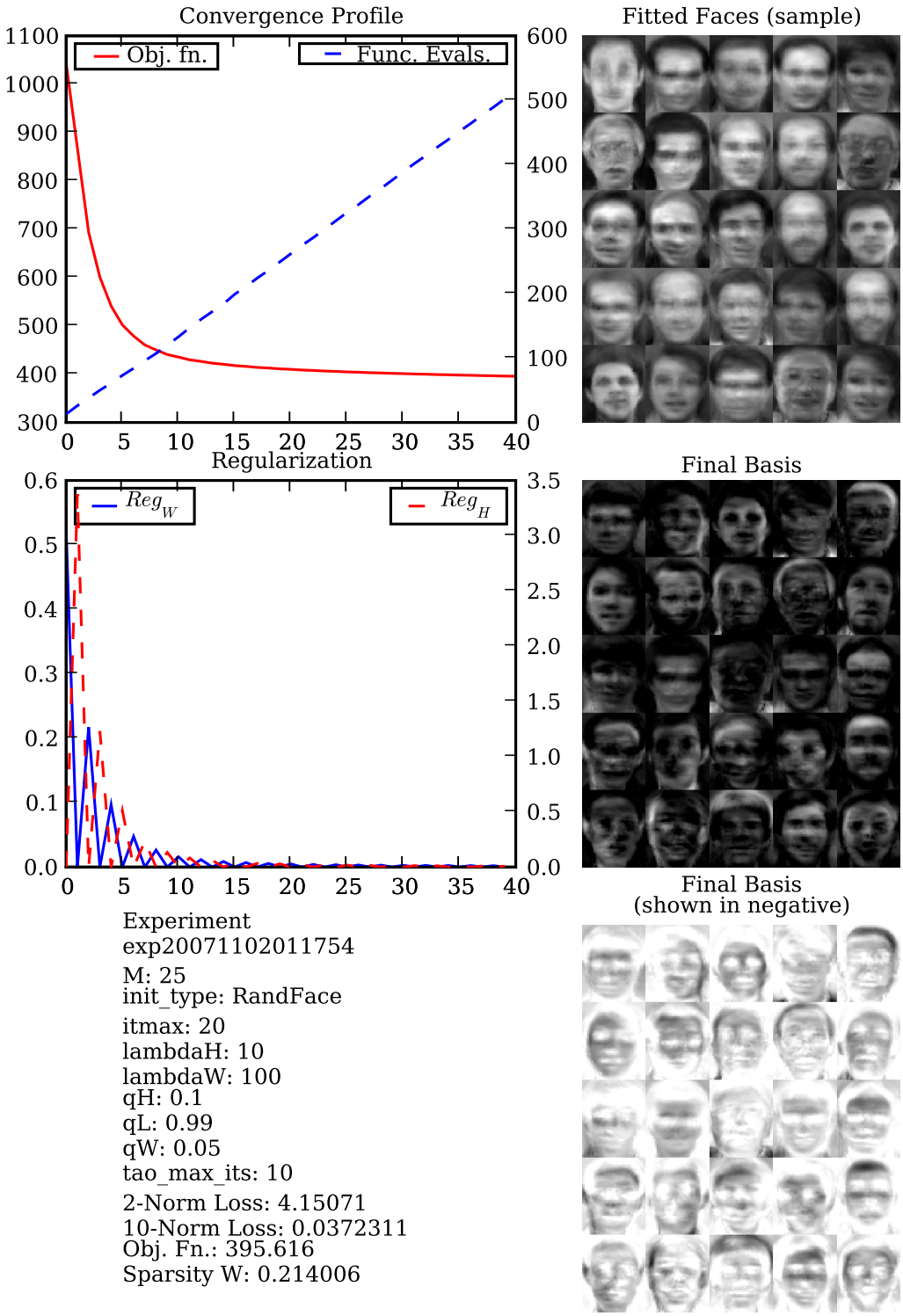


Figure 6.6: Sample Experiment 2

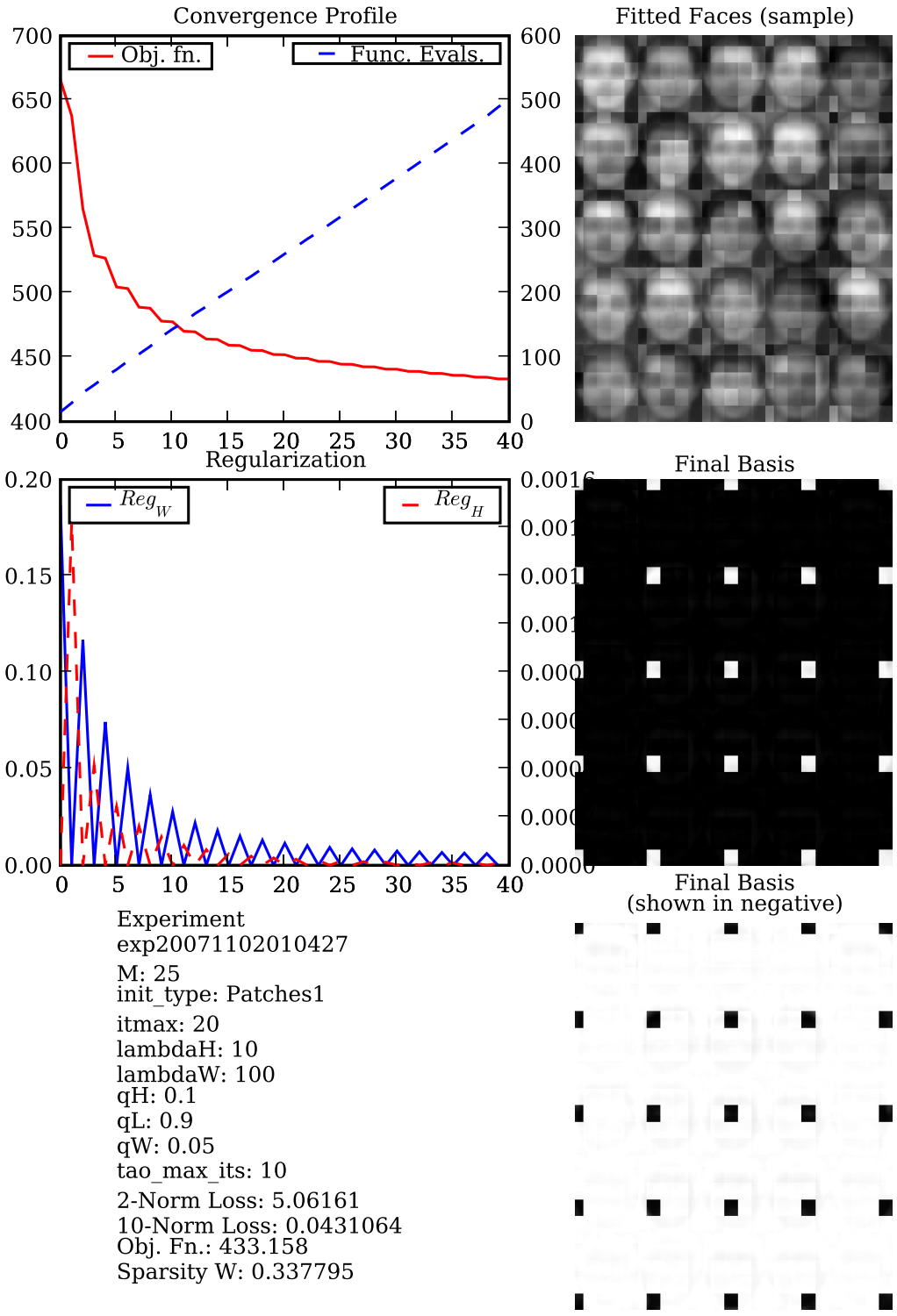


Figure 6.7: Sample Experiment 3

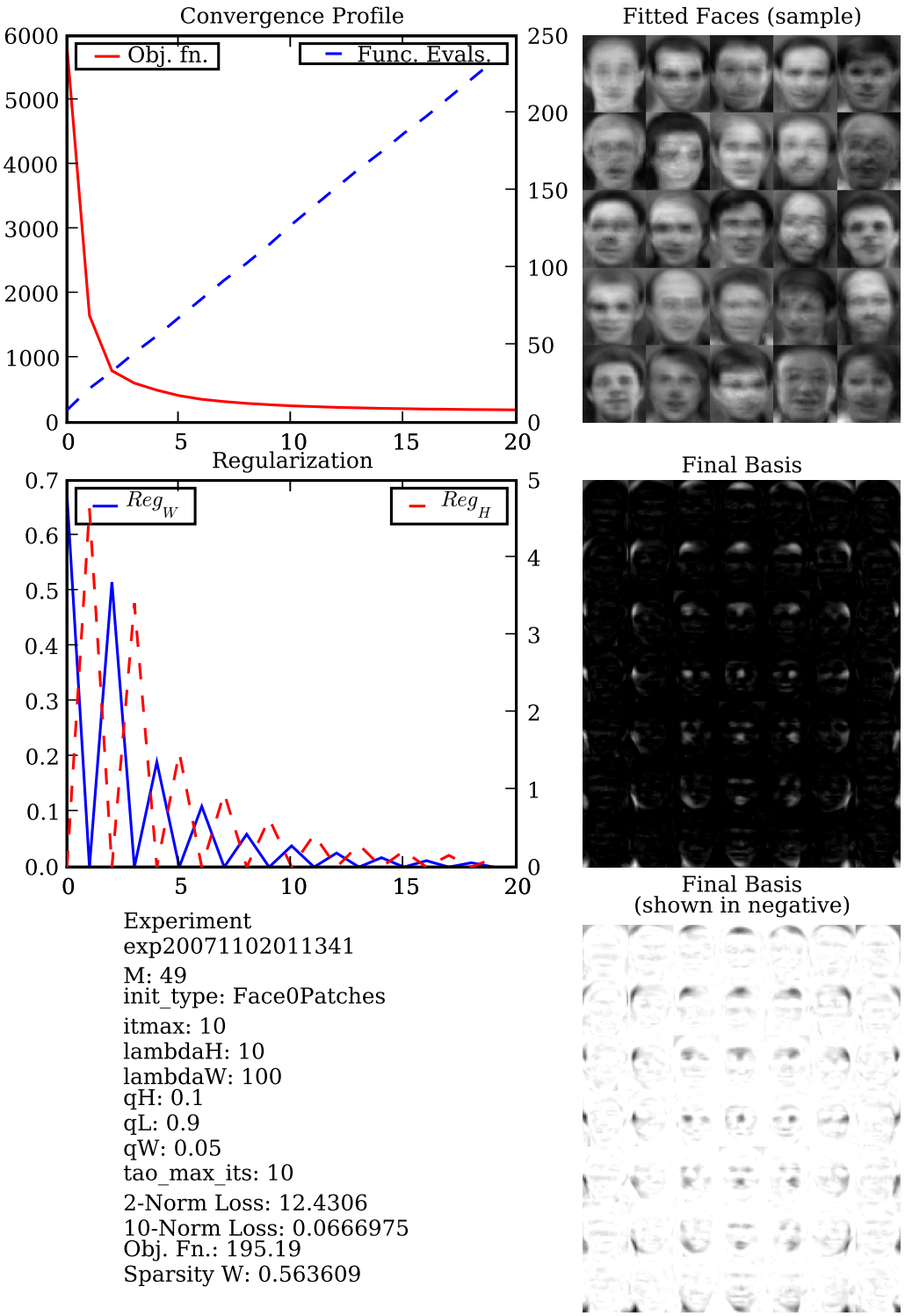


Figure 6.8: Sample Experiment 4

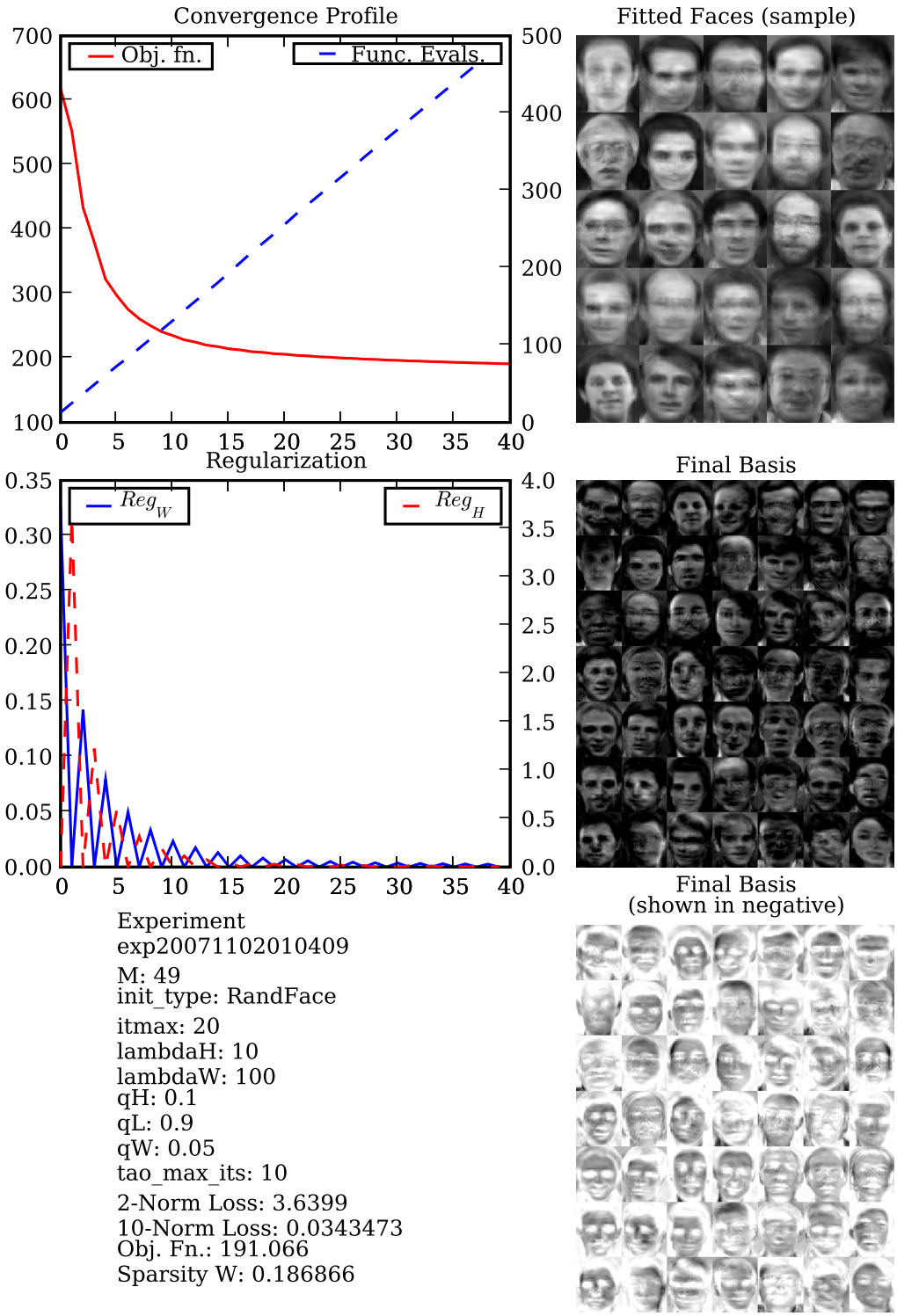


Figure 6.9: Sample Experiment 5

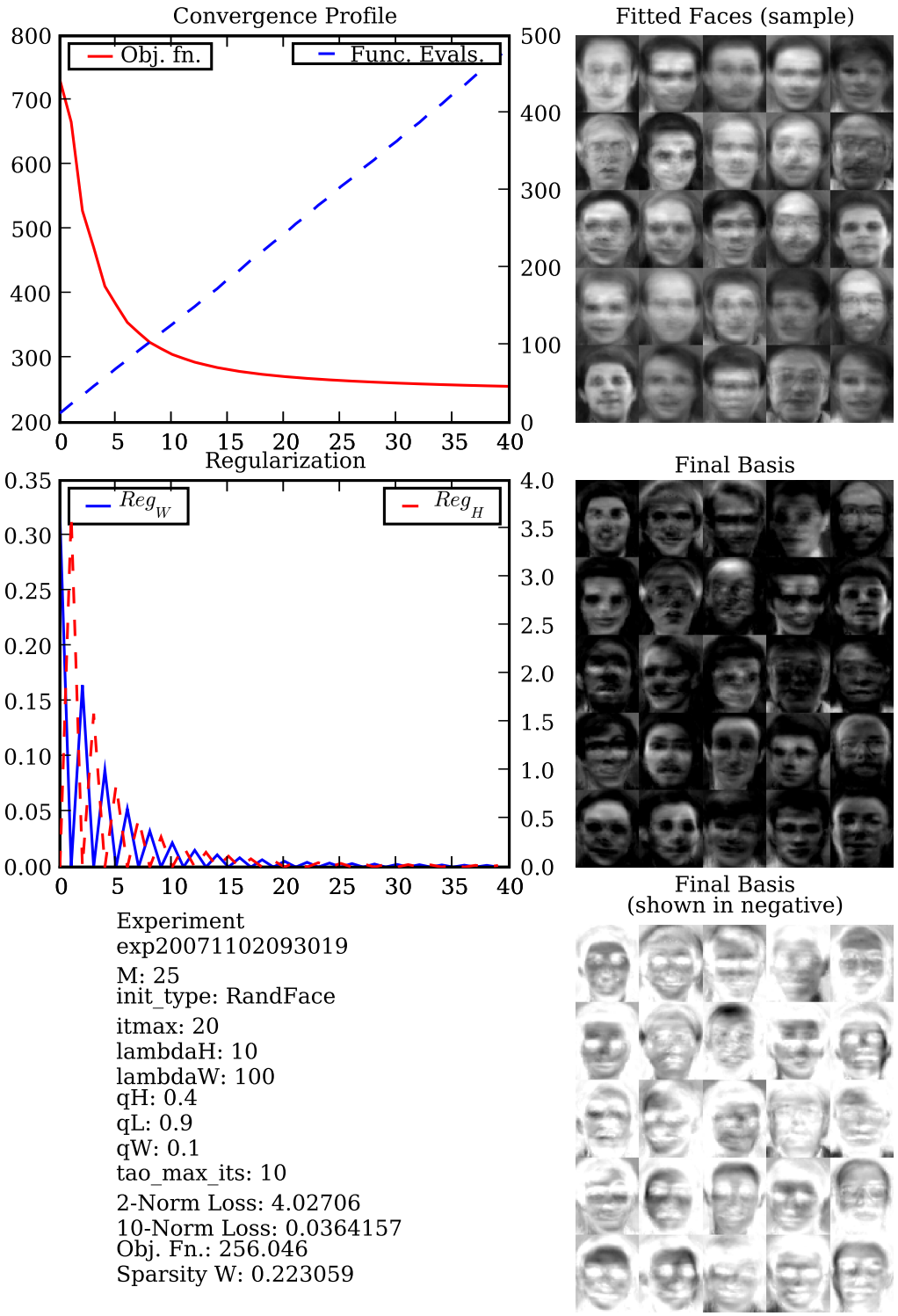


Figure 6.10: Sample Experiment 6

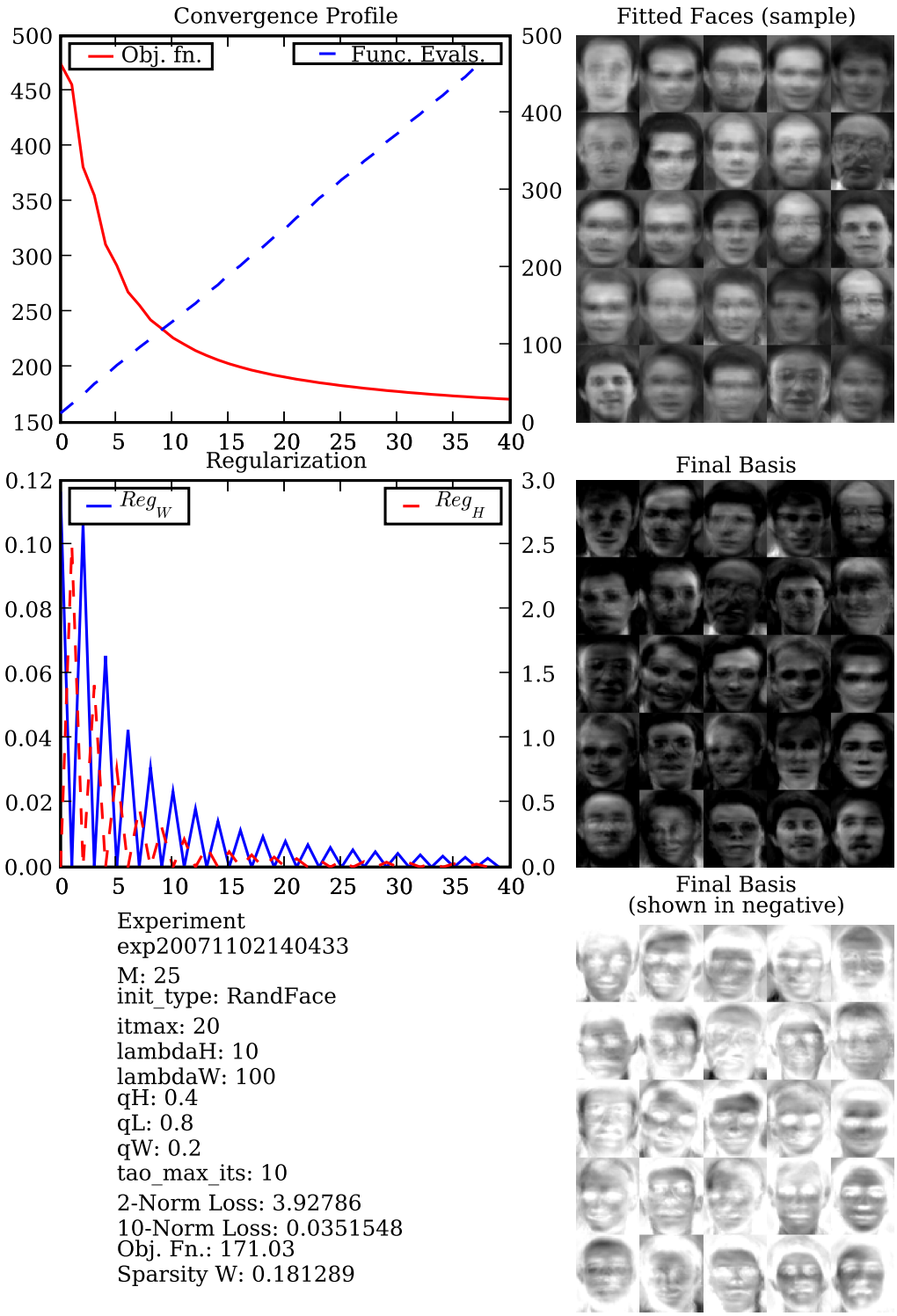


Figure 6.11: Sample Experiment 7

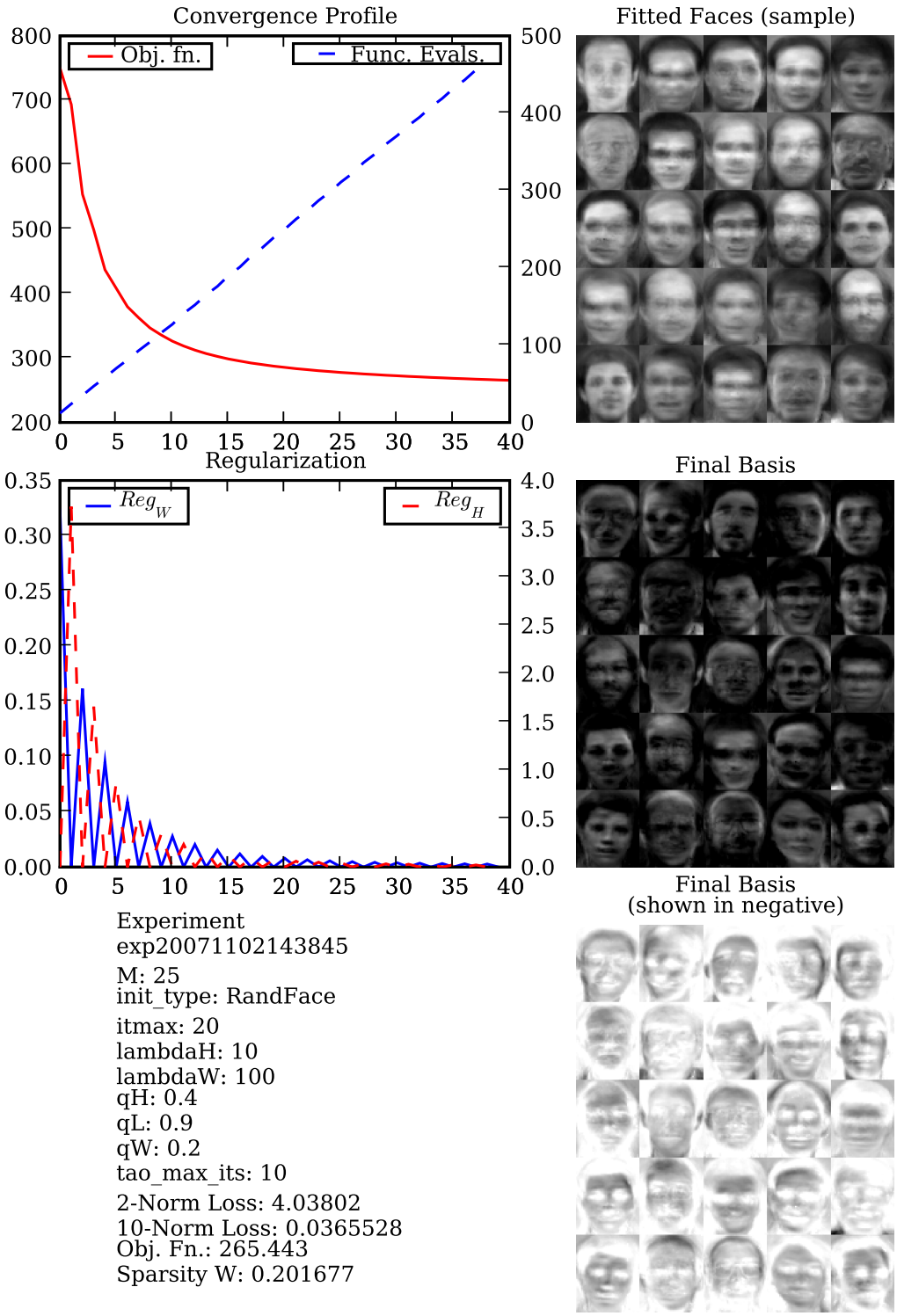


Figure 6.12: Sample Experiment 8

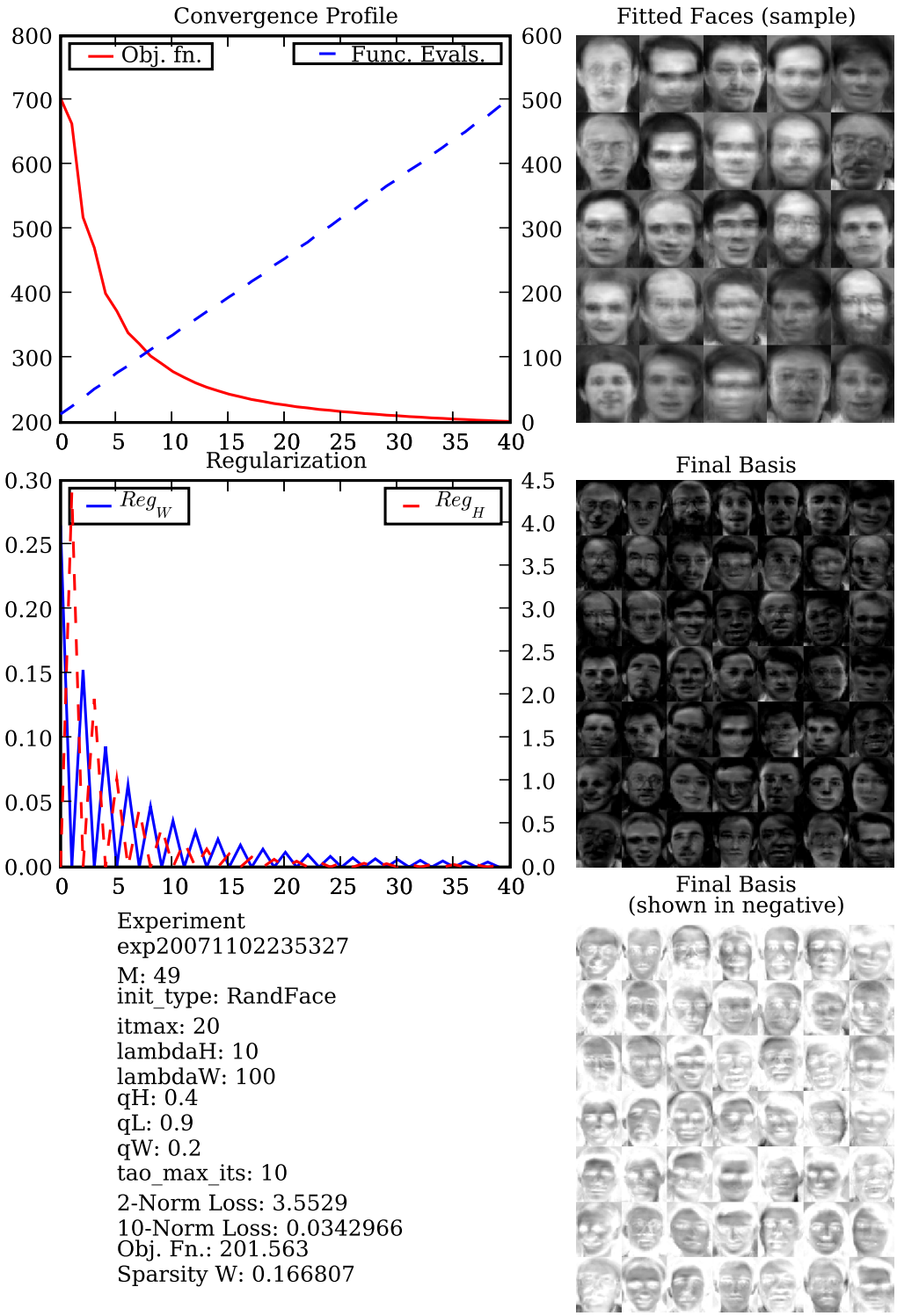


Figure 6.13: Sample Experiment 9

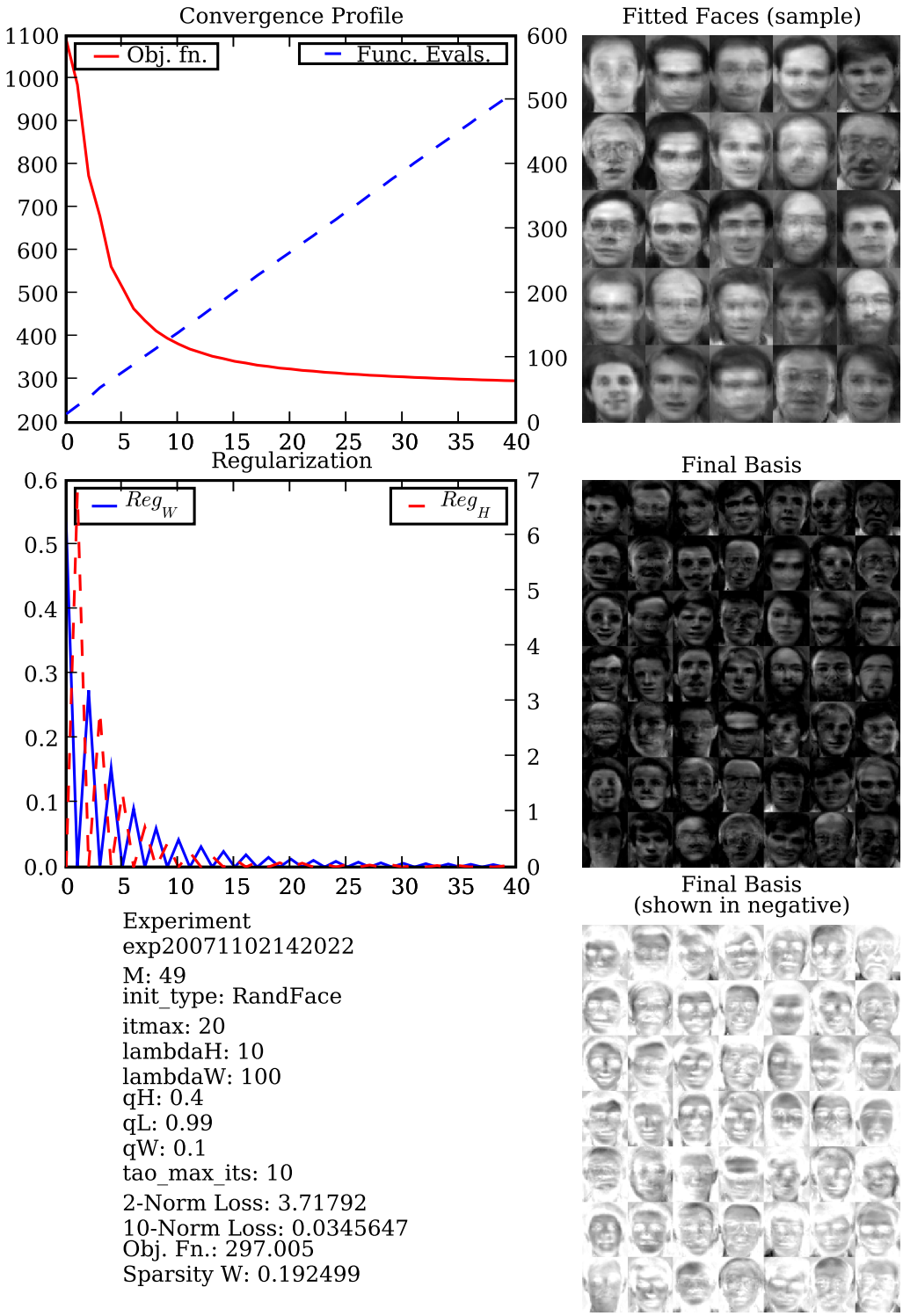


Figure 6.14: Sample Experiment 10

6.2 The 108th U.S. Senate

The THOMAS database of legislative information (thomas.loc.gov/home/rollcallvotes.html) contains the roll call votes for each senator. In a roll call vote, as opposed to a voice vote, the vote of each senator is recorded as 'yea' , 'nay' or not voting. These are coded as +1, 0 and -1, respectively in the THOMAS database. Assume for the moment that there are no absences or abstentions, as these will be addressed later. In this case T = 459 roll call votes and there a N = 100 senators, two for each state. This results in a $100 \times 459 = N \times T$ data matrix, reflecting all the votes of the 108th Senate. Data such as these are fodder for lobbyists, pundits and social scientists alike. They represent an important way to measure individual senators' positions on the important questions of the day and also the behavior of groups in voting blocs. See Jakulin and Buntine (2004) for a wide range of analyses and perspectives on this dataset[2].

For this data set the goal of the NMF problem is to approximate the voting data as:

$$\underbrace{\boldsymbol{V}}_{N\times T} \approx \underbrace{\boldsymbol{W}}_{N\times M}\underbrace{\boldsymbol{H}}_{M\times N}$$

An NMF study of $\boldsymbol{V}$ offers an appealingly simple interpretation of the output, which can be used to perform a variety of tasks.

The input variable M can be considered as the factors which dispose a senator to cast a yea or nay. At the same time, each vote may bear on a combination of the same factors. In the language of politics the factors may be thought of as 'issues', although there is no semantic requirement on the factors. That must be left to interpreters.

In constructing an objective function, one immediate issue arises. The coding of 'yea' as 1 and 'nay' as 0 is arbitrary. Many methods, including the q-entropy-based model of the previous section will see these votes very differently and are not invariant to the opposite encoding. Not only that, in most legislative bodies the nature of the bills and amendments requires casting a rather arbitrary pattern of yeas and nays in order to consistently support a position. The effect on the data can be explained by a simple example. Suppose there are only two senators, Rep and Dem. Rep is 'pro-gun' and Dem is 'anti-gun'. Furthermore there are only two votes in the session. The first is on a bill to arm all citizens and the second is on a bill to ban all bullets. Listing Rep's votes in the first row and

[2]Thanks also to Wray Buntine for supplying a pre-parsed version of the data for this study.

Dem's in the second we get the data matrix:

$$\boldsymbol{V} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

In a sense there is only one issue before this mock senate, but it results in a voting data matrix that is rank-2. The following proposal can address both of the aforementioned issues: Augment the data matrix with a matrix of 'reverse' votes (on fictitious 'anti-bills') and introduce corresponding reverse $\boldsymbol{W}$ in the first matrix of the factorization. This results in the approximation problem:

$$\boldsymbol{V} = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix} \approx \underbrace{(\boldsymbol{W} \mid \widetilde{\boldsymbol{W}})}_{N \times 2M} \underbrace{\boldsymbol{H}}_{2M \times 2 \cdot T}, \tag{6.4}$$

where $\widetilde{\boldsymbol{W}}$ is a matrix of anti-factors, An anti-factor matrix is formed by converting all missing votes to 1/2 and then applying the map $\tilde{v} = |v - 1|$, which flips all the 1s to zeros and vice-versa. The anti-factor matrix is then used to augment the normal one in the initialization. After the initialization the factor/antifactor relationship is not enforced. Also, to calculate the statistical loss component of the objective function all non-votes were ignored. In other respects the objective function and algorithm is identical to that of the previous section.

For this data set, two initialization methods were employed. The first uses small groups of senators that might be presumed to have similar voting patterns, based on very preliminary guesswork. Pairs of senators are selected based on same-party relationships. Several also have same-state or neighboring-state relationships. These are listed in Table 6.1. Included in the pair was one, Kennedy-Santorum, who presumably take opposite stances on many issues. This was done to investigate the stickiness of the initial groups. For each factor a pair of senators had their corresponding entries in $\boldsymbol{W}_0$ set to 1, and the reverse setting was applied to the anti-factor matrix. This method results in an initial sparsity (fraction of entries equal to zero) of .5 for $\boldsymbol{W}_0$. Clearly some, but not very much prior knowledge is required to use the method. The method was tagged as 'seed_pairs_0' in the experimental runs.

An alternative method of initialization of $\boldsymbol{W}_0$ is based on a suggestion from Langville et al. (2006). A random selection (without replacement) of 40 columns of $\boldsymbol{V}$ representing 40 random votes are averaged to construct an initial factor. These factor supplies a column of $\boldsymbol{W}_0$, and the process is repeated as often as required by the input M. An antifactor matrix is constructed and used to augment $\boldsymbol{W}_0$ as described above. The motivation for this method is to use a sample small enough to create variation between the columns, but large enough for averaging, to reveal some common factors. The method is simple to implement, requiring no

prior knowledge, but there is little theory to guide our choice of sample size. The resulting initialization matrix is dense. This method was tagged as 'randomVcol' in the run sheets.

The $\boldsymbol{H}$ matrix is initialized to zero. The first step taken by the algorithm is an H-step, so this choice affects mainly the first regularization, and so reflects a preference for sparse solution values for $\boldsymbol{H}$. In preliminary analysis, this appeared to work better than a positive constant.

Sarbanes (D-MD)	Mikulski (D-MD)
Alexander (R-TN)	McCain (R-AZ)
Byrd (D-WV)	Rockefeller (D-WV)
Craig (R-ID)	Crapo (R-ID)
Schumer (D-NY)	Clinton (D-NY)
Hatch (R-UT)	Bennett (R-UT)
Frist (R-TN)	McConnell (R-KY)
Kennedy (D-MA)	Santorum (R-PA)
Enzi (R-WY)	Thomas (R-WY)
Dayton (D-MN)	Feingold (D-WI)
Roberts (R-KS)	Brownback (R-KS)
Feinstein (D-CA)	Reid (D-NV)

Table 6.1: Seeding the matrix $\boldsymbol{W}_0$ based on pairs of senators with same-party affiliation.

6.2.1 Experiment Results

In order to explore the effect of possible parameter settings a large number of runs were undertaken. A few of the summary sheets appear as Figures 6.15 - 6.23. In the discussion below comparisons are made between them and also to the group of face data experiments discussed in the previous section.

The results of the factorization can be used as input to many post processing tasks such as visualization, clustering and prediction. In the discussion we focus on two such tasks. The first measures the quality of fit based on the ratio of *mistakes* to total votes. Here a raw estimated data vector was calculated as $\boldsymbol{V} = \bar{\boldsymbol{W}}\bar{\boldsymbol{H}}$, this was converted to vote estimates (1 or 0) based on a cutoff of 0.5. The same procedure was used to eliminate missing votes in the actual data matrix. This mistake ratio, along with the 2-norm of the raw error was tracked during the estimation process and is shown in the lower left panel of each experiment run. The mistake ratio is an intuitively sensible statistical error measure for measuring statistical fit, but one that is difficult to optimize directly because it is composed of discrete variables.

The second method uses the $\boldsymbol{W}$ matrix to perform a *hard clustering* of the senators. To perform the clustering, each row of $\boldsymbol{W}$ is normalized, and all but the largest element of each row is zeroed out, yielding a hard assignment based on the strongest factor for each senator, with a weight that is dependent on the original strength of the factor. The magnitude of the weight is also used as a measure of affinity strength to sort the senators within the cluster. It is reflected by the order of their names within each cluster. (The overall order of the clusters has no particular meaning.) The clustering results are given in the lower right panel of each experiment run. See e.g. Figure 6.15.

In the upper left panel, the objective function path and cumulative number of gradient/function calls are given. In the middle left panel, the regularization function values are shown. Both of these panels correspond to similar panels from the face data experiments of the previous section.

In the upper right, the input parameters are given. The value M reflects the number of columns *after* augmentation with anti-vote data, so it is always double the number of the input M. In other respects these extra factors are no different than ordinary factors and no further distinction is maintained.

Comparisons between sample runs serve to indicate a number of findings. The aim is not to find the 'best' run, as this is ultimately task dependent. It is used instead to make comparisons across a number of 'good' runs to examine sensitivity and other issues.

First, the 'randomVcol' slightly outperforms the 'seed-pair' method (Figure 6.15 v. Figure 6.16 are typical.) Deleting the bad pair, Kennedy-Santorum, did not reverse this relationship. Nor did Kennedy and Santorum ever appear in the same cluster, a result which would likely be of comfort to both Senators. Subsequent examples all use the 'randomVcol' method.

The mistake ratios represent an intuitively sensible statistical error measure. In most runs, the mistake ratio curve flattens out well before the objective function. The 2-norm of the error is not optimized and oscillates between H- and W-steps. The lower envelope of the series usually, but not always, has a downward trend (see e.g. Figure 6.16). This effect was also seen before with the face data. The usefulness of the 2-norm measure seems questionable here again, unless it can be motivated as a loss function in its own right.

Fit was improved by increasing the number of factors, although little statistical improvement occurred after about M = 24 (input M=12). One should expect fit with M until $\boldsymbol{W}$ is square or there are no errors. See Figure 6.15 versus Figures 6.17 - 6.19.

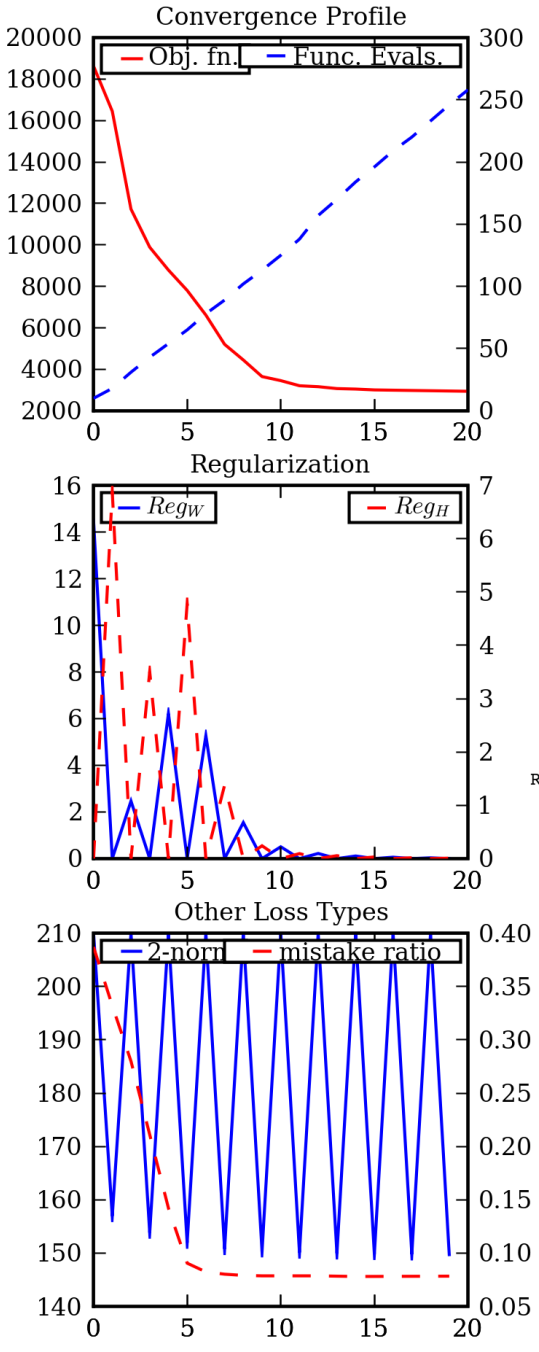
As one would expect, reducing the degree of regularization (λ_H, λ_W) decreased the measures of statistical error (ΔS_{q_L} and the mistake ratio). Comparing Figure 6.17 versus 6.21 indicates a reduction in mistakes by over 40% to close to a .4% mistake ratio.

The hard clustering method produces more clusters as M is increased, but often falls short of utilization all available slots Figure(6.19). Judging the right level of M or the quality of the clusters seems inherently subjective, however a few observations may be pertinent. When the number of clusters was two, Republicans and Democrats were neatly split, with the sole independent, Jeffords clustered with Democrats. Larger numbers of clusters usually included large blocs consisting of all Democrats or all Republicans, with some smaller blocs sometimes containing a mixture. These more obvious findings are in accord with Jakulin and Buntine (2004), but further discussion is effectively out of the scope of the thesis.

A final but important question for this thesis is the effect of the various q_L q_W and q_H values and the connection, if any, between sparsity and performance. Here, good fits were found with relatively low values of all three parameters. Better fits were found with q_W set to values higher than q_L or q_W. Figures 6.22 and 6.23 show that raising q_L to the level of q_W or higher made the fit noticeably worse. A natural interpretation of this situation is that the data matrix contains many zeros. This is problematic for KL-divergence ($q = 1$) and measures approaching it. In the face data set, a sparse basis was more desirable. In the case of the Senate data, more of a mixture is favored. In contrast each vote is more likely, though not guaranteed to concern a single issue. This makes it plausible that $\boldsymbol{W}$ should be less sparse than $\boldsymbol{H}$. All of the 'good' fits shown here reflect that property in the sparsity results given on each summary sheet. Once again the variability of the sparsity is fairly high. In this algorithm direct control of the sparsity is not offered and whether that is ideal remains an open question.

6.2.2 Senate Experiments Summary Sheets

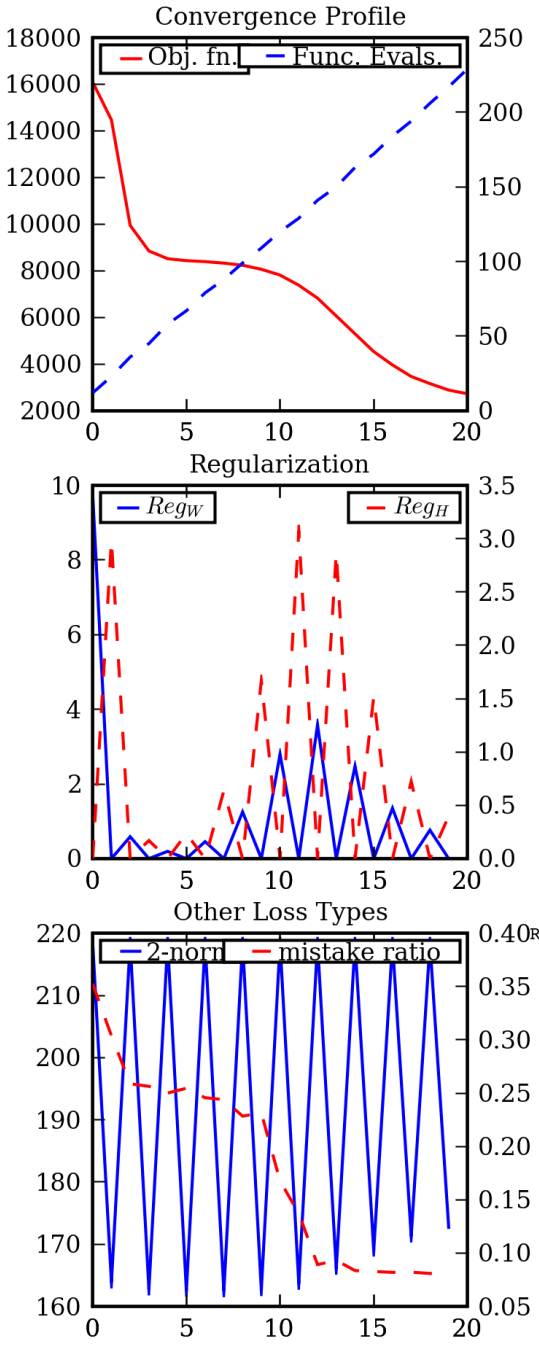
This section contains the experiment summary sheets referred to in the previous section, beginning on the next page.



Experiment exp20071119095613
M: 2
P: 1
init_type: random_Vcol
itmax: 10
lambdaH: 100
lambdaW: 100
qH: 0.05
qL: 0.1
qW: 0.2
tao_max_its: 10
2-Norm Loss: 149.716
1.0-Norm Loss: 38029
Obj. Fn.: 2948.5
Pred. mistakes 7185
Sparsity W: 0.13
Sparsity H: 0.332244

___Cluster 0_____	Bayh (D-IN)	Bennett (R-UT)
Stabenow (D-MI)	Jeffords (I-VT)	Bunning (R-KY)
Levin (D-MI)	Hollings (D-SC)	Alexander (R-TN)
Akaka (D-HI)	Lincoln (D-AR)	Bond (R-MO)
Rockefeller (D-WV)	Carper (D-DE)	Talent (R-MO)
Reid (D-NV)	Inouye (D-HI)	Graham (R-SC)
Cantwell (D-WA)	Landrieu (D-LA)	Brownback (R-KS)
Durbin (D-IL)	Baucus (D-MT)	Stevens (R-AK)
Leahy (D-VT)	Graham (D-FL)	Inhofe (R-OK)
Murray (D-WA)	Breaux (D-LA)	Warner (R-VA)
Clinton (D-NY)	Nelson (D-NE)	Shelby (R-AL)
Reed (D-RI)	Edwards (D-NC)	McConnell (R-KY)
Schumer (D-NY)	Lieberman (D-CT)	Lott (R-MS)
Sarbanes (D-MD)	Kerry (D-MA)	Fitzgerald (R-IL)
Mikulski (D-MD)	___Cluster 1_____	DeWine (R-OH)
Daschle (D-SD)	Cornyn (R-TX)	Coleman (R-MN)
Kohl (D-WI)	Sessions (R-AL)	Hagel (R-NE)
Johnson (D-SD)	Burns (R-MT)	Ensign (R-NV)
Dodd (D-CT)	Craig (R-ID)	Hutchison (R-TX)
Boxer (D-CA)	Chambliss (R-GA)	Voinovich (R-OH)
Corzine (D-NJ)	Cochran (R-MS)	Sununu (R-NH)
Wyden (D-OR)	Enzi (R-WY)	Gregg (R-NH)
Lautenberg (D-NJ)	Frist (R-TN)	Murkowski (R-AK)
Feingold (D-WI)	Grassley (R-IA)	Domenici (R-NM)
Dorgan (D-ND)	Thomas (R-WY)	Smith (R-OR)
Bingaman (D-NM)	Hatch (R-UT)	Campbell (R-CO)
Dayton (D-MN)	Allen (R-VA)	McCain (R-AZ)
Kennedy (D-MA)	Nickles (R-OK)	Specter (R-PA)
Feinstein (D-CA)	Kyl (R-AZ)	Collins (R-ME)
Nelson (D-FL)	Crapo (R-ID)	Snowe (R-ME)
Harkin (D-IA)	Allard (R-CO)	Miller (D-GA)
Byrd (D-WV)	Dole (R-NC)	Chafee (R-RI)
Pryor (D-AR)	Roberts (R-KS)	
Conrad (D-ND)	Santorum (R-PA)	
Biden (D-DE)	Lugar (R-IN)	

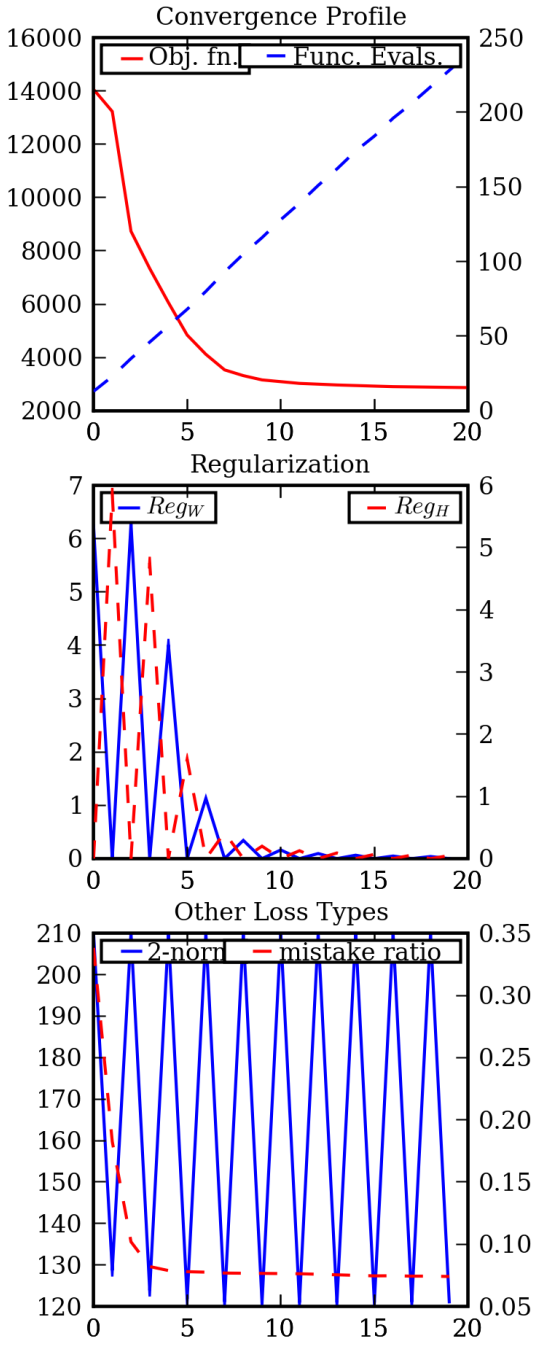
Figure 6.15: Senate Experiment 1



Experiment exp20071119165928
M: 2
P: 1
init_type: seed_pairs_0
itmax: 10
lambdaH: 100
lambdaW: 100
qH: 0.05
qL: 0.1
qW: 0.2
tao_max_its: 10
2-Norm Loss: 172.702
1.0-Norm Loss: 40459.8
Obj. Fn.: 2728.68
Pred. mistakes 7388
Sparsity W: 0.225
Sparsity H: 0.183551

___Cluster 0_____	Jeffords (I-VT)	Sessions (R-AL)
Sarbanes (D-MD)	Pryor (D-AR)	DeWine (R-OH)
Mikulski (D-MD)	Conrad (D-ND)	Bunning (R-KY)
Boxer (D-CA)	Graham (D-FL)	Coleman (R-MN)
Levin (D-MI)	Bayh (D-IN)	Shelby (R-AL)
Reed (D-RI)	Carper (D-DE)	Graham (R-SC)
Durbin (D-IL)	Lincoln (D-AR)	Hagel (R-NE)
Lautenberg (D-NJ)	Edwards (D-NC)	McConnell (R-KY)
Clinton (D-NY)	Landrieu (D-LA)	Lott (R-MS)
Stabenow (D-MI)	Baucus (D-MT)	Santorum (R-PA)
Corzine (D-NJ)	Lieberman (D-CT)	Inhofe (R-OK)
Leahy (D-VT)	Kerry (D-MA)	Thomas (R-WY)
Rockefeller (D-WV)	Breaux (D-LA)	Enzi (R-WY)
Akaka (D-HI)	Nelson (D-NE)	Fitzgerald (R-IL)
Harkin (D-IA)	___Cluster 1_____	Allard (R-CO)
Murray (D-WA)	Burns (R-MT)	Hutchison (R-TX)
Cantwell (D-WA)	Cochran (R-MS)	Voinovich (R-OH)
Schumer (D-NY)	Grassley (R-IA)	Kyl (R-AZ)
Kennedy (D-MA)	Chambliss (R-GA)	Domenici (R-NM)
Dodd (D-CT)	Roberts (R-KS)	Nickles (R-OK)
Reid (D-NV)	Frist (R-TN)	Murkowski (R-AK)
Daschle (D-SD)	Alexander (R-TN)	Smith (R-OR)
Feingold (D-WI)	Cornyn (R-TX)	Ensign (R-NV)
Kohl (D-WI)	Lugar (R-IN)	Sununu (R-NH)
Byrd (D-WV)	Bennett (R-UT)	Gregg (R-NH)
Johnson (D-SD)	Hatch (R-UT)	Campbell (R-CO)
Wyden (D-OR)	Dole (R-NC)	Specter (R-PA)
Dayton (D-MN)	Talent (R-MO)	McCain (R-AZ)
Bingaman (D-NM)	Stevens (R-AK)	Collins (R-ME)
Feinstein (D-CA)	Allen (R-VA)	Snowe (R-ME)
Dorgan (D-ND)	Brownback (R-KS)	Miller (D-GA)
Nelson (D-FL)	Warner (R-VA)	Chafee (R-RI)
Biden (D-DE)	Bond (R-MO)	
Inouye (D-HI)	Crapo (R-ID)	
Hollings (D-SC)	Craig (R-ID)	

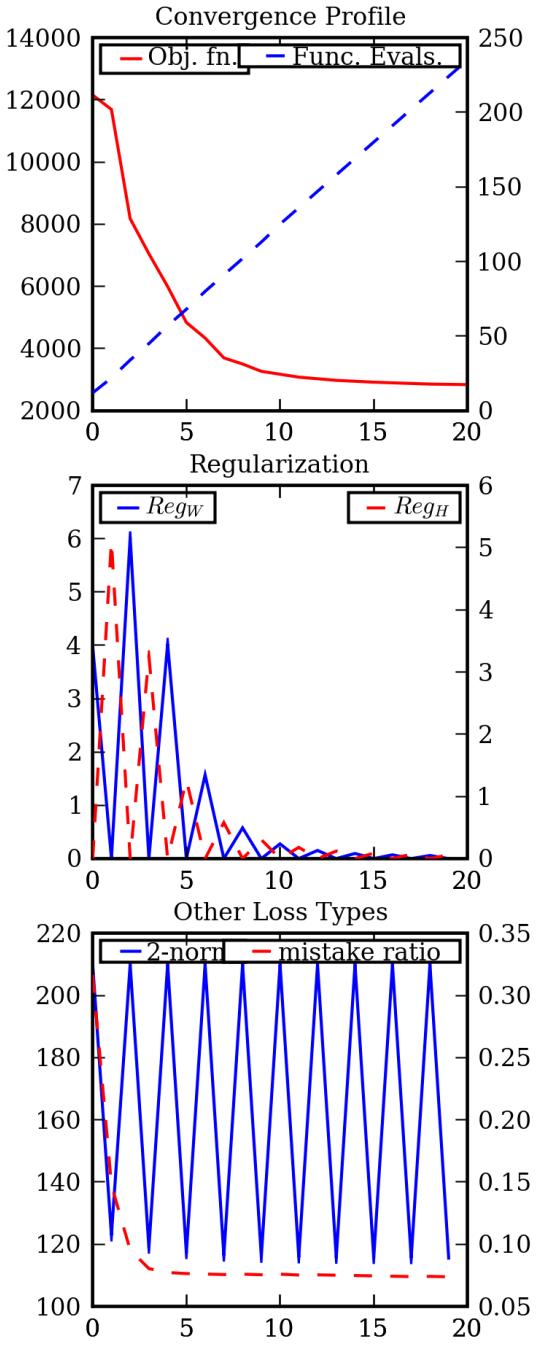
Figure 6.16: Senate Experiment 2



Experiment exp20071119095833
M: 8
P: 1
init_type: random_Vcol
itmax: 10
lambdaH: 100
lambdaW: 100
qH: 0.05
qL: 0.1
qW: 0.2
tao_max_its: 10
2-Norm Loss: 121.097
1.0-Norm Loss: 31631.5
Obj. Fn.: 2871.52
Pred. mistakes 6797
Sparsity W: 0.12
Sparsity H: 0.442266

___Cluster 0_____	Warner (R-VA)	Harkin (D-IA)
Snowe (R-ME)	Bond (R-MO)	Leahy (D-VT)
___Cluster 1_____	McConnell (R-KY)	Akaka (D-HI)
Landrieu (D-LA)	Lugar (R-IN)	Rockefeller (D-WV)
Breaux (D-LA)	Stevens (R-AK)	Murray (D-WA)
Baucus (D-MT)	Roberts (R-KS)	Dodd (D-CT)
Nelson (D-NE)	Talent (R-MO)	Cantwell (D-WA)
___Cluster 2_____	Lott (R-MS)	Schumer (D-NY)
Kyl (R-AZ)	Hutchison (R-TX)	Feingold (D-WI)
Nickles (R-OK)	Fitzgerald (R-IL)	Reid (D-NV)
Thomas (R-WY)	Voinovich (R-OH)	Kennedy (D-MA)
Enzi (R-WY)	DeWine (R-OH)	Byrd (D-WV)
Allard (R-CO)	Murkowski (R-AK)	Daschle (D-SD)
Sessions (R-AL)	Hagel (R-NE)	Wyden (D-OR)
Santorum (R-PA)	McCain (R-AZ)	Johnson (D-SD)
Cornyn (R-TX)	Specter (R-PA)	Kohl (D-WI)
Gregg (R-NH)	Collins (R-ME)	Dayton (D-MN)
Allen (R-VA)	___Cluster 3_____	Bingaman (D-NM)
Craig (R-ID)	Coleman (R-MN)	Feinstein (D-CA)
Graham (R-SC)	Smith (R-OR)	Dorgan (D-ND)
Sununu (R-NH)	Domenici (R-NM)	Biden (D-DE)
Crapo (R-ID)	Miller (D-GA)	Jeffords (I-VT)
Hatch (R-UT)	___Cluster 4_____	Hollings (D-SC)
Frist (R-TN)	Campbell (R-CO)	Nelson (D-FL)
Chambliss (R-GA)	Chafee (R-RI)	Inouye (D-HI)
Inhofe (R-OK)	___Cluster 5_____	Pryor (D-AR)
Dole (R-NC)	Boxer (D-CA)	Conrad (D-ND)
Bunning (R-KY)	Levin (D-MI)	Graham (D-FL)
Cochran (R-MS)	Sarbanes (D-MD)	Bayh (D-IN)
Bennett (R-UT)	Lautenberg (D-NJ)	Carper (D-DE)
Ensign (R-NV)	Durbin (D-IL)	Lincoln (D-AR)
Shelby (R-AL)	Reed (D-RI)	Edwards (D-NC)
Alexander (R-TN)	Corzine (D-NJ)	Lieberman (D-CT)
Grassley (R-IA)	Mikulski (D-MD)	Kerry (D-MA)
Burns (R-MT)	Clinton (D-NY)	
Brownback (R-KS)	Stabenow (D-MI)	

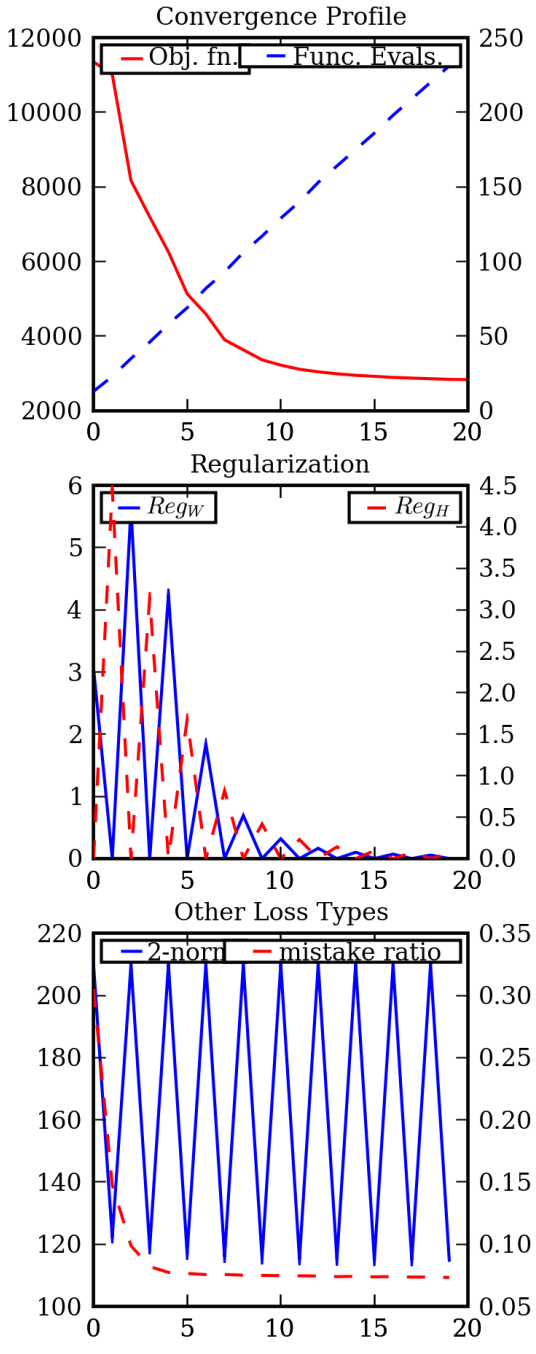
Figure 6.17: Senate Experiment 3



Experiment exp20071119100118
M: 16
P: 1
init_type: random_Vcol
itmax: 10
lambdaH: 100
lambdaW: 100
qH: 0.05
qL: 0.1
qW: 0.2
tao_max_its: 10
2-Norm Loss: 115.528
1.0-Norm Loss: 30038.9
Obj. Fn.: 2841.09
Pred. mistakes 6775
Sparsity W: 0.099375
Sparsity H: 0.520289

___Cluster 0_____	Alexander (R-TN)	Clinton (D-NY)
Breaux (D-LA)	Coleman (R-MN)	Dodd (D-CT)
Baucus (D-MT)	Grassley (R-IA)	Leahy (D-VT)
Nelson (D-NE)	Voinovich (R-OH)	Feingold (D-WI)
___Cluster 1_____	Lugar (R-IN)	Cantwell (D-WA)
Thomas (R-WY)	Hatch (R-UT)	Byrd (D-WV)
Nickles (R-OK)	Brownback (R-KS)	Murray (D-WA)
Kyl (R-AZ)	Talent (R-MO)	Daschle (D-SD)
Enzi (R-WY)	Roberts (R-KS)	Kohl (D-WI)
Allard (R-CO)	Fitzgerald (R-IL)	Johnson (D-SD)
Sessions (R-AL)	Dole (R-NC)	Dorgan (D-ND)
Gregg (R-NH)	Burns (R-MT)	Bingaman (D-NM)
Sununu (R-NH)	DeWine (R-OH)	Dayton (D-MN)
Cornyn (R-TX)	Stevens (R-AK)	Biden (D-DE)
Santorum (R-PA)	McCain (R-AZ)	Jeffords (I-VT)
Allen (R-VA)	Smith (R-OR)	Conrad (D-ND)
Craig (R-ID)	Snowe (R-ME)	Pryor (D-AR)
McConnell (R-KY)	Chafee (R-RI)	Bayh (D-IN)
Ensign (R-NV)	___Cluster 3_____	Lincoln (D-AR)
Graham (R-SC)	Domenici (R-NM)	Inouye (D-HI)
Bennett (R-UT)	___Cluster 4_____	Carper (D-DE)
Crapo (R-ID)	Miller (D-GA)	___Cluster 7_____
Bunning (R-KY)	___Cluster 5_____	Wyden (D-OR)
Inhofe (R-OK)	Campbell (R-CO)	Feinstein (D-CA)
Frist (R-TN)	___Cluster 6_____	Rockefeller (D-WV)
Shelby (R-AL)	Levin (D-MI)	Schumer (D-NY)
Chambliss (R-GA)	Stabenow (D-MI)	Kennedy (D-MA)
Cochran (R-MS)	Durbin (D-IL)	Nelson (D-FL)
Warner (R-VA)	Sarbanes (D-MD)	Hollings (D-SC)
Bond (R-MO)	Harkin (D-IA)	Landrieu (D-LA)
Lott (R-MS)	Mikulski (D-MD)	___Cluster 8_____
Hutchison (R-TX)	Boxer (D-CA)	Edwards (D-NC)
Murkowski (R-AK)	Corzine (D-NJ)	Lieberman (D-CT)
Specter (R-PA)	Lautenberg (D-NJ)	Kerry (D-MA)
Collins (R-ME)	Akaka (D-HI)	Graham (D-FL)
___Cluster 2_____	Reid (D-NV)	
Hagel (R-NE)	Reed (D-RI)	

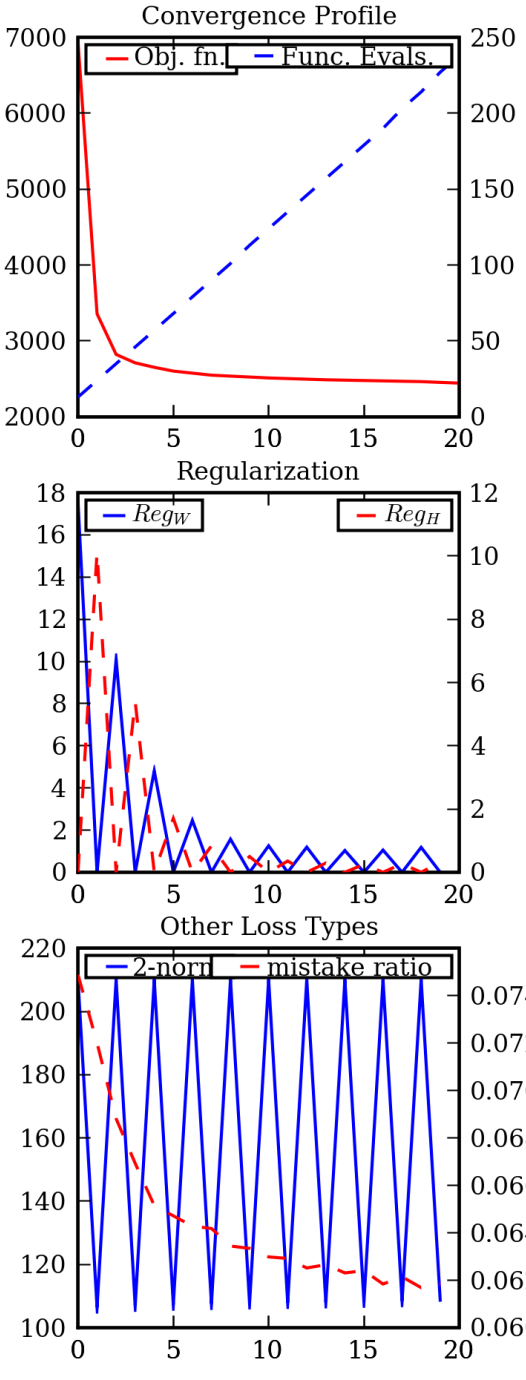
Figure 6.18: Senate Experiment 4



Experiment exp20071119100436
M: 24
P: 1
init_type: random_Vcol
itmax: 10
lambdaH: 100
lambdaW: 100
qH: 0.05
qL: 0.1
qW: 0.2
tao_max_its: 10
2-Norm Loss: 114.835
1.0-Norm Loss: 29650.5
Obj. Fn.: 2834.13
Pred. mistakes 6728
Sparsity W: 0.075
Sparsity H: 0.589052

___Cluster 0_____	Hagel (R-NE)	Dodd (D-CT)
Breaux (D-LA)	Voinovich (R-OH)	Leahy (D-VT)
Nelson (D-NE)	Coleman (R-MN)	Cantwell (D-WA)
___Cluster 1_____	Grassley (R-IA)	Feingold (D-WI)
Thomas (R-WY)	Alexander (R-TN)	Murray (D-WA)
Nickles (R-OK)	Lugar (R-IN)	Byrd (D-WV)
Enzi (R-WY)	Hatch (R-UT)	Daschle (D-SD)
Allard (R-CO)	Brownback (R-KS)	Kohl (D-WI)
Kyl (R-AZ)	Talent (R-MO)	Johnson (D-SD)
Sessions (R-AL)	Roberts (R-KS)	Dorgan (D-ND)
Gregg (R-NH)	Fitzgerald (R-IL)	Bingaman (D-NM)
Sununu (R-NH)	DeWine (R-OH)	Dayton (D-MN)
Cornyn (R-TX)	Stevens (R-AK)	Biden (D-DE)
Santorum (R-PA)	McCain (R-AZ)	Jeffords (I-VT)
Allen (R-VA)	Smith (R-OR)	Conrad (D-ND)
McConnell (R-KY)	Snowe (R-ME)	Pryor (D-AR)
Craig (R-ID)	Chafee (R-RI)	Bayh (D-IN)
Bennett (R-UT)	___Cluster 3_____	Lincoln (D-AR)
Graham (R-SC)	Domenici (R-NM)	Carper (D-DE)
Crapo (R-ID)	___Cluster 4_____	Baucus (D-MT)
Cochran (R-MS)	Miller (D-GA)	___Cluster 7_____
Ensign (R-NV)	___Cluster 5_____	Wyden (D-OR)
Bunning (R-KY)	Campbell (R-CO)	Rockefeller (D-WV)
Frist (R-TN)	___Cluster 6_____	Feinstein (D-CA)
Inhofe (R-OK)	Levin (D-MI)	Schumer (D-NY)
Chambliss (R-GA)	Stabenow (D-MI)	Kennedy (D-MA)
Burns (R-MT)	Durbin (D-IL)	Hollings (D-SC)
Dole (R-NC)	Sarbanes (D-MD)	Nelson (D-FL)
Warner (R-VA)	Mikulski (D-MD)	Inouye (D-HI)
Shelby (R-AL)	Corzine (D-NJ)	Landrieu (D-LA)
Bond (R-MO)	Akaka (D-HI)	___Cluster 8_____
Lott (R-MS)	Reid (D-NV)	Lieberman (D-CT)
Hutchison (R-TX)	Boxer (D-CA)	Edwards (D-NC)
Murkowski (R-AK)	Harkin (D-IA)	Kerry (D-MA)
Specter (R-PA)	Lautenberg (D-NJ)	Graham (D-FL)
Collins (R-ME)	Clinton (D-NY)	
___Cluster 2_____	Reed (D-RI)	

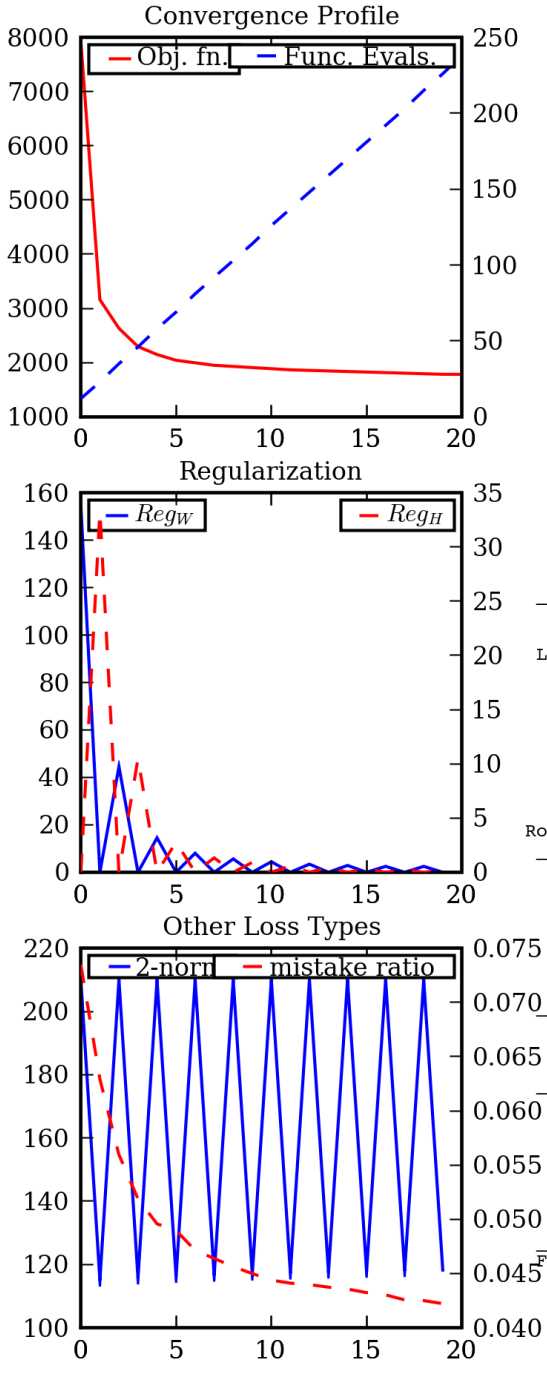
Figure 6.19: Senate Experiment 5



Experiment exp20071119095720
M: 8
P: 1
init_type: random_Vcol
itmax: 10
lambdaH: 1
lambdaW: 1
qH: 0.05
qL: 0.1
qW: 0.2
tao_max_its: 10
2-Norm Loss: 108.704
1.0-Norm Loss: 27476.3
Obj. Fn.: 2445.53
Pred. mistakes 5676
Sparsity W: 0.31125
Sparsity H: 0.511029

___Cluster 0_____	Dole (R-NC)	Hagel (R-NE)
Kyl (R-AZ)	McConnell (R-KY)	Burns (R-MT)
Santorum (R-PA)	Lugar (R-IN)	Campbell (R-CO)
Nickles (R-OK)	Alexander (R-TN)	Hatch (R-UT)
Sununu (R-NH)	Bunning (R-KY)	Lott (R-MS)
Allard (R-CO)	Chambliss (R-GA)	Lieberman (D-CT)
Sessions (R-AL)	Talent (R-MO)	___Cluster 6_____
Ensign (R-NV)	Crapo (R-ID)	Miller (D-GA)
Enzi (R-WY)	Grassley (R-IA)	Graham (D-FL)
Thomas (R-WY)	Bennett (R-UT)	Smith (R-OR)
McCain (R-AZ)	Roberts (R-KS)	___Cluster 7_____
Gregg (R-NH)	Brownback (R-KS)	Boxer (D-CA)
Graham (R-SC)	Inhofe (R-OK)	Durbin (D-IL)
Reed (D-RI)	Warner (R-VA)	Sarbanes (D-MD)
Allen (R-VA)	Fitzgerald (R-IL)	Lautenberg (D-NJ)
Hollings (D-SC)	Hutchison (R-TX)	Corzine (D-NJ)
Kennedy (D-MA)	Murkowski (R-AK)	Clinton (D-NY)
___Cluster 1_____	Voinovich (R-OH)	Cantwell (D-WA)
Dayton (D-MN)	___Cluster 4_____	Mikulski (D-MD)
Levin (D-MI)	Breaux (D-LA)	Harkin (D-IA)
Stabenow (D-MI)	Nelson (D-NE)	Akaka (D-HI)
Coleman (R-MN)	Baucus (D-MT)	Rockefeller (D-WV)
DeWine (R-OH)	Landrieu (D-LA)	Feingold (D-WI)
___Cluster 2_____	Lincoln (D-AR)	Murray (D-WA)
Snowe (R-ME)	Pryor (D-AR)	Leahy (D-VT)
Collins (R-ME)	Bayh (D-IN)	Reid (D-NV)
Biden (D-DE)	Carper (D-DE)	Dodd (D-CT)
Inouye (D-HI)	Conrad (D-ND)	Schumer (D-NY)
Wyden (D-OR)	Dorgan (D-ND)	Daschle (D-SD)
Feinstein (D-CA)	___Cluster 5_____	Kohl (D-WI)
Specter (R-PA)	Edwards (D-NC)	Byrd (D-WV)
Jeffords (I-VT)	Stevens (R-AK)	Johnson (D-SD)
Chafee (R-RI)	Domenici (R-NM)	Bingaman (D-NM)
___Cluster 3_____	Cochran (R-MS)	Nelson (D-FL)
Cornyn (R-TX)	Kerry (D-MA)	
Craig (R-ID)	Bond (R-MO)	
Frist (R-TN)	Shelby (R-AL)	

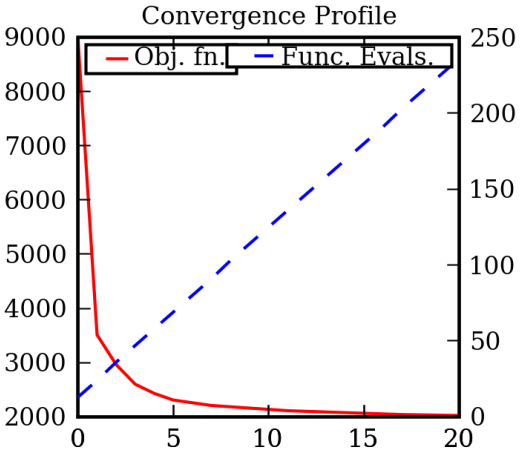
Figure 6.20: Senate Experiment 6

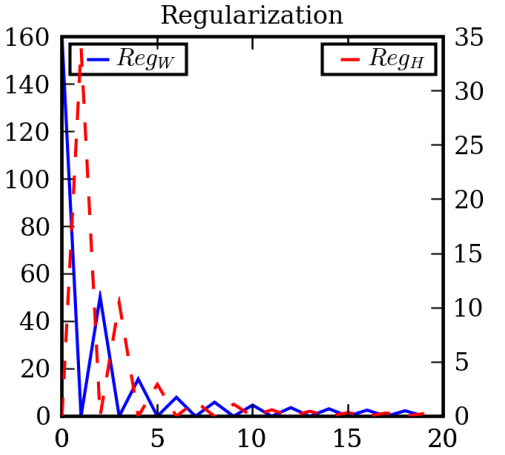


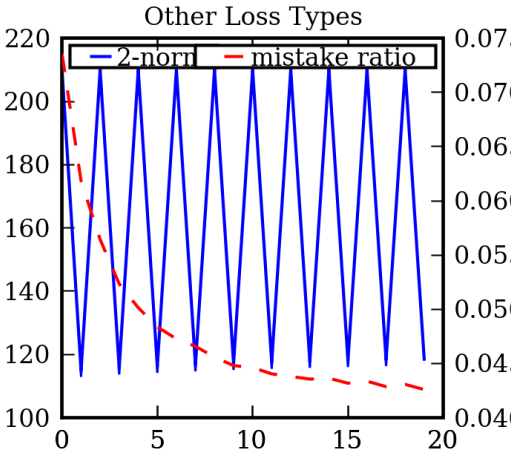
Experiment exp20071119100251
M: 24
P: 1
init_type: random_Vcol
itmax: 10
lambdaH: 1
lambdaW: 1
qH: 0.05
qL: 0.1
qW: 0.2
tao_max_its: 10
2-Norm Loss: 118.235
1.0-Norm Loss: 29694.2
Obj. Fn.: 1783.23
Pred. mistakes 3876
Sparsity W: 0.4775
Sparsity H: 0.600808

___Cluster 0_____	Hagel (R-NE)	Cochran (R-MS)
Hollings (D-SC)	Hatch (R-UT)	___Cluster 13_____
Kennedy (D-MA)	Bond (R-MO)	Durbin (D-IL)
Lautenberg (D-NJ)	___Cluster 5_____	Kohl (D-WI)
Boxer (D-CA)	Lincoln (D-AR)	___Cluster 14_____
Reed (D-RI)	Conrad (D-ND)	Dayton (D-MN)
Harkin (D-IA)	___Cluster 6_____	Feinstein (D-CA)
Akaka (D-HI)	Sununu (R-NH)	Lieberman (D-CT)
Corzine (D-NJ)	Nickles (R-OK)	Pryor (D-AR)
Leahy (D-VT)	Kyl (R-AZ)	Inouye (D-HI)
Clinton (D-NY)	Allard (R-CO)	Wyden (D-OR)
Sarbanes (D-MD)	Enzi (R-WY)	___Cluster 15_____
Rockefeller (D-WV)	Lott (R-MS)	Warner (R-VA)
___Cluster 1_____	___Cluster 7_____	___Cluster 16_____
Murray (D-WA)	Domenici (R-NM)	Edwards (D-NC)
Stabenow (D-MI)	Burns (R-MT)	Dorgan (D-ND)
Cantwell (D-WA)	___Cluster 8_____	Johnson (D-SD)
Levin (D-MI)	Dodd (D-CT)	Daschle (D-SD)
Reid (D-NV)	___Cluster 9_____	Mikulski (D-MD)
Feingold (D-WI)	Inhofe (R-OK)	___Cluster 17_____
Schumer (D-NY)	Crapo (R-ID)	Kerry (D-MA)
___Cluster 2_____	Craig (R-ID)	Biden (D-DE)
Breaux (D-LA)	Bunning (R-KY)	Byrd (D-WV)
Baucus (D-MT)	Cornyn (R-TX)	___Cluster 18_____
Jeffords (I-VT)	Frist (R-TN)	Graham (D-FL)
___Cluster 3_____	___Cluster 10_____	___Cluster 19_____
Specter (R-PA)	Nelson (D-NE)	McCain (R-AZ)
Gregg (R-NH)	Coleman (R-MN)	Smith (R-OR)
Santorum (R-PA)	Bayh (D-IN)	Ensign (R-NV)
Sessions (R-AL)	Alexander (R-TN)	Brownback (R-KS)
Campbell (R-CO)	___Cluster 11_____	Graham (R-SC)
Thomas (R-WY)	Landrieu (D-LA)	Chambliss (R-GA)
Allen (R-VA)	Voinovich (R-OH)	___Cluster 20_____
___Cluster 4_____	___Cluster 12_____	Nelson (D-FL)
Fitzgerald (R-IL)	Miller (D-GA)	___Cluster 21_____
DeWine (R-OH)	McConnell (R-KY)	Chafee (R-RI)
Lugar (R-IN)	Murkowski (R-AK)	Snowe (R-ME)
Roberts (R-KS)	Stevens (R-AK)	Collins (R-ME)
Talent (R-MO)	Shelby (R-AL)	Carper (R-DE)
Grassley (R-IA)	Hutchison (R-TX)	Bingaman (D-NM)
Dole (R-NC)	Bennett (R-UT)	

Figure 6.21: Senate Experiment 7



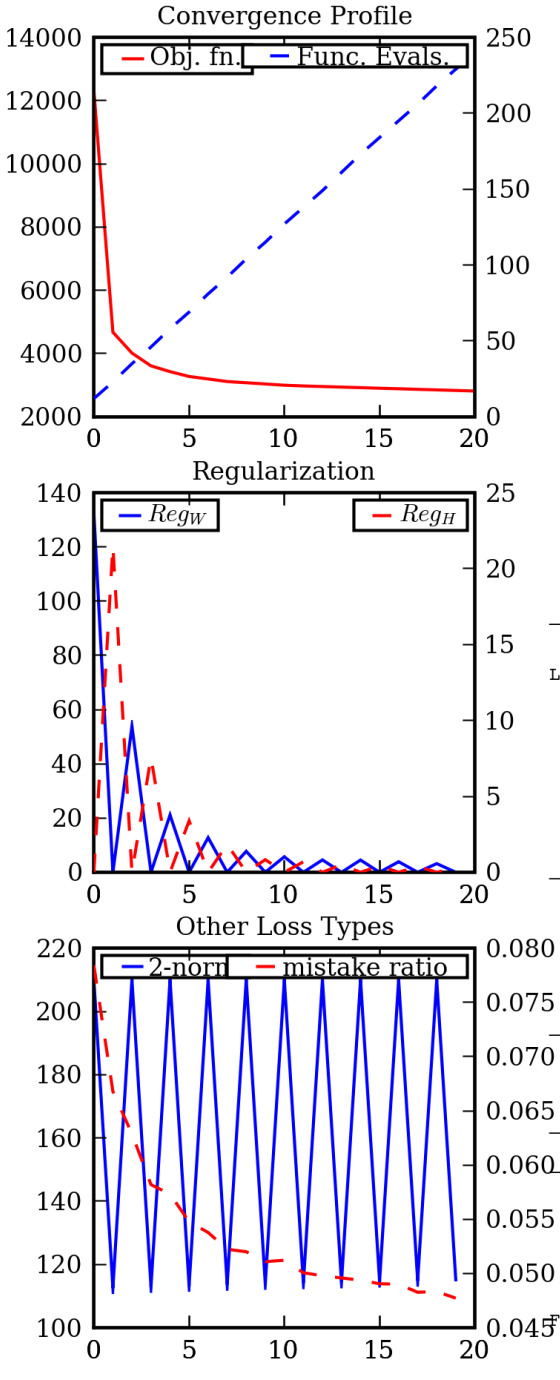




Experiment exp20071119100325
M: 24
P: 1
init_type: random_Vcol
itmax: 10
lambdaH: 1
lambdaW: 1
qH: 0.05
qL: 0.2
qW: 0.2
tao_max_its: 10
2-Norm Loss: 118.504
1.0-Norm Loss: 30247.3
Obj. Fn.: 2021.84
Pred. mistakes 3910
Sparsity W: 0.507083
Sparsity H: 0.593455

```
___Cluster 0_____
   Hollings (D-SC)
     Kennedy (D-MA)
  Lautenberg (D-NJ)
      Harkin (D-IA)
       Boxer (D-CA)
        Reed (D-RI)
   Feinstein (D-CA)
      Akaka (D-HI)
   Sarbanes (D-MD)
Rockefeller (D-WV)
___Cluster 1_____
      Dayton (D-MN)
   Stabenow (D-MI)
   Cantwell (D-WA)
      Murray (D-WA)
       Levin (D-MI)
     Clinton (D-NY)
   Feingold (D-WI)
        Reid (D-NV)
     Corzine (D-NJ)
      Durbin (D-IL)
     Schumer (D-NY)
        Dodd (D-CT)
___Cluster 2_____
      Breaux (D-LA)
   Jeffords (I-VT)
___Cluster 3_____
   Voinovich (R-OH)
  Fitzgerald (R-IL)
      DeWine (R-OH)
       Lugar (R-IN)
   Grassley (R-IA)
      Talent (R-MO)
___Cluster 4_____
      Conrad (D-ND)
      Dorgan (D-ND)
     Bunning (R-KY)
___Cluster 5_____
      Ensign (R-NV)
     Sununu (R-NH)
      Nickles (R-OK)
         Kyl (R-AZ)
     Allard (R-CO)
       Gregg (R-NH)
    Sessions (R-AL)
   Santorum (R-PA)
       Enzi (R-WY)
     Thomas (R-WY)
     Graham (R-SC)
       Lott (R-MS)
___Cluster 6_____
    Domenici (R-NM)
      Burns (R-MT)
___Cluster 7_____
    Bingaman (D-NM)
___Cluster 8_____
      Inouye (D-HI)
    Lincoln (D-AR)
       Byrd (D-WV)
       Kohl (D-WI)
___Cluster 9_____
      Inhofe (R-OK)
      Crapo (R-ID)
      Craig (R-ID)
     Cornyn (R-TX)
      Hatch (R-UT)
      Frist (R-TN)
      Allen (R-VA)
       Bond (R-MO)
___Cluster 10____
     Nelson (D-NE)
      Smith (R-OR)
    Coleman (R-MN)
       Dole (R-NC)
  Alexander (R-TN)
___Cluster 11____
   Landrieu (D-LA)
___Cluster 12____
      Miller (D-GA)
  McConnell (R-KY)
   Hutchison (R-TX)
      Shelby (R-AL)
   Murkowski (R-AK)
     Stevens (R-AK)
     Roberts (R-KS)
     Bennett (R-UT)
     Cochran (R-MS)
       Hagel (R-NE)
___Cluster 13____
   Lieberman (D-CT)
      Pryor (D-AR)
___Cluster 14____
     Specter (R-PA)
     Warner (R-VA)
      Leahy (D-VT)
___Cluster 15____
     Edwards (D-NC)
   Campbell (R-CO)
    Daschle (D-SD)
    Johnson (D-SD)
   Mikulski (D-MD)
___Cluster 16____
      Kerry (D-MA)
     Graham (D-FL)
      Biden (D-DE)
___Cluster 17____
     McCain (R-AZ)
  Brownback (R-KS)
  Chambliss (R-GA)
___Cluster 18____
     Nelson (D-FL)
      Bayh (D-IN)
___Cluster 19____
     Chafee (R-RI)
     Snowe (R-ME)
    Baucus (D-MT)
   Collins (R-ME)
    Carper (D-DE)
     Wyden (D-OR)
```

Figure 6.22: Senate Experiment 8



Experiment exp20071119100359
M: 24
P: 1
init_type: random_Vcol
itmax: 10
lambdaH: 1
lambdaW: 1
qH: 0.05
qL: 0.4
qW: 0.2
tao_max_its: 10
2-Norm Loss: 115.142
1.0-Norm Loss: 29156.9
Obj. Fn.: 2817.73
Pred. mistakes 4381
Sparsity W: 0.472917
Sparsity H: 0.646106

```
___Cluster 0_____
   Hollings (D-SC)
     Kennedy (D-MA)
Lautenberg (D-NJ)
  Feinstein (D-CA)
        Reed (D-RI)
       Boxer (D-CA)
       Wyden (D-OR)
     Clinton (D-NY)
     Corzine (D-NJ)
       Akaka (D-HI)
       Leahy (D-VT)
     Schumer (D-NY)
___Cluster 1_____
       Snowe (R-ME)
      Nelson (D-FL)
      Dayton (D-MN)
  Bingaman (D-NM)
      Murray (D-WA)
   Stabenow (D-MI)
        Kohl (D-WI)
___Cluster 2_____
      Breaux (D-LA)
      Baucus (D-MT)
   Landrieu (D-LA)
   Jeffords (I-VT)
___Cluster 3_____
       Allen (R-VA)
___Cluster 4_____
     Collins (R-ME)
      Chafee (R-RI)
     McCain (R-AZ)
      Smith (R-OR)
  Voinovich (R-OH)
    Coleman (R-MN)
     DeWine (R-OH)
 Fitzgerald (R-IL)
      Lugar (R-IN)
    Grassley (R-IA)
___Cluster 5_____
    Lincoln (D-AR)
      Pryor (D-AR)
     Conrad (D-ND)
    Johnson (D-SD)
    Daschle (D-SD)
     Dorgan (D-ND)
  Brownback (R-KS)
___Cluster 6_____
      Gregg (R-NH)
     Sununu (R-NH)
      Ensign (R-NV)
    Nickles (R-OK)
   Santorum (R-PA)
   Sessions (R-AL)
     Allard (R-CO)
    Specter (R-PA)
___Cluster 7_____
   Domenici (R-NM)
  Hutchison (R-TX)
    Cochran (R-MS)
  McConnell (R-KY)
      Burns (R-MT)
___Cluster 8_____
     Inouye (D-HI)
     Harkin (D-IA)
___Cluster 9_____
        Kyl (R-AZ)
    Thomas (R-WY)
     Craig (R-ID)
      Enzi (R-WY)
    Graham (R-SC)
    Inhofe (R-OK)
     Crapo (R-ID)
    Cornyn (R-TX)
   Bunning (R-KY)
___Cluster 10_____
    Nelson (D-NE)
      Bayh (D-IN)
___Cluster 11_____
    Carper (D-DE)
      Lott (R-MS)
    Warner (R-VA)
___Cluster 12_____
     Miller (D-GA)
  Chambliss (R-GA)
    Roberts (R-KS)
     Shelby (R-AL)
  Alexander (R-TN)
      Hagel (R-NE)
    Stevens (R-AK)
     Frist (R-TN)
___Cluster 13_____
       Reid (D-NV)
   Cantwell (D-WA)
Rockefeller (D-WV)
___Cluster 14_____
     Edwards (D-NC)
  Lieberman (D-CT)
      Kerry (D-MA)
___Cluster 15_____
       Bond (R-MO)
     Talent (R-MO)
___Cluster 16_____
    Bennett (R-UT)
      Dole (R-NC)
  Campbell (R-CO)
 Murkowski (R-AK)
      Hatch (R-UT)
___Cluster 17_____
    Graham (D-FL)
      Byrd (D-WV)
     Biden (D-DE)
      Dodd (D-CT)
  Sarbanes (D-MD)
    Durbin (D-IL)
     Levin (D-MI)
  Feingold (D-WI)
___Cluster 18_____
  Mikulski (D-MD)
```

Figure 6.23: Senate Experiment 9

6.3 Discussion

In this chapter, the q-entropy functions introduced much earlier proved to be useful in constructing an effective and efficient algorithm to perform non-negative matrix factorization. In particular non-Legendre entropies ($q < 1$) were found to be quite useful for the purpose. The potential motivations of the various entropic formulations faded into the background. Instead the properties of the solutions came to the fore. The dual characterization made available via the convex analysis of earlier chapters remains useful even in empirical work, since it helps to indicate how and when sparsity will show up in the solution.

NMF converges only to a local minimum and little theory about these solutions is available. Hopefully the treatment represents a contribution by describing an algorithm in terms of steps involving well-behaved generalized maxent problems, with some describable properties at each step. In addition, the parametrization given here has a sensible interpretation across a range of data sets.

6.4 Appendix: Derivation of Equation 6.3

This appendix contains the derivation of the dual characterization of the W-step referred to in Section 6.0.2. Problems W and H are almost entirely symmetric in our setup, so we will switch to a neutral notation and later apply the result to both problems. Consider the NMF problem as the following approximation:

$$\boldsymbol{BC} \approx \boldsymbol{D}$$

Symmetry will allow us to focus solely on Problem 'B' which will be solved using the objective function

$$\Delta S_{q'}(\boldsymbol{BC}, \boldsymbol{D}) + \lambda \Delta S_q(\boldsymbol{B}, \boldsymbol{B}_0).$$

The theory of Chapter 3 was developed for vectors, not matrices, so we rewrite the preceding line indicating for convenience how the terms map to the original notation.

$$\Delta S_{q'}(\underbrace{(\boldsymbol{C} \otimes \boldsymbol{I}_N)}_{'\boldsymbol{A}'} \underbrace{\operatorname{vec} \boldsymbol{B}}_{'\boldsymbol{p}'}, \underbrace{\operatorname{vec} \boldsymbol{D}}_{'\boldsymbol{b}'}) + \lambda \, \Delta S_q(\operatorname{vec} \boldsymbol{B}, \underbrace{\operatorname{vec} \boldsymbol{B}_0}_{'\boldsymbol{p}_0'})$$

This identification makes the dual objective function (with argument $\operatorname{vec} \boldsymbol{D}^*$) straightforward to calculate using Fenchel duality, the conjugation formula for Bregman divergence, and the conjugate calculus rules to handle λ in succession.

The dual objective is:

$$\lambda S_q^*(\lambda^{-1}(\boldsymbol{C}^{\mathrm{T}} \otimes \boldsymbol{I}_N) \operatorname{vec} \boldsymbol{D}^* + \boldsymbol{B}_0^*) + S_{q'}^*(\operatorname{vec} \boldsymbol{D}^* - \boldsymbol{D}_0^*),$$

where $\boldsymbol{B}_0^*$ is the image of the initial guess $\boldsymbol{B}_0$ under the gradient of S_q, and where $\boldsymbol{D}_0^*$ is the image of the data matrix $\boldsymbol{D}$ under the gradient of $S_{q'}$. Given an optimal solution for this problem, $\bar{\boldsymbol{D}}^*$, we can use the recovery formula to obtain:

$$\bar{B} = \nabla S_q^*(\, (\boldsymbol{C}^{\mathrm{T}} \otimes \boldsymbol{I}_N) \operatorname{vec} \bar{\boldsymbol{D}}^* + \boldsymbol{B}_0^*),$$

which is perhaps easier to parse without the vec operations:

$$\bar{B} = s_q'^*[\, \bar{\boldsymbol{D}}^* \boldsymbol{C} + \boldsymbol{B}_0^*\,].$$

For Problem W, substitution of $\boldsymbol{W}$ for $\boldsymbol{B}$, $\boldsymbol{H}$ for $\boldsymbol{C}$ and $\boldsymbol{V}$ for $\boldsymbol{D}$ gives the result in the text.

As an aside, the symmetry of Problem H is nearly complete if we consider the approximation problem $\boldsymbol{H}^{\mathrm{T}}\boldsymbol{W}^{\mathrm{T}} \approx \boldsymbol{V}^{\mathrm{T}}$. By substituting HvT for $\boldsymbol{B}$, etc., the same argument reveals the solution to the H-step as another q-exponential namely:

$$\bar{\boldsymbol{H}} = \exp_{q_H}[\, \boldsymbol{W}^{\mathrm{T}} \bar{\boldsymbol{V}}^* + \boldsymbol{H}_0^*\,].$$

Finally, the dual problem has the choice variable $\boldsymbol{V}^*$, which has the same shape as $\boldsymbol{V}$, much larger than $\boldsymbol{W}$ or $\boldsymbol{H}$. This makes the dual a poor choice of problem to solve, in comparison to, say, the classic maxent problem.

Chapter 7

Induced Semantics for Undirected Graphs

7.1 Another Look at the Hammersley-Clifford Theorem

In this chapter the aim is to utilize the entropy functions introduced earlier in the framework of graphical models. The Hammersley-Clifford (H-C) theorem relates the factorization properties of a probability distribution to the clique structure of an undirected graph. If a density factorizes according to the clique structure of an undirected graph, the theorem guarantees that the distribution satisfies the Markov property and vice versa. In graphical models, connections to the exponential family are strong, mainly because of the factorization property of the exponential function. Do we have to forgo graphical models when we step away from SBG entropy? Not entirely. Here we generalize the H-C theorem to different notions of decomposability and the corresponding generalized-Markov property. Finally we discuss how our technique might be used to arrive at other generalizations of the H-C theorem, inducing a graph semantics adapted to the modeling problem. This represents a first step in incoporating generalized entropies in the setting of graphical models.

Statistical distributions of the q-exponential form can be motivated by a generalization of maximum entropy termed Tsallis entropy. This is one possible generalization of Shannon-Boltzmann-Gibbs (SBG) entropy. An important property of Tsallis entropy is its capability of generating distributions with power-law behaviour and distributions with finite support. One of these is the q-Gaussian distribution. Its density, assuming it is centered around the origin is

$$f(x) = \exp_q(-\beta x^2 - \alpha_q(\beta)),$$

where $\exp_q$ is the q-exponential, defined later in the section on q-analogues. For now, we just note that the usual Gaussian is recovered when $q \to 1$. See Naudts

(2004b); Vignat and Plastino (2006); Sears and Vishwanathan (2007); Gell-Mann and Tsallis (2004) for more discussion and references along these lines. Here, our main concern is to judge the compatibility of distributions of the form $\exp_q(\cdot)$ with the tools and techniques of graphical models. We are motivated by the possibility of using joint distributions of q-exponential form to model situations where a collection of random variables correlated to a greater degree (or less) under dramatic circumstances than they do in an everyday environment.

If we have N independent, zero-mean, Gaussian random variables (therefore $q = 1$), the joint density is

$$\prod_{i=1}^{N} \exp(-\beta_i x_i^2 - \alpha_i) = \exp(-\sum_{i=1}^{N}(\beta_i x_i^2) - (\sum_{i=1}^{N} \alpha_i)).$$

To have a corresponding formula for the q-exponentials, an operation called the q-product($\otimes_q$) which satisfies

$$\exp_q(x_1) \otimes_q \exp_q(x_2) \otimes_q \ldots \otimes_q \exp_q(x_N) = \exp_q(\sum_{i=1}^{N} x_i)$$

must be introduced. The q-product has been studied elsewhere (Suyari et al., 2005; Umarov et al., 2006; Borges, 2004). It leads to a definition of q-independence that says that X_1 and X_2 are q-independent if their joint density, f, q-factorizes, i.e. if for some g and h

$$f(x_1, x_2) = g(x_1) \otimes_q h(x_2).$$

Note when $q \neq 1$, g and h are not the marginals of f. Even so the q-product can be used to create joint distributions from univariate distributions. Recently, a central limit theorem for q-independent variables was proved (Umarov et al., 2006). Bayesian updating of such distributions is a different matter.

Introducing the idea of q-independence to the setting of graphical models induces a different semantics for the edges of a graph. It is possible to demonstrate this by formulating a q-Markov condition and prove a version of the Hammersley-Clifford theorem that says that a distribution is q-Markov with respect to the graph in question, if and only if it q-factorizes over the maximal cliques of that graph. We will conclude with a brief discussion of how to use this type of graphical model to perform typical tasks involving graphical models, focusing on a corresponding extension of the Viterbi algorithm.

7.2 Background

The proof of the Hammersley-Clifford Theorem relies on terminology and results from both number theory and graph theory. We give the necessary elements here as well as on q-analogues of some elementary mathematical operations.

7.2.1 Graphical Models

A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consists of a set of vertices (or nodes), $\mathcal{V}$, and a set of edges $\mathcal{E}$. An edge is an ordered pair of nodes. A clique, $c \in \mathcal{V}$ is a fully connected subgraph of $\mathcal{G}$. The vertices, $\{ a_i \}_{i=1}^{M}$ in a graphical model typically correspond to the variables of a distribution, with density $f(\boldsymbol{x})$. We will move between these representations assuming the order of the nodes is the same.

We will say a function $\mathcal{T}$-decomposes according to a graph for the transformation $\mathcal{T} : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^{\mathcal{X}}$ if there are ψ_c such that

$$\mathcal{T} f(\boldsymbol{x}) = \sum_{c \in \mathcal{C}} \psi_c(\boldsymbol{x}_c), \tag{7.1}$$

where $\mathcal{C}$ is the set of cliques of $\mathcal{G}$ and where $\boldsymbol{x}_c$ is equal to $\boldsymbol{x}$ with the entries corresponding to $\mathcal{V} \backslash c$ removed.

Let $\boldsymbol{x}^c$ denote a vector equal to $\boldsymbol{x}$ with the m-th and n-th entries deleted. We say that a function has the pairwise generalized-Markov property if, whenever the nodes a_m and a_n do not share an edge, then there exists functions with appropriate domains and ranges, h_1 and h_2, such that

$$\mathcal{T} f(\boldsymbol{x}) = h_1(x_m, \boldsymbol{x}^c) + h_2(x_n, \boldsymbol{x}^c). \tag{7.2}$$

In the classic formulation of the Hammersley-Clifford Theorem $\mathcal{T} = \log$ is implicitly assumed. Also in this setting decomposition is equivalent to multiplicative factorization. The role of $\mathcal{T}$ is therefore to convert factorization (of some kind) into addition. The generalized-Markov property then asserts the separability of the effects of x_m and x_n when the corresponding nodes are separated. More in-depth background in graph theory, especially as it relates to graphical models is given by Lauritzen (1996).

7.2.2 Number Theory Essentials

Number theoretic tools play an important role in accounting for all the operations that could be defined on a graph. Although the subject is rather deep, we only require a modest gathering of results for our problem.

The prime numbers will be denoted p_i, which represents the i-th prime number. The number 1 will be considered to be the 0-th prime number.

An arithmetic function is one of the form $g : \mathbb{N} \to \mathbb{C}$. A multiplicative function is an arithmetic function that also satisfies $g(m \cdot n) = g(m)g(n)$ if the greatest common divisor of m and n is 1. If it holds for all m, n the arithmetic function is said to be totally multiplicative. The sum-function of an arithmetic function g is defined as

$$S_g(n) := \sum_{d|n} g(d),$$

where the summation notation $\sum_{d|n}$ is standard notation for the sum over the divisors of n.

The fundamental theorem of arithmetic states that any $1 < n \in \mathbb{N}$ decomposes into a unique product of powers of primes: $n = p_{i_1}^{\alpha_{i_1}} \cdot, \cdots, \cdot p_{i_M}^{\alpha_{i_M}}$. The function $\lambda(n)$ will be used to denote the number of prime factors of n.

We require a few of the important number theoretic functions. One is the Möbius function: For $n = p_{i_1}^{\alpha_{i_1}} \cdot, \cdots, \cdot p_{i_M}^{\alpha_{i_M}}$,

$$\mu(n) = \begin{cases} 0 & \text{if any } \alpha_i > 1 \\ 1 & \text{if } n = 1 \\ (-1)^{\lambda(n)} & \text{otherwise} \end{cases} \tag{7.3}$$

The first condition tests whether or not the factorization of n contains a square number.

The constant function will be denoted $1(n)$ or simply 1 when the context makes it clear that a function, not a number, is required. The Dirichlet identity is defined as

$$\epsilon(n) = \begin{cases} 1 & \text{if } n = 1 \\ 0 & \text{otherwise.} \end{cases}$$

The Dirichlet convolution of two arithmetic functions, f and g, is another arithmetic function. It is denoted $f * g$ and defined as

$$(f * g)(n) = \sum_{d|n} f(d)g(\frac{n}{d}). \tag{7.4}$$

This operation is commutative and associative. The sum-function, defined earlier, is in fact the Dirichlet convolution $S_f = f * 1$. Two other important identities are: $\mu * 1 = \epsilon$, and, for all arithmetic functions, f, $f * \epsilon = f$ holds. Using these, we can easily derive the famous Möbius Inversion theorem:

$$\mu * S_f = \mu * (f * 1) = (\mu * 1) * f = \epsilon * f = f \tag{7.5}$$

This theorem plays an important role in the analysis. Our account of the number theoretic tools has been terse. To place them in their proper context one should consult a reference such as Wilf (1994) or Graham et al. (1989). The notes of Stankova-Frenkel (1999) are also helpful.

7.2.3 q-logarithm

q-log and q-exponential is included to make the chapter self-contained. This material is presented with much more background in the previous chapter. The q-logarithm is defined for $q > 0$ as

$$\log_q(p) := \begin{cases} \log(p) & \text{if } q = 1 \\ \frac{p^{1-q}-1}{1-q} & \text{otherwise} \end{cases} \tag{7.6}$$

We let the notation $(v)_+$ mean v if $v > 0$ and 0 otherwise. The inverse of the q-logarithm is

$$\exp_q(v) = (1 + (1-q)v)_+^{\frac{1}{1-q}} .$$

Using these two functions we can define an analogue to multiplication:

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

if $x^{1-q} + y^{1-q} - 1 > 0$ and otherwise it is 0. It is associative, commmutative and it has 1 as its neutral elements. Under this definition of $\otimes_q$ we have the identities:

$$\exp_q(x + y) = \exp_q(x) \otimes_q \exp_q(y)$$
$$\log_q(x \otimes_q y) = \log_q(x) + \log_q(y) \text{ (whenever the left-hand-side is defined).}$$

Thus a q-exponential does not factorize, but it does "q-factorize".

7.3 Main Result

With the background of the previous section we are ready to state the main result. We let the transformation $\mathcal{T}$ correspond to $\log_q$, leaving q as a parameter. Then it makes sense to refer to the decomposition and generalized Markov properties as q-decomposition and q-Markov properties. Later we will discuss other transformations $\mathcal{T}$ leading to different generalizations of the H-C theorem.

Theorem 7.1 (q-Hammersley-Clifford) *Let P be a positive measure with non-negative density f. Then f satisfies the pairwise q-Markov property iff f q-factorizes according to the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.*

Proof ($\Leftarrow$-direction) If f q-factorizes according to the graph and $a_m \neq a_n$ do not share an edge, then the expression in (7.1) can be broken apart into

$$\sum_{\{\, c | c \in \mathcal{C},\, a_m \in c \,\}} \psi_c(\boldsymbol{x}_c) + \sum_{\{\, c | c \in \mathcal{C},\, a_n \in c \,\}} \psi_c(\boldsymbol{x}_c) + \sum_{\{\, c | c \in \mathcal{C},\, a_m \notin c,\, a_n \notin c \,\}} \psi_c(\boldsymbol{x}_c)$$

The first term can serve as h_1 and the last two terms can serve as h_2 which verifies in the Markov condition (7.2).

($\Rightarrow$-direction) *Guide:* First we define a map or index from sets of vertices of the graph (equivalently the argument indexes of f) to the natural numbers. This map is simple to define, but special, because its value encodes the membership of the vertices in the set. In other words, knowing only a value of the index, we can determine the identity of all of the vertices which are contained in the set which has that index value. Using this index, we define a generic potential function, ψ, over the index values of each subset of $\mathcal{V}$. Next we employ the results from number theory. The function ψ is seen to be the Dirichlet convolution of the Möbius function μ and another function k. The Möbius inversion theorem provides that k is the sum-function S_ψ, also defined over the index values of all the subsets of $\mathcal{V}$. The inversion theorem provides a representation of k as a sum of potentials taken over *all* all possible subsets of $\mathcal{V}$.

Next we define a special version of f that switches the evaluation of f at a given point $\boldsymbol{x}$ and an arbitrary alternative point $\widehat{\boldsymbol{x}}$, depending on a supplied set of vertices, b. This function evaluates a vector using the coordinate values of $\boldsymbol{x}$ corresponding to b and the coordinate values of $\widehat{\boldsymbol{x}}$ corresponding to the complement, b^c. This function must therefore agree with f when $\mathcal{V}$ is the set under consideration. Otherwise it tests the sensitivity of f to arbitrary settings of variables outside the given set of coordinate indexes corresponding to b.

Next we set k to be this modified f evaluated at $\mathcal{V}$, and use the sum function representation of k to express f in terms of a sum of terms involving ψ. In this way, the modified f, which agrees with f at the level of the whole graph, is represented as a sum of local versions of itself. However, this representation contains *all* subsets of $\mathcal{V}$ in the sum. We then use the q-Markov property of f to show that any subset which does not correspond to a clique can be omitted from the sum. This will conclude the proof.

Proof: In the following a and b will be arbitrary subsets of $\mathcal{V}$ unless indicated otherwise, while a_m, a_n represent vertices of the graph and $M = |\mathcal{V}|$.

Define an index map $\iota : \mathcal{P}(\mathcal{V}) \to \mathbb{N}$ as follows. Let $\iota(\emptyset) = 1$. Let us consider 1 as the 0-th prime number. Choose an arbitrary ordering of the nodes $a_i, i = 1 \ldots M$, in $\mathcal{V}$. Define an index map ι as follows. For singleton sets, $b = \{\, a_i \,\}$, let

$\iota(\{\, a_i \,\}) = p_i$, the i-th prime number. For larger subsets, $b = \{a_{i_1} \ldots a_{i_j}\}$, let

$$\iota(b) = p_{i_1} \cdot p_{i_2} \cdots p_{i_{j-1}} \cdot p_{i_j}.$$

This map has certain helpful properties. The range of ι is square-free since no prime appears twice in the definition of $\iota(b)$. If $a \subset b \subset \mathcal{V}$ then $\iota(a)$ divides $\iota(b)$. The Möbius function (7.3), μ, takes on only the values ± 1 when evaluated on the range of ι, *i.e.,* if $j = \iota(a)$, then $\mu(\iota(a)) = (-1)^{|a|} = (-1)^{\lambda(j)}$, where λ is defined in (7.3).

Next, define a generic potential function over any subset of $\mathcal{V}$. Let

$$\psi(\iota(a)) = \sum_{j|\iota(a)} \mu\left(\frac{\iota(a)}{j}\right) k(j) = \sum_{j|\iota(a)} (-1)^{\frac{\iota(a)}{j}} k(j),$$

where k is an arithmetic function. This definition is equivalent to $\psi = \mu * k$ where "$*$" denotes Dirichlet convolution (7.4). The Möbius inversion theorem (7.5) implies that k is the sum-function of ψ, S_ψ. Thus

$$k(n) = S_\psi(n) = \sum_{j|n} \psi(j). \tag{7.7}$$

We will fix a particular choice of k in a moment.

Turning to $f(\boldsymbol{x})$, consider an arbitrary alternative setting of the variables, denoted $\widehat{\boldsymbol{x}}$. We define a *test of settings* as follows: $\boldsymbol{v} : \mathcal{X} \times \mathcal{X} \times \mathcal{P}(\mathcal{V}) \to \mathcal{X}$

$$\boldsymbol{v}(\boldsymbol{x}, \widehat{\boldsymbol{x}}, a) = (v_1 \ldots v_M) \text{ where} \tag{7.8}$$

$$v_i := \begin{cases} x_i & \text{if } \iota(a_i)|\iota(a) \\ \widehat{x}_i & \text{otherwise.} \end{cases} \tag{7.9}$$

The vector $\boldsymbol{v}$ matches $\boldsymbol{x}$ on coordinate indexes that correspond to vertices in a and matches $\widehat{x}$ on the other coordinates. Thus the test. $\boldsymbol{v}$ is not the usual notion of a mixture of values, rather it represents a mixture of settings. In this way $\boldsymbol{v}$ will permit measurement of the sensitivity of f to settings to groups of different entries in $\boldsymbol{x}$ outside of a given set. At the top level of the graph, $a = \mathcal{V}$, all such tests are exhausted, since we have $\boldsymbol{v}(\boldsymbol{x}, \widehat{\boldsymbol{x}}, \mathcal{V})) = \boldsymbol{x}$, and therefore $f(\boldsymbol{v}(\boldsymbol{x}, \widehat{\boldsymbol{x}}, \mathcal{V})) = f(\boldsymbol{x})$.

Now we are ready to set k concretely. For $a \subset \mathcal{V}$, let

$$k(\iota(a)) = \log_q f(\boldsymbol{v}(\boldsymbol{x}, \widehat{\boldsymbol{x}}, a)).$$

At the top level of the graph, we have (using (7.7))

$$k(\iota(\mathcal{V})) = \log_q f(\boldsymbol{x}) = \log_q f(\boldsymbol{v}(\boldsymbol{x}, \widehat{\boldsymbol{x}}, \mathcal{V})) = \sum_{j|\iota(\mathcal{V})} \psi(j).$$

Thus we have an expression for f as a sum-function. However, the summand in the last term is evaluated once for *each* subset of $\mathcal{V}$. Next we aim to show if j is not the index of a clique then $\psi(j)$ is zero and the sum can therefore be taken over cliques only.

Let $j = \iota(a)$ for $a \subset \mathcal{V}$, which is not a clique. Then a contains vertices a_m and a_n, which do not have an edge between them. Denote the remainder of the set, $c = a \backslash \{ a_m, a_n \}$. Noting that $\iota(a) = p_m \cdot p_n \cdot \iota(c)$, the potential for this subset can be written as

$$\begin{aligned}\psi(\iota(a)) &= \sum_{j|\iota(a)} (-1)^{\lambda\left(\frac{\iota(a)}{j}\right)} k(j) \\ &= \sum_{j|\iota(c)} (-1)^{\lambda\left(\frac{\iota(a)}{j}\right)} \left(k(j) - k(j \cdot p_m) - k(j \cdot p_n) + k(j \cdot p_m \cdot p_n)\right). \qquad (7.10)\end{aligned}$$

It is sufficient to show that each term in the sum is zero. Recall the pairwise q-Markov condition (7.2) as it applies to $f(\boldsymbol{v})$. Let $\bar{\boldsymbol{v}} = \boldsymbol{v}_{\mathcal{V} \backslash \{ a_m, a_n \}}$ correspond to $\boldsymbol{v}$ with the m- and n-th entries deleted. Since a_m and a_n are not connected, there exist functions h_1 and h_2 such that

$$\log_q f(\boldsymbol{v}(\boldsymbol{x}, \widehat{\boldsymbol{x}}, j)) = \begin{cases} h_1(\widehat{x}_m, \bar{\boldsymbol{v}}) + h_2(\widehat{x}_n, \bar{\boldsymbol{v}}) & \text{if NEITHER } p_m \text{ nor } p_n \text{ is a factor of } j \\ h_1(x_m, \bar{\boldsymbol{v}}) + h_2(\widehat{x}_n, \bar{\boldsymbol{v}}) & \text{if } p_m \text{ IS a factor of } j \text{ and } p_n \text{ is NOT a factor of } j \\ h_1(\widehat{x}_m, \bar{\boldsymbol{v}}) + h_2(x_n, \bar{\boldsymbol{v}}) & \text{if } p_m \text{ is NOT a factor of j and } p_n \text{ IS a factor of } j \\ h_1(x_m, \bar{\boldsymbol{v}}) + h_2(x_n, \bar{\boldsymbol{v}}) & \text{if BOTH } p_m \text{ and } p_n \text{ are factors of } i. \end{cases}$$

In (7.10), we note that each of the conditions above holds exactly once. Therefore the summand is

$$\begin{aligned} & k(j) - k(j \cdot p_m) - k(j \cdot p_n) + k(j \cdot p_m \cdot p_n) \\ &= h_1(\widehat{x}_m, \bar{\boldsymbol{v}}) + h_2(\widehat{x}_n, \bar{\boldsymbol{v}}) - h_1(x_m, \bar{\boldsymbol{v}}) - h_2(\widehat{x}_n, \bar{\boldsymbol{v}}) \\ &\quad -h_1(\widehat{x}_m, \bar{\boldsymbol{v}}) - h_2(x_n, \bar{\boldsymbol{v}}) + h_1(x_m, \bar{\boldsymbol{v}}) + h_2(x_n, \bar{\boldsymbol{v}}) \qquad (7.11) \\ &= 0. \end{aligned}$$

Since this is true for each term in (7.10), the proof is complete. ■

The theorem specializes to the Hammersley Clifford (Theorem 3.9 Lauritzen

(1996)) exactly when f is assumed to be a probability distribution and $q = 1$.

Remark: A corollary to H-C reduces the decomposition to one over *maximal* cliques, since each clique is a subset of at least one maximal clique.

7.4 Discussion

Other Graph Semantics. The proof technique of Theorem 7.1 can be adjusted to accommodate other graph semantics. One way is to replace $\log_q$ with a member of a larger class of *deformed logarithms*. A variety of non-SBG entropies studied in statistical physics fit into this framework Naudts (2004b). The families of distributions associated with these entropies are known as phi-exponential families. Let $\phi : [0, \infty) \to [0, \infty)$ be strictly positive and non-decreasing on $(0, \infty)$. Define $\log_\phi$ via

$$\log_\phi(p) := \int_1^p \frac{1}{\phi(y)}\, dy$$

If this integral converges for all finite $p > 0$, then $\log_\phi$ is called a *deformed logarithm*. Note that $\phi(p) = p^q$ recovers $\log_q$.

Dynamic Programming The method of optimization by combining optimal solutions to simpler subproblems, is often called dynamic programming. The key step to use dynamic programming involves expressing the desired computation as a sum of products in a semi-ring, after which the distributive law can be used to diminish the required number of multiplications Aji and McEliece (2000). The choice of semi-ring corresponds to the task being performed with the graph. The goal of the Viterbi algorithm is to find the most likely setting of variables, given a joint distribution. For that task it performs dynamic programming using the max-product semi-ring. The q-product discussed here is compatible with a Viterbi-like algorithm since q-factorization is compatible with max *i.e.,* $a \otimes_q \max(b, c) = \max(a \otimes_q b, a \otimes_q c)$. This is true since if $z > 0$, then $x \leq y$ implies $z \otimes_q x \leq z \otimes_q y$.

'Broadcast node' extension. There are other decompositions where not all the contributions of non-cliques get set to zero. The proof technique of Theorem 7.1 can also be adjusted to accommodate these graph semantics. In general, we do this by letting $\psi = \psi_1 + \psi_2$, in which case we will have $k = S_{\psi_1+\psi_2} = 1 * (\psi_1 + \psi_2) = 1 * \psi_1 + 1 * \psi_2 = S_{\psi_1} + S_{\psi_2} = k_1 + k_2$. In this case k becomes a mixture of sum-potentials. The "Markov condition" condition for each type of potential can then be stated independently. The sum expression for k can be taken separately over the two types of potentials, will hopefully (or by design) further reduce to sums over different types of sets. They do not necessarily have to decompose over the cliques of the graph.

For example ψ_1 might correspond to an ordinary Markov model, with, say $T = log$, decomposing over cliques. On the other hand ψ_2 could include an effect that has each node influenced by an exchangeable model over all of its non-neighbors. In the latter submodel, the sum-potential would decompose over the non-neighbors of each node, effectively adding a single extra 'broadcast' node attached to each of the original nodes of the graph, representing some summary statistic of all of its non-neighbors. This aspect is left for future potential research.

Conclusion. The theory that has been presented in this chapter suggests the possibility of using non-exponential-family distributions, such as those derived via generalized maximum entropy, in conjunction with graphical models.

To do so we have had to introduce the the notion of q-factorization. Some authors have been intrigued by the algebraic issues suggested by this operation see e.g. Borges (2004) and have extended it to other operations such as q-derivation, q-division, etc. The difficulty here is that the operations severely restrict the domain where the algebraic properities hold[1]. It is difficult to see how to maintain these properties since they would depend on the value of q and other model inputs. Such algebraic approaches are even more difficult to sustain if we are ultimately interested in applying generalization based on ϕ-logarithms discussed in Chapter 4.

To employ the q-exponential distributions, or ϕ-exponentials for that matter, in a graphical model setting requires reinterpretation of the information encoded in the graph adjacency structure and the node distributions, away from conditional independence, and the marginals of the joint distribution, respectively. We believe that in some applications, this may correspond to the type of prior information that is available. In the future it may be possible to extend the idea to utilize more of the techniques of graphical models and to address the issue of updating.

[1]Thanks to one of my examiners for pointing this out

Chapter 8

Conclusion

This chapter summarizes the thesis and reflects on some research directions suggested by it. I also indulge in a few impressions I have about the field. Any opinions expressed are my current opinion only, and the reader should feel free to attempt to change my mind.

8.1 Synopsis

Understanding the duality between exponential families and SBG entropy sparked the original motivation for this research. In generalizing this connection, my aim was simply to "borrow" some alternative relative entropy functions from the physics literature to generate better fitting models for statistical applications. Digging deeper, I found that the class of functions proposed by Naudts (Chapter 4) to be appealing, since they are carefully constructed to have certain "nice" properties. In particular, the gradients have a closed form.

This class of functions helped to illuminate connections between three different fields: convex analysis (*e.g.* Pietra et al. (2002a), Bauschke (2003)), statistical physics (*e.g.* Naudts (2004a,b, 2002)), and information geometry (*e.g.* Murata et al. (2004); Eguchi (2005)). These connections were explored in Chapter 5. This was made possible with elementary tools from convex duality and Fenchel duality, which allowed us to replicate the main result from each paper. Certainly it was straightforward to construct the normalizing function T and discover the need for escort probabilities in this framework. This speaks to the elegance and simplicity of the duality that was laid out in Chapters 1 to 3.

Most of the elegant results of information geometry seem to rely on the Legendre property of the entropy. But this property is not really necessary for optimization. In the compressed sensing literature, here represented by the Dantzig selector example (Chapter 5), and also in the NMF example (Chapter 6), using non-Legendre entropy functions yields distributions with limited support, something that exponential families are not 'designed' to do. (See also Grunwald and Dawid (2004); Dawid and Lauritzen (2005) for some initial results in this direc-

tion.) The non-Legendre functions correspond to aggressive learning and sparse models, features which are useful to machine learners as Chapter 6 indicates.

Stepping outside of the exponential families has its costs. For one thing, they correspond well to graphical models, with the Hammersley-Clifford theorem establishing a strong theoretical connection, and many algorithms have been designed to exploit the factorization properties of this type of model (*e.g.* Wainwright and Jordan, 2003b,a; Lauritzen, 1996; Pakzad and Anantharam, 2004; Aji and McEliece, 2000). Obvious questions to pursue are do phi-exponential families also have such a natural correspondence, and in what sense does the distribution decompose? An initial answer is given in Chapter 7. A form of decomposition can be established, but the resulting distribution is no longer a product of marginals. Independence assumptions encoded in the graph are replaced by 'q-independence', and q becomes an index of global correlation.

In addition, the use of explicit normalization, rather than embedding it in a partition function, never to be relaxed, makes it possible to consider models that are non-probabilistic. In fact, this distinction could be considered an artificial characterization of what is actually a continuum. For example, in a finance setting, it is common to identify prices with probabilities. On the other hand, a market-maker often faces the task of setting prices for a number of securities simultaneously, not knowing which will result in a transaction. In that case, the task for a model might be to output *two* sets of prices: bids and offers, with the stipulation that they sum to < 1 and > 1, respectively. As we all are taught, probabilities must sum to one, or else they are not probabilities. It is fine to be pedantic about this as long as we recognize how useful summing to *approximately* 1 can be.

8.2 Sequels

The setting explored in this thesis could be furthered in a number of directions. Though it has been termed 'generalized maxent', some aspects do not represent the ultimate generalization that is possible. It remains to future work to push these aspects further, or incorporate more ideas that have already appeared or been hinted at in the literature.

Infinite dimensions. By building the framework in a discrete setting we gain the considerable advantage of a simplified yet powerful theory about convex optimization. One way to address infinite state spaces is to reformulate the problem with $\boldsymbol{p}$ as an element of a Banach space, and $\boldsymbol{A}$ as an operator. Convex optimization theory does extend to such spaces, although constraint qualifications become much harder to state and verify (Bauschke et al., 2000; Altun and Smola, 2006). However, this reformulation does not comprehensively address methods of optimization, nor does it currently address non-Legendre entropy functions.

If we start from the finite dimensional problem and let its dimension (N) grow, it is possible to identify exactly where the difficulties are likely to arise and focus the discussion on techniques from approximation theory that could solve or mitigate the problem.

To review, a typical (dual) objective function will have the following form:

$$F^*(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\mu} + \boldsymbol{p}_0^*) + \langle \boldsymbol{\mu}, \boldsymbol{b} \rangle + G^*(-\boldsymbol{\mu}).$$

where G^* is meant as a proxy to stand in for part of the conjugate of the relaxation term in the primal (generalized maxent) problem. Next, it is is useful to examine the first order condition for optimality. A nearly identical equation is used in the update step for optimization algorithms:

$$\boldsymbol{A}\nabla F^*(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\mu} + \boldsymbol{p}_0^*) - \boldsymbol{b} + \nabla G^*(-\boldsymbol{\mu}) = \boldsymbol{0}. \tag{8.1}$$

Note that $\boldsymbol{b}$ is to be regarded as a fixed parameter of the last term and the gradient is taken with respect to $\boldsymbol{\mu}$. Calculating this condition is critical for evaluating algorithm speed. In fact, the software implemented for this thesis includes code that performs this calculation. The key issue for scalability is maintaining the ability to represent $\boldsymbol{A}$ and its transpose in memory and performing the calculations necessary to evaluate (8.1).

Let us take a moment to carefully identify what calculations are needed.

$$\boldsymbol{v}_1 = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\mu}, \text{ a } N \times 1 \text{ vector} \tag{8.2}$$

$$\boldsymbol{v}_2 = F(\boldsymbol{v}_1) = f[\boldsymbol{v}_1], \text{ another } N \times 1 \text{ vector} \tag{8.3}$$

$$\boldsymbol{v}_3 = \boldsymbol{A}\boldsymbol{v}_2, \text{ an } M \times 1 \text{ vector} \tag{8.4}$$

The equations represent potential bottlenecks in the computation. Each of these may be a place to approximate or achieve some type of speedup. Any approximations would be passed downstream to the next calculation listed. There are a number of approaches to deal with the approximation problem, e.g. sparse grids (Garcke et al., 2006; Hegland, 2003).

An alternative idea commonly used in machine learning is to use a discriminative model. This limits the dimensionality of the model to the sample size, albeit at the cost of using a conditional model and over-weighting the parts of the state space that are actually sampled.

Entropy functions and divergence measures. Other proposals for entropy functions have been mentioned in the thesis. There are also many candidates, besides Bregman divergence that could be used to construct a distance measure (see Zhang, 2004). In some cases these can be easily accommodated in the analysis. For example, let us briefly reconsider Csiszar's divergence, briefly introduced in section 2.3.2. We note in passing that only for $q = 1$ do Csizar's and Bregman

divergence coincide as the special case of KL-divergnce. If we replace the Bregman divergence with Csiszar the conjugate term of the dual problem is $\boldsymbol{p}_0^{\mathrm{T}} f^*[\boldsymbol{p}^*]$. This give distributions of the form:

$$\bar{\boldsymbol{p}}_i = (\boldsymbol{p}_0)_i f^{*'}[\boldsymbol{A}^{\mathrm{T}}\bar{\boldsymbol{\mu}}]_i.$$

Comparing this to the Bregman version (3.26), where the initial guess is always on the inside ($\boldsymbol{p}_0^*$), we can see that the Csiszar's formulation preserves the multiplicative factorization with respect to the initial guess, $\boldsymbol{p}_0$.

We can draw the following qualitative conclusions comparing the q-exponential families. In the case of $q > 1$, the q-exponential family can still place positive weight on states that $\boldsymbol{p}_0$ places zero weight on. This is ruled out when using Csiszar's divergence. In the case of $q < 1$, the q-exponential family can place zero weight on some states. This is true under either the Csiszar or Bregman formulation.

Relaxations/priors/losses. As we have seen, when exact constraint matching is relaxed, new terms begin to appear in the dual problem. In statistical terminology, these can be thought of as priors or statistical loss measures, depending on one's perspective. Many forms of constraint relaxation have been explored in statistics and machine learning (e.g. Chen and Rosenfeld, 2000; Golan and Gyzl, 2002; Kazama and Tsujii, 2005; Goodman, 2004; Brand, 1998; Cheeseman and Stutz, 2005) and is probably the most fully explored generalization addressed in empirical work.

Machine learning algorithms make heavy use of many other *non-smooth* loss functions, besides the SVM loss presented. Even the simple norm relaxation employed in (1.4) is not differentiable at zero. In Teo et al. (2007) a table of 20 other common loss functions is presented and only 5 are $\mathbb{C}^1$ or smoother. Non-smoothness hinders some analytical statements, but does not represent overwhelming practical problems. Smooth solvers can be modified to handle some non-smooth problems (Haarala et al., 2004, 2007) and ongoing work in unifying non-smooth solvers for problems of the type discussed is promising (Teo et al., 2007). One can reasonably expect to see more software combining Bregman divergences and non-smooth loss functions in the near future.

8.3 Prequels

Our focus throughout has been on generalized maxent, but we could have focused on other induction principles. The most important question for machine learning is ultimately how to build an intelligent agent that can model its environment (see Russell and Norvig, 2003). Unfortunately no induction principle offers a complete solution to this problem, which remains wide open.

It is, however, a fascinating exercise to compare some of the proposals. Many principles have been put forth and applied in machine learning. Minimum Description Length (see Grünwald, 1998, for a tuorial), Kolmogorov Complexity (Li and Vitányi, 1997), Bayesian methods (Tipping, 2004), and Statistical Learning Theory (Vapnik, 1998) represent a few of the alternatives. Working out the relationships between these seems a field in itself. Even the relationship between two closely related fields, maxent and Bayesian methods, is still an active area (Cheeseman and Stutz, 2004; Giffin and Caticha, 2007; Caticha, 2003).

As for maxent, the appeal of building an induction principle around SBG entropy has been taken up by philosophers (Paris, 1994; Williamson, 2005) and physicists alike (Caticha, 2003; Giffin and Caticha, 2007). It has been motivated axiomatically (Shore and Johnson, 1980; Csiszar, 1975, 1991), via game theory (Harremoes and Topsøe, 2001) and various other ways (Grendar and Grendar, 2000; Knuth, 2005). The historical development of SBG entropy points to combinatorial arguments as a foundation (see Jaynes, 1982b; Niven, 2007), with the entropy function seen as an approximation to a combinatorial expression. In the Tsallis literature, similar motivations have been investigated (Almeida, 2001; Borges and Roditi, 1998; Furuichi, 2005b,a).

These motivations have led some to the characteristic algebraic property of the q-logarithm: $\log_q(x\,y) = \log_q(x) + \log_q(y) + (1-q)\log_q(x)\log_q(y)$ and a corresponding one for the q-exponential: $\exp_q(x+y+(1-q)xy) = \exp_q(x)\exp_q(y)$. These properties in turn lead to many of the properties of found in solutions to problems that utilize these terms in an objective function. However, better motivations for choosing an entropy function would be highly desirable and is an ongoing area of research (see Niven, 2007) and the topic of a recent conference (Facets of Entropy, Copenhagen, 2007). While the selection of an objective function is important, this topic is both deep and wide, so here we have emphasized the consequences of that choice, rather than its motivation. In practice the consequences are likely to be just as important to the machine learning community.

Despite the large effort necessary to understand and integrate different model building principles, it is striking how similar many turn out to be in practice. The similarity of different induction principles when they are put into practice has been noted elsewhere (Lanterman, 2000).

One reason models can be hard to compare is that they might be parameterized in widely different ways. This can often disguise the elements they have in common. Since virtually all machine learning algorithms produce outputs, if we can discover a selection criteria for the outputs, the model formulations might be comparable on this basis. This point of view was pursued by Rifkin and Lippert (2007), who found that some common techniques in machine learning can be

viewed as solving

$$\min_{\boldsymbol{y}} \quad R(\boldsymbol{y}) + L(\boldsymbol{y}),$$

where $\boldsymbol{y}$ represents an *output* of the model. The two functions involved are a convex regularizer R, which favors smooth (small norm) outputs, regardless of the data vector, $\boldsymbol{b}$. Meanwhile the loss L, which typically makes use of the data vector $\boldsymbol{b}$ as a parameter, is used to enforce a close fit to the data. They term this view of the model *value regularization*.

Although maxent is not addressed in Rifkin and Lippert (2007), it is not hard to work out how the generalized maxent problem (1.4) can be mapped into this framework. In fact, the dual problem is unaffected. For the primal, the first step is a change of variables $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{p}$. This makes it clear that the 'outputs' under discussion are expectations of the rows of $\boldsymbol{A}$ in the generalized maxent context. The loss is then just $L(\boldsymbol{y}) = \delta_{\epsilon B_p}(\boldsymbol{y} - \boldsymbol{b})$. Describing the regularizer requires another device from convex analysis, the *image function*, denoted by the compound symbol $A\Delta F$:

$$R(\boldsymbol{y}) = (A\Delta F)(\boldsymbol{y}) := \min_{\boldsymbol{p}} \quad \Delta F(\boldsymbol{p}, \boldsymbol{p}_0) + \delta_{\{0\}}(\boldsymbol{A}\boldsymbol{p} - \boldsymbol{y}).$$

It should be noted that $A\Delta F$ is also a well-behaved convex function on the set of feasible points for $\boldsymbol{y}$. Combining the two components yields a new primal problem:

$$\min_{\boldsymbol{y}} \quad (A\Delta F)(\boldsymbol{y}) + \delta_{\epsilon B_p}(\boldsymbol{y} - \boldsymbol{b}) \tag{8.5}$$

We are now in a position to compare the maxent model to one of the great workhorses of machine learning, the Support Vector Machine (SVM). Two key definitions are required. The first is $\boldsymbol{K}$, a positive semidefinite kernel matrix, constructed from the training examples. The second is the loss function employed by SVMs, the soft-margin hinge loss: $L(y, b) = \max(0, 1 - yb)$, where y is the output label and b is the training target. Under the value regularization framework (see Rifkin and Lippert (2007) for details) it can be shown that SVMs solve the primal problem

$$\min_{\boldsymbol{y}} \quad \underbrace{\frac{1}{2}\lambda \boldsymbol{y}^{\mathrm{T}} \boldsymbol{K}^{-1} \boldsymbol{y}^{\mathrm{T}}}_{R} + \underbrace{\sum_i L(y_i, b_i)}_{L}. \tag{8.6}$$

The SVM regularizer is minimized if the output vector, $\boldsymbol{y}$, is smooth, since $\boldsymbol{K}^{-1}$ is positive semi-definite. However this is balanced against the SVM loss, which is minimized only when the training output match the labels. The mathematical

techniques used to originate this model and the philosophical orientation of its originators are rather distinct from the maxent approach (see Schölkopf and Smola (2001)). But as this example makes clear, all those considerations, different as they may be, lead to models that are conceptually not too far apart, differing only in the choice of regularizer and loss function.

I do not mean to suggest that *all* machine learning problems fit into this framework. Many problems lead to non-convex optimization (*e.g.,* missing variables, NMF). Nor has a fully Bayesian treatment been brought into the picture.

After examining various model formulations, hopefully the reader will agree that convex analysis helps clear away the clutter and expose what they have in common. Having done so, by comparing (8.5) and (8.6) it can be seen that SVMs and maxent models are only "two substitutions" apart. In the language of machine learning the crux of the distinction seems to be in how to choose a regularizer and how to choose a loss. The loss can be viewed as a means of relinquishing trust in the data, while the regularizer reflects how smooth the answer is required to be. This operation seems to be fundamental in some sense. Techniques for trading off these two competing goals have been invented independently and repeatedly in the past. As optimization algorithms continue to improve, perhaps we can look forward to machine learning researchers making greater use of entropic regularization variants such as the ones described in this thesis, while maxent researchers may benefit from designing loss functions directly as is common in machine learning.

Bibliography

Srinivas M. Aji and Robert J. McEliece. The generalized distributive law. *IEEE Transactions on Inforamtion Theory*, 46(2):325–343, March 2000.

MP Almeida. Generalized entropies from first principles. *Physica A: Statistical Mechanics and its Applications*, 300(3-4):424–432, 2001.

Yasemin Altun and Alexander J. Smola. Unifying divergence minimization and statistical inference via convex duality. In *COLT*, pages 139–153, 2006.

Shun-Ichi. Amari and H. Nagaoka. *Methods of Information Geometry*. Oxford University Press, 1993.

K.S. Azoury and M. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43:211–246, 2001.

Satish Balay, William D. Gropp, Lois Curfman McInnes, and Barry F. Smith. Efficient management of parallelism in object oriented numerical software libraries. In E. Arge, A. M. Bruaset, and H. P. Langtangen, editors, *Modern Software Tools in Scientific Computing*, pages 163–202. Birkhäuser Press, 1997.

Satish Balay, Kris Buschelman, William D. Gropp, Dinesh Kaushik, Matthew G. Knepley, Lois Curfman McInnes, Barry F. Smith, and Hong Zhang. PETSc Web page, 2001. http://www.mcs.anl.gov/petsc.

Satish Balay, Kris Buschelman, Victor Eijkhout, William D. Gropp, Dinesh Kaushik, Matthew G. Knepley, Lois Curfman McInnes, Barry F. Smith, and Hong Zhang. PETSc users manual. Technical Report ANL-95/11 - Revision 2.1.5, Argonne National Laboratory, 2004.

A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh. Clustering with Bregman Divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

H. H. Bauschke. Duality for Bregman projections onto translated cones and affine subspaces. *Journal of Approximation Theory*, 121(1):1–12, 2003.

H. H. Bauschke and J. M. Borwein. Legendre functions and the method of random Bregman projections. *Journal of Convex Analysis*, 4(1):27–67, 1997.

Heinz Bauschke and Martin von Mohrenschildt. Fenchel conjugates and subdifferentials in maple. *MapleTech*, 1999.

Heinz H. Bauschke and Martin v. Mohrenschildt. Symbolic computation of fenchel conjugates. *ACM Commun. Comput. Algebra*, 40(1):18–28, 2006. ISSN 1932-2240.

Heinz H. Bauschke, Jonathan M. Borwein, and Patrick L. Combettes. Essential smoothness, essential strict convexity, and Legendre functions in Banach spaces, 2000.

H.H. Bauschke and J.M. Borwein. Joint and separate convexity of the Bregman distance. *Inherently parallel algorithms in feasibility and optimization and their applications*, 8:23–36, 2000.

Steve Benson, Lois Curfman McInnes, Jorge Moré, Todd Munson, and Jason Sarich. TAO user manual (revision 1.9). Technical Report ANL/MCS-TM-242, Mathematics and Computer Science Division, Argonne National Laboratory, 2007. http://www.mcs.anl.gov/tao.

Adam L Berger, Stephen A. della Pietra, and Vincent J. della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71, 1996.

Guy Le Besnerais, Jean-Francois Bercher, and Guy Demoment. A new look at entropy for solving linear inverse problems. *IEEE Trans. on Information Theory*, 45, 1999.

E.P. Borges. A possible deformed algebra and calculus inspired in nonextensive thermostatistics. *Physica A: Statistical Mechanics and its Applications*, 340 (1-3):95–101, 2004.

Ernesto Borges and Itzhak Roditi. A family of nonextensive entropies. *Physics Letters A*, 246:399–402, 1998.

J. Borwein, R. Choksi, and P Marechal. Probability distributions of assets inferred from option prices via the principle of maximum entropy. *SIAM J, Optim.*, 14(2), 2003.

Jonathan Borwein and Adrian Lewis. *Convex Analysis and Nonlinear Optimization: Theory and Examples*. Springer, 2000.

Jonathan M. Borwein and Mark A. Limber. On entropy maximization via convex programming, 1996.

Jonathan M. Borwein and Qiji J. Zhu. *Techniques of Variational Analysis*. CMS. Springer, 2005.

S. Boyd and L. Vandenberghe. *S. Boyd and L. Vandenberghe*. Cambridge University Press, 2004.

Matthew Brand. Pattern discovery via entropy minimization. Technical Report TR-98-21, MERL, 1998.

E. Candes and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Stat.*, accepted, 2007.

E.J. Candes. Compressive sampling. *Proceedings of the International Congress of Mathematicians, Madrid, Spain*, 2006.

Ariel Caticha. Relative entropy and inductive inference. In *MaxEnt 2003*, 2003.

Y. Censor and S. A. Zenios. *Parallel Optimization: Theory, Algorithms and Applications*. Oxford University Press, 1997.

P. Cheeseman and J. Stutz. On the relationship between Bayesian and maximum entropy inference. *24th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, 735:445–461, 2004.

P. Cheeseman and J. Stutz. Generalized Maximum Entropy. *AIP Conference Proceedings*, 803:374, 2005.

SF Chen and R. Rosenfeld. A survey of smoothing techniques for ME models. *Speech and Audio Processing, IEEE Transactions on*, 8(1):37–50, 2000.

A. Cichocki, R. Zdunek, and S. Amari. Csiszar's Divergences for Non-Negative Matrix Factorization: Family of New Algorithms. *Lect. Notes Comput. Sci*, 3889:32–39, 2006.

A. Cohen, W. Dahmen, and R. DeVore. A Taste of Compressed Sensing. In *Proceedings of SPIE*. SPIE, 2007.

Michael Collins, Robert E. Schapire, and Yoram Singer. Logistic regression, adaboost and bregman distances. *Machine Learning*, 48(1-3):253–285, 2002. ISSN 0885-6125.

Thomas Cover and Joy Thomas. *Elements of Information Theory*. Wiley, 1991.

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

Imre Csiszar. Why least squares and maximum entropy? an axiomatic approach to inference for linear inverse problems. *Ann. Stat.*, 19(4):2032–2066, 1991.

A. Philip Dawid and Steffen L. Lauritzen. The geometry of decision theory. In *2nd International Symposium on Information Geometry and its Applications*, 2005.

R.A. DeVore. Deterministic Constructions of Compressed Sensing Matrices, 2007.

I.S. Dhillon and S. Sra. Generalized nonnegative matrix approximations with Bregman divergences. *Proceeding of the Neural Information Processing Systems (NIPS) Conference, Vancouver, BC*, 2005.

David Donoho and Victoria Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *NIPS 16*, Cambridge, MA, 2003. MIT Press.

David L. Donoho, Iain M. Johnstone, Jeffrey C. Hoch, and Alan S. Stern. Maximum entropy and the nearly black object. *Journal of the Royal Statistical Society*, 54(1), 1992.

DL Donoho. Compressed Sensing. *Information Theory, IEEE Transactions on*, 52(4):1289–1306, 2006.

Miroslav Dudík and Robert E. Schapire. Maximum entropy distribution estimation with generalized regularization. In *COLT*, pages 123–138, 2006.

S. Eguchi. Information Geometry and Statistical Pattern Recognition. *Sugaku Exposition, Amer. Math. Soc.*, 2005.

T. Feng, SZ Li, H.Y. Shum, and H.J. Zhang. Local non-negative matrix factorization as a visual representation. *Development and Learning, 2002. Proceedings. The 2nd International Conference on*, pages 178–183, 2002.

S. Furuichi. On uniqueness Theorems for Tsallis entropy and Tsallis relative entropy. *Information Theory, IEEE Transactions on*, 51(10):3638–3645, 2005a.

Shigeru Furuichi. Fundamental properties on Tsallis entropies, 2005b.

J. Garcke, M. Hegland, and O. Nielsen. Parallelisation of sparse grids for large scale data analysis. *ANZIAM J*, 48:11–22, 2006.

Murray Gell-Mann and Constantino Tsallis, editors. *Nonextensive Entropy*. Sante Fe Institute Studies in the Sciences of Complexity. Oxford University Press, 2004.

A. Giffin and A. Caticha. Updating Probabilities with Data and Moments. *27th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Saratoga Springs, NY*, 2007.

A. Golan and Henryk Gyzl. A generalized maxentropic inversion procedure for noisy data. *Applied Math. and Computation*, 127(2-3), 2002.

A. Golan, G.G. Judge, and D. Miller. *Maximum entropy econometrics: robust estimation with limited data*. Wiley, 1996.

J. Goodman. Exponential Priors for Maximum Entropy Models. North American ACL, 2004, 2004.

R.L. Graham, D.E. Knuth, and O. Patashnik. *Concrete mathematics*. Addison-Wesley Reading, Mass, 1989.

Marian Grendar and Marian Grendar. What is the question that maxent answers? : A probabilistic interpretation, 2000.

T.L. Griffiths and M. Steyvers. Prediction and semantic association. *Advances in Neural Information Processing Systems*, 15:11–18, 2003.

P. Grunwald and A. Dawid. Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. *Annals of Statistics*, 32(4): 1367–1433, 2004.

P.D. Grünwald. *The Minimum Description Length Principle and Reasoning Under Uncertainty*. Institute for Logic, Language and Computation, 1998.

N. Haarala, K. Miettinen, and M. M. Mäkelä. New limited memory bundle method for large-scale nonsmooth optimization. *Optimization Methods and Software*, 19(6):673–692, 2004.

N. Haarala, K. Miettinen, and M. M. Mäkelä. Globally convergent limited memory bundle method for large-scale nonsmooth optimization. *Mathematical Programming*, 109(1):181–205, 2007.

P. Harremoes and F. Topsøe. Maximum entropy fundamentals. *Entropy*, 3:191–226, 2001.

T. Hastie, R. Tibshirani, and J.H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2001.

M. Hegland. Adaptive sparse grids. *ANZIAM J*, 44, 2003.

M. Heiler and C. Schnörr. Learning Sparse Representations by Non-Negative Matrix Factorization and Sequential Cone Programming. *The Journal of Machine Learning Research*, 7:1385–1407, 2006.

Jean-Baptiste Hiriart-Urruty and Claude Lemarechal. *Fundamentals of Convex Analysis*. Springer, 2001.

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 1985.

P.O. Hoyer. Non-negative Matrix Factorization with Sparseness Constraints. *The Journal of Machine Learning Research*, 5:1457–1469, 2004.

A. Hyvarinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. J. Wiley New York, 2001.

A. Jakulin and W. Buntine. Analyzing the US Senate in 2003: Similarities, Networks, Clusters and Blocs, 2004.

E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106 (4):620–630, 1957a.

E. T. Jaynes. On the rationale of maximum entropy methods. *IEEE*, 70(9): 939–952, 1982a.

E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106 (4):620–630, May 1957b.

E. T. Jaynes. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70(9):939–952, September 1982b.

Tony Jebara. *Machine Learning : Discriminative and Generative*. Kluwer international series in engineering and computer science. Springer, Boston, 2003. ISBN 1402076479.

G. Kaniadakis, M. Lissia, and A.M. Scarfone. Deformed logarithms and entropies, 2004.

G. Kaniadakis, M. Lissia, and A. M. Scarfone. Two-parameter deformations of logarithm, exponential, and entropy: A consistent framework for generalized statistical mechanics. *Phys. Rev. E*, 71(046128):1–12, 2005.

J. Kazama and J. Tsujii. Maximum Entropy Models with Inequality Constraints: A Case Study on Text Categorization. *Machine Learning*, 60(1):159–194, 2005.

J. Kivinen and M.K. Warmuth. Exponentiated Gradient versus Gradient Descent for Linear Predictors. *Information and Computation*, 132(1):1–63, 1997.

K.H. Knuth. Lattice duality: The origin of probability and entropy. *Neurocomputing*, 67:245–274, 2005.

J. Lafferty. Additive models, boosting, and inference for generalized divergences. *Proceedings of the twelfth annual conference on Computational learning theory*, pages 125–133, 1999.

J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conf. on Machine Learning*, pages 282–289. Morgan Kaufmann, San Francisco, CA, 2001.

A.N. Langville, C.D. Meyer, and R. Albright. Initializations for the Nonnegative Matrix Factorization. In *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006.

Aaron D. Lanterman. Schwarz, Wallace, and Rissanen: Intertwining Themes in Theories of Model Selection, 2000.

S.L. Lauritzen. *Graphical models: Clarendon Press*. Clarendon Press, 1996.

Guy Lebanon and John Lafferty. Boosting and maximum likelihood for exponential models. In *NIPS*, volume 14, 2001.

D.D. Lee and H.S. Seung. Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems*, 13:556–562, 2001.

D.D. Lee and H.S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

M. Li and P. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer, 1997.

J.W. Lloyd and T.D. Sears. An architecture for rational agents. In *Declarative Agent Languages and Technologies*, Lecture Notes in Artificial Intelligence (forthcoming). Springer, 2005. URL http://www.doc.ic.ac.uk/~ue/DALT-2005/home.shtml.

David G. Luenberger. *Optimization By Vector Space Methods*. John Wiley & Sons, 1969.

M. Masi. A step beyond Tsallis and Renyi entropies. *Physics Letters A*, 338: 217–224, 2005.

J.J. Moré and D.J. Thuente. Line search algorithms with guaranteed sufficient decrease. *ACM Transactions on Mathematical Software (TOMS)*, 20(3):286–307, 1994.

N. Murata, T. Takenouchi, T. Kanemori, and S. Eguchi. Information geometry of u-boost and divergence. *Neural Compuatation*, 16:1437–1481, 2004.

Jan Naudts. Deformed exponentials and logarithms in generalized thermostatistics. *Physica A*, 316:323–334, 2002.

Jan Naudts. Generalized thermostatistics based on deformed exponential and logarithmic functions. *Physica A*, 340(1-3):32–40, 2004a.

Jan Naudts. Estimators, escort proabilities, and ϕ-exponential families in statistical physics. *Journal of Inequalities in Pure and Applied Mathematics*, 5 (4) 102(4), 2004b.

Robert K. Niven. Origins of the combinatorial basis of entropy. *eprint arXiv: 0708.1861*, 2007. URL http://www.citebase.org/abstract?id=oai:arXiv.org:0708.1861.

P. Paatero. Least squares formulation of robust non-negative factor analysis. *Chemometrics and Intelligent Laboratory Systems*, 37(1):23–35, 1997.

P. Paatero and U. Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(111-126):1, 1994.

Payam Pakzad and Venkat Anantharam. A new look at the generalized distributive law. *IEEE Transactions on Information Theory*, 50(6):1132–1155, June 2004. URL http://www.eecs.berkeley.edu/~ananth/2002+/Payam/cisspaper1.pdf.

J. B. Paris. *The Uncertain Reasoner's Companion*. Cambridge Univ. Press, 1994.

A. Pascual-Montano, JM Carazo, K. Kochi, D. Lehmann, and RD Pascual-Marqui. Nonsmooth Nonnegative Matrix Factorization (nsNMF). *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(3):403–415, 2006.

S.J. Phillips, R.P. Anderson, and R.E. Schapire. Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190(3-4):231–259, 2006.

Steven J. Phillips, Miroslav Dudik, and Robert E. Schapire. A maximum entropy approach to species distribution modeling. In *Proceedings of the 21 st International Conference on Machine Learning*, 2004.

S. Della Pietra, V. Della Pietra, and J. D. Lafferty. Duality and auxillary functions for Bregman distances. Technical Report CMU-CS-01-109R, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, 2002a.

Stephen Della Pietra, Vincent Della Pietra, and John Lafferty. Inducing features of random fields. Technical report, CMU, Pittsburgh, PA, USA, 1995.

Stephen Della Pietra, Vincent Della Pietra, and John Lafferty. Duality and auxiliary functions for Bregman distances. Technical Report CMU-CS-01-109R, CMU, 2002b.

MD Plumbley. Algorithms for nonnegative independent component analysis. *Neural Networks, IEEE Transactions on*, 14(3):534–543, 2003.

R.M. Rifkin and R.A. Lippert. Value regularization and fenchel duality. *Journal of Machine Learning Research*, 8:441–479, 2007.

R. T. Rockafellar. *Convex Analysis*, volume 28 of *Princeton Mathematics Series*. Princeton University Press, 1970.

R. Tyrrell Rockafellar and Roger J-B. Wets. *Variational Analysis*. Springer, 1998.

Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2nd edition, 2003.

I. Sandberg. Global implicit function theorems. *Circuits and Systems, IEEE Transactions on*, 28(2):145–149, 1981.

R. Schapire. The boosting approach to machine learning: An overview. In *MSRI Workshop on Nonlinear Estimation and Classification, Berkeley, CA, Mar. 2001*, 2001.

Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Adaptive computation and machine learning. The MIT Press, 2001.

Timothy D. Sears. From Maxent to Machine Learning and Back. In Kevin Knuth, editor, *26th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, 2007.

Timothy D. Sears and Peter Sunehag. Induced semantics for undirected graphs. In Kevin H. Knuth, editor, *26th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, 2007.

Timothy D. Sears and S.V.N. Vishwanathan. Fenchel duality and generalized maxent. working paper, 2007.

F. Shahnaz, M.W. Berry, V.P. Pauca, and R.J. Plemmons. Document Clustering Using Nonnegative Matrix Factorization. *Information Processing & Management*, 42(2):373–386, 2006.

J. Shore and R. Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEE Tran. on Info. Theory*, IT-26, 1980.

Michael Spivak. *Calculus on Manifolds*. Westview Press, 1965.

N. Srebro and T. Jaakkola. Weighted low-rank approximations. *Proceedings of the Twentieth International Conference on Machine Learning*, pages 720–727, 2003.

Zvezdelina Stankova-Frenkel. Mobius inversion formula. muliplicative functions, 1999.

Michael J. Stutzer. Simple entropic derivation of a generalized black-scholes option pricing model. *Entropy*, 2:70–77, 2000.

H. Suyari, M. Tsukada, and Y. Uesaka. Mathematical structures derived from the q-product uniquely determined by tsallis entropy. In *ISIT 2005*, 2005.

I. Takeuchi, Q.V. Lee, T. Sears, and A. Smola. Nonparamentric quantile estimation. *Journal of Machine Learning Research*, 7:1231–1264, July 2006.

C.H. Teo, Q. Le, A.J. Smola, and S.V.N. Vishwanathan. A scalable modular convex solver for regularized risk minimization. In *Conference on Knowledge Discovery and Data Mining*, 2007.

M.E. Tipping. Bayesian Inference: An Introduction to Principles and Practice in Machine Learning. *Lecture Notes in Artificial Intelligence*, 3176:41–62, 2004.

J. Tropp. Algorithms for simultaneous sparse approximation. Part II: Convex relaxation. *Signal Processing*, 86(3):589–602, 2006.

J.A. Tropp, A.C. Gilbert, and M.J. Strauss. Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit star, open. *Signal Processing*, 86(3): 572–588, 2006.

Yaakov Tsaig and David L. Donoho. Extensions of compressed sensing. *Signal Processing*, 86(3):549–571, 2006.

C. Tsallis. Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, 52(1):479–487, 1988.

S. Umarov, C. Tsallis, and S. Steinberg. A generalization of the central limit theorem consistent with nonextensive statistical mechanics. *Arxiv preprint cond-mat/0603593*, 2006.

V.N. Vapnik. *Statistical learning theory*. Wiley New York, 1998.

C. Vignat and A. Plastino. Poincare's observation and the origin of tsallis generalized canonical distributions. *Physica A*, 365(1), 2006. URL http://arxiv.org/pdf/cond-mat/0509689.

M. J. Wainwright and M. I. Jordan. Semidefinite relaxations for approximate inference on graphs with cycles. Technical Report UCB/CSD-3-1226, University of California, Berkeley, 2003a.

M.J. Wainwright and M.I. Jordan. Graphical models, exponential families, and variational inference. *UC Berkeley, Dept. of Statistics, Technical Report*, 649, 2003b.

Marnfred Warmuth and S.V.N Vishwanathan. Divergence based inference. Unpublished notes, 2006.

Herbert S. Wilf. *generatingfunctionology*. Academic Press, 1994.

J. Williamson. *Bayesian Nets and Causality: Philosophical and Computational Foundations*. Oxford University Press, 2005.

Jun Zhang. Divergence function, duality and convex analysis. *Neural Computation*, 16(159-195), 2004.