A Comparison of Entropies in the Underdetermined Moment Problem

Jonathan M. Borwein

Department of Mathematics and Statistics
Simon Fraser University
Burnaby, British Columbia V5A 1S6, CANADA

Robert K. Goodrich
Mark A. Limber

Department of Mathematics
University of Colorado at Boulder
Boulder, CO 80309, USA

Abstract

The problem of function reconstruction from a small number of accurate moments is reviewed and studied numerically. Since this problem is underdetermined, a number of selections have been proposed to pick out a single function. We consider those selections of the "entropic" variety, i.e. those of the form

$$f(x) = \int \phi(x(t))\mathrm{d}t$$

for some convex function ϕ. We present a summary of our attempts to determine which "entropy" is best for reconstructing x in a variety of model problems. This presumes we have a measure of "best", which in itself is a difficult issue.

We include a review of the pertinent theory and the details of our numerical implementation. We give numerous numerical examples to support our conclusions.

1. Introduction

The problem of reconstructing a function from a finite set of data is very common in applied problems. Usually the finite data is not sufficient to uniquely determine the underlying function and some sort of "best" function must be chosen from the collection

[1]Key words: Convex analysis, maximum entropy reconstructions, numerical analysis.

of all functions that have the correct data. By best, we mean that the reconstruction should be as "close as possible" to the original function that generated the data. We study a particular instance of such a situation: the problem of reconstruction from a finite collection of moments.

The set up is as follows. Suppose we are given data $b_1, \ldots, b_n \in \mathbb{R}^n$ generated from an unknown function $x^* \in L^1(\Omega)$, via

$$b_k = \int_\Omega x^*(t)\psi_k(t)\mathrm{d}t. \tag{1.1}$$

where $\Omega \subset \mathbb{R}^d$ is a compact set and the functions $\psi_k \in L^\infty(\Omega)$ are known. Let b be the vector of data. We wish to find the function $\hat{x}$ that has the same data and minimizes some convex functional f. In this way we are choosing the "best" feasible function out of the potential infinity of feasible functions.

For notational simplicity, we shall write $A : L^1(\Omega) \to \mathbb{R}^n$ for the operator that maps a function $x(t)$ into its data vector:

$$(Ax)_k = \int_\Omega x(t)\psi_k(t)\mathrm{d}t. \tag{1.2}$$

We say a function x is feasible if $Ax = b$.

Our basic problem is then, find the function $\hat{x}(t)$ with $A\hat{x} = b$ that minimizes the convex functional f. That is, find the feasible $\hat{x}$ such that

$$\hat{x} = \operatorname{argmin}\left\{f(x) : Ax = b\right\}. \tag{1.3}$$

Here we use the notation argmin to denote the feasible function where the minimum is attained.

In many applications a positivity constraint is natural and the problem becomes, find

$$\hat{x} = \operatorname{argmin}\left\{f(x) : Ax = b, x \geq 0\right\}. \tag{1.4}$$

Now we say a function x is feasible if $Ax = b$ and $x \geq 0$ a.e. We note in passing that it is possible to impose further lattice constraints $\alpha \leq x \leq \beta$ as in [7, 9].

Such problems can be found in power spectrum estimation[16], image reconstruction from projections[14, 19, 25], crystallography[11, 23], information theory[17, 27], fluid flow problems[29], speech processing[18] and many other applied fields[28]. In different fields different objective functions f are used with a variety of rationale ranging from statistical arguments to simply the observation that the reconstruction "looks good". This paper attempts to quantitatively judge several objective functions in one dimensional reconstruction problems by measuring certain properties of the reconstructed functions.

Despite the size of the literature dealing with these sorts of problems, it can be argued that the theoretical foundations of many of the solution derivations are suspect, and in many cases incorrect. Section 2 of this paper presents a heuristic derivation of the Fenchel duality theorem and cites references where a mathematically sound foundation for our solution technique can be found.

In section 3 we discuss several popular objective functionals, or *entropies*, derive the forms of their solutions and discuss some of their immediately apparent properties.

In section 4 we discuss the numerical aspects of the minimization problem. We show that it is better to discretize the integrals in the continuous dual problem rather than in the primal problem, and describe our numerical implementation.

Section 5 defines several measures of what we think makes one reconstructed function in some way better than another and we give some numerical evidence for choosing one objective over another.

It should be emphasized that no single entropy is going to be the best in every situation. For a particular application, one should consider the kind of solution one expects to get, and pick the entropy that best produces that type of function. This comment will become more clear in later sections.

For a comparison of the Burg and Boltzmann-Shannon entropies in speech processing, see [18]. For a brief comparison of several entropies, see [1]. For a heuristic method to solve the Boltzmann-Shannon maximum entropy, see [3].

2. Theoretical Results

For reasons that will become clear, characterizing solutions to our minimization problem is not as straight forward as one might suspect. The facts that we are working over the infinite dimensional space $L^1(\Omega)$ and that our objective function typically is nondifferentiable make a correct analysis difficult. We present a heuristic derivation of our solution technique and quote a sound treatment of the problem.

For brevity we refer the reader to the cited literature for proofs of most of the assertions. We point out along the way the problem steps that must be addressed in a complete treatment of the problem. A mathematically complete treatment can be found in [4].

2.1 The Problem Statement

We consider the following reconstruction problem. Let

$$f(x) = \int_\Omega \phi(x(t))\mathrm{d}t \tag{2.1}$$

for some strictly convex function $\phi : I\!R \to (-\infty, +\infty]$. Recall a function ϕ is *strictly convex* if

$$\phi(\lambda u + (1-\lambda)v) < \lambda\phi(u) + (1-\lambda)\phi(v) \tag{2.2}$$

for all $u \neq v$ in $I\!R$ and all $\lambda \in (0,1)$. We wish to characterize solutions to

$$\inf\left\{\int_\Omega \phi(x(t))\mathrm{d}t : Ax = b, \quad x \geq 0\right\}. \tag{2.3}$$

By defining

$$\phi_+(t) = \begin{cases} \phi(t) & t \geq 0 \\ +\infty & \text{else.} \end{cases} \tag{2.4}$$

we can write (2.3) as

$$\inf\left\{\int_\Omega \phi_+(x(t))\mathrm{d}t : Ax = b\right\}. \tag{2.5}$$

We call this the *primal* problem. Note that in many cases the function $x \mapsto \int \phi_+(x)\mathrm{d}t$ is not Frechet or even Gateaux differentiable. This fact is commonly forgotten in many derivations of the solutions. For example, if

$$f(x) = \int_\Omega x(t)\ln[x(t)]\mathrm{d}t \tag{2.6}$$

then f is $+\infty$ whenever $x < 0$ on a set of positive measure, and since in $L^1(\Omega)$ the set of all functions that are negative on a set of positive measure is dense, the function f is not continuous, much less differentiable. This is the motivation for studying a dual formulation of the problem.

2.2 Dual Problems

We first give a heuristic approach to the dual problem to motivate what follows. This is not meant to be complete, only to reassure those unfamiliar with Fenchel duality.

2.2.1 Heuristic Derivation of Fenchel Duality

First consider the general problem

$$p = \inf_x \{f(x) : g(x) = 0\} \tag{2.7}$$

where we leave the spaces in which x, f and g live undetermined for this discussion. Suppose the minimum is attained at the feasible function $\hat{x}$. A common approach to solving such a problem is to attach a Lagrange multiplier vector $\hat{\lambda}$ to the equality constraint. Then assuming such a vector exists we get

$$p = \inf_x \left\{f(x) + \langle g(x), \hat{\lambda}\rangle\right\}. \tag{2.8}$$

If λ is any vector, we have

$$\begin{aligned} \inf_x \{f(x) + \langle g(x), \lambda\rangle\} &\leq \inf \{f(x) + \langle g(x), \lambda\rangle : g(x) = 0\} && (2.9)\\ &= \inf \{f(x) : g(x) = 0\} && (2.10)\\ &= f(\hat{x}) && (2.11) \end{aligned}$$

and since we have assumed a Lagrange multiplier vector exists we see that

$$\inf_x \{f(x) : g(x) = 0\} = \max_\lambda \inf_x \{f(x) + \langle g(x), \lambda\rangle\}. \tag{2.12}$$

This result is known as Lagrange duality [22]. One common mistake in the literature is to write the Lagrangian

$$L(x, \lambda) = f(x) + \langle g(x), \lambda\rangle \tag{2.13}$$

and argue that the optimal $\hat{x}$ must occur where the derivative of L with respect to x is zero, that is,

$$\nabla_x L(\hat{x}, \hat{\lambda}) = 0. \tag{2.14}$$

Unfortunately, as we mentioned above, the function f need not be differentiable, and in that case neither is L.

If we now consider the problem

$$\inf_x \{f(x) : Ax = b\} \tag{2.15}$$

where A is a linear map into $\mathbb{R}^n$, then this fits the form of (2.7) with $g(x) = b - Ax$. We rewrite (2.12) with this g.

$$\begin{aligned} \inf_x \{f(x) : Ax = b\} &= \max_\lambda \inf_x \{f(x) + \langle b - Ax, \lambda\rangle\} && (2.16) \\ &= \max_\lambda \inf_x \left\{f(x) + \langle b, \lambda\rangle - \langle x, A^T\lambda\rangle\right\} && (2.17) \\ &= \max_\lambda \left\{\langle b, \lambda\rangle + \inf_x \left\{f(x) - \langle x, A^T\lambda\rangle\right\}\right\}. && (2.18) \end{aligned}$$

If we know define

$$f^*(y) = \sup_x \{\langle x, y\rangle - f(x)\} \tag{2.19}$$

to be the Fenchel conjugate, or Legendre transform, of f then we see that

$$\inf \{f(x) : Ax = b\} = \max_\lambda \left\{\langle b, \lambda\rangle - f^*(A^T\lambda)\right\} \tag{2.20}$$

and the maximum on the right hand side is attained. The right hand side of (2.20) is what is known as the Fenchel dual of problem (2.15):

$$\max_\lambda \left\{\langle b, \lambda\rangle - f^*(A^T\lambda)\right\} \tag{2.21}$$

Note we have that the *values* of the primal and dual are equal. We shall see that we can use the dual to directly determine the solution $\hat{x}$ to the primal problem, and can disregard the heuristic derivation of the dual (2.21). To do that we need a little convex analysis.

For a convex function f, the *subgradients* of f at the point $\hat{x}$ is defined to be the set

$$\partial f(\hat{x}) = \{y : \langle x - \hat{x}, y\rangle \leq f(x) - f(\hat{x}) \quad \forall x\}. \tag{2.22}$$

This is a convenient generalization of the derivative of a convex function. It can be shown[26] that for a lower semicontinuous convex function f,

$$x \in \partial f^*(y) \iff y \in \partial f(x). \tag{2.23}$$

It can also be shown, that for many convex functionals f, f^* is Frechét differentiable, even though f may not be.

These facts are useful in solving the dual problem (2.20). Since the function

$$G(\lambda) = \langle b, \lambda\rangle - f^*(A^T\lambda) \tag{2.24}$$

is concave, there is at most one maximum value (though this value may be attained at several λ's) and since G is differentiable the maximum must occur where

$$\nabla_\lambda G(\hat{\lambda}) = 0 = b - A\nabla f^*(A^T\hat{\lambda}) \tag{2.25}$$

assuming, of course, that $\nabla f^*(A^T\hat{\lambda})$ lies in the correct space, a condition that must be checked. Let $\hat{x} = \nabla f^*(A^T\hat{\lambda})$ where $\hat{\lambda}$ solves the Fenchel dual problem. Then by (2.25) we see that $A\hat{x} = b$ and $\hat{x}$ is feasible. Also for any x with $Ax = b$ we have

$$\begin{aligned}
\hat{x} = \nabla f^*(A^T\hat{\lambda}) \quad &\Longleftrightarrow \quad A^T\hat{\lambda} \in \partial f(\hat{x}) && (2.26)\\
&\Rightarrow \quad \langle A^T\hat{\lambda}, x - \hat{x}\rangle \le f(x) - f(\hat{x}) && (2.27)\\
&\Rightarrow \quad \langle\hat{\lambda}, Ax\rangle - \langle\hat{\lambda}, A\hat{x}\rangle \le f(x) - f(\hat{x}) && (2.28)\\
&\Rightarrow \quad 0 \le f(x) - f(\hat{x}) && (2.29)
\end{aligned}$$

since $Ax = b = A\hat{x}$. This implies that $\hat{x}$ is the solution to the primal problem.

2.2.2 Fenchel Duality

Here we quote a theorem that makes the above conclusion rigorous, though not the derivation. This theorem gives conditions under which the optimal solution exists and characterizes the solution in terms of a solution to a finite dimensional smooth problem, as in the heuristic development.

It should first be mentioned that the mere question of existence of a solution to (2.5) involves a delicate argument because the problem is formulated in $L^1(\Omega)$. Certain conditions must be imposed on the integrand function ϕ: ϕ must be a proper lower semicontinuous, strictly convex function with (necessarily essentially smooth) conjugate ϕ^*. This last condition ensures that the gradient of ϕ^* tends towards infinity as x tends towards the boundary of the domain of ϕ^*, the set of points where ϕ^* is finite valued. This guarantees that the conjugate f^* of f is differentiable. To prove existence we must impose a growth condition on ϕ. These issues are discussed in [4].

From [4] we quote a Fenchel duality theorem which circumvents the above derivation.

Theorem 2.1 *Let $\Omega = [\alpha, \beta]$, with Lebesgue measure, and each ψ_i be continuously differentiable, $i = 1, \ldots, n$. Suppose that ϕ_+ is lower semicontinuous and strictly convex on its domain. Suppose*

$$\exists \mu \in I\!R^n \text{ such that } \sum_{i=1}^{n} \mu_i \psi_i(t) < d \quad \forall t \in [\alpha, \beta], \tag{2.30}$$

where $d = \lim_{u\to\infty}(\phi(u)/u)$, and if $d < \infty$ define $c = \lim_{y\uparrow d}(d-y)((\phi_+)^*)'(y)$. Suppose that either $d = \infty$ or $d < \infty$ and $c > 0$. Finally, suppose the following constraint qualification holds:

$$\exists x_0 \in \mathrm{dom}(f) \text{ such that } Ax_0 = b \text{ and } x_0 > 0 \text{ a.e.} \tag{2.31}$$

Then the unique primal optimal solution is given by

$$\hat{x}(t) = (\phi_+^*)' \left(\sum_{i=1}^{n} \hat{\lambda}_i \psi_i(t) \right) \tag{2.32}$$

where $\overline{\lambda}$ *is any solution to*

$$\max_{\lambda \in I\!R^n} \left\{ \langle b, \lambda \rangle + \int_\Omega \phi_+^* \left(\sum_{i=1}^{n} \hat{\lambda}_i \psi_i(t) \right) \right\} \tag{2.33}$$

Note that $(A^T\lambda)(t) = \sum_{i=1}^n \lambda_i \psi_i(t)$ so (2.33) is the Fenchel dual (2.20) as in the discussion. This gives us a prescription for finding solutions to (2.5): solve the finite dimensional dual problem (2.33) for a vector $\hat{\lambda}$ and use (2.32) for the optimal function.

It should also be mentioned that in the case the functions ψ_k are linearly independent on every set of positive measure, called pseudo Haar[4], and if $(0, \infty) \subseteq \mathrm{dom}\phi$ then the *constraint qualification* (2.31) can be weakened to

$$\exists x_0 \in \mathrm{dom}(f) \text{ such that } Ax_0 = b.$$

In general (CQ) is *not* the same as assuming there is an x_0 in the interior of the positive cone in $L^1(\Omega)$. The fact that the interior of the positive cone in $L^1(\Omega)$ is empty is one of the main difficulties in giving a mathematically sound derivation of the Fenchel duality result. From this fact, f typically is not continuous and standard Lagrange multiplier theorems do not apply. Efforts have been made to get around this problem in [15] and [20], but the Fenchel approach seems the most useful.

This constraint qualification (2.31) is essentially that there is a feasible function x_0 in the *quasi interior* of the positive cone, a generalization of the interior which has some of the same properties of the interior, but is nonempty in cases where the interior may be empty. For a more extensive discussion of the quasi interior, see [5] and [20].

It can also be shown that if the ψ's are linearly independent then there is a unique solution $\hat{\lambda}$ to the dual problem (2.33). Also, if

$$x_0(t) = (\phi_+^*)(A^T\lambda) \tag{2.34}$$

is feasible for some λ discovered by whatever method, then x_0 is in fact primal optimal. That is, any feasible function of the correct form is optimal.

3. The Objective Functions

In this section we describe several popular entropies and their solution forms as derived from the results in the previous section. We also discuss some of the difficulties in working with some of these entropies.

3.1 Entropies

There are three entropies that are most commonly found in the literature. These are the L^2-norm, the Boltzmann-Shannon and Burg entropies. We discuss these entropies and a simple generalization which gives a natural directed distance version of these. We also describe the Fermi-Dirac entropy which enforces lattice bounds on feasible functions. We give formulas for ϕ_+, ϕ_+^* and the numbers d and c from Theorem 2.1 as appropriate.

L^2 norm: This is just the L^2 norm of the function x, used in power spectrum estimation [16].

$$\phi_+(u) = \begin{cases} \frac{1}{2}u^2 & u \geq 0 \\ +\infty & \text{else,} \end{cases} \tag{3.1}$$

$$\phi_+^*(y) = \frac{1}{2}(y_+)^2, \quad d = +\infty. \tag{3.2}$$

where y_+ is the function y truncated at zero.

Boltzmann-Shannon: Made popular by Shannon in information theory [27] and Boltzmann in thermodynamics.

$$\phi_+(u) = \begin{cases} u\ln(u) & u > 0 \\ 0 & u = 0 \\ +\infty & \text{else,} \end{cases} \tag{3.3}$$

$$\phi_+^*(y) = \exp(y-1), \quad d = +\infty. \tag{3.4}$$

Burg: Popularized in geophysics by Burg[8] and used in speech processing[18].

$$\phi_+(u) = \begin{cases} -\ln(u) & u > 0 \\ +\infty & \text{else,} \end{cases} \tag{3.5}$$

$$\phi_+^*(y) = -1 - \ln(-y), \quad d = 0, \quad c = 1. \tag{3.6}$$

Fermi-Dirac: This entropy has the added feature that it forces feasible functions to lie between 0 and 1. This has been found useful in fluid flow problems[29].

$$\phi_+(u) = \begin{cases} u\ln(u) + (1-u)\ln(1-u) & 0 < u < 1 \\ 0 & u = 0 \text{ or } u = 1 \\ +\infty & \text{else,} \end{cases} \tag{3.7}$$

$$\phi_+^*(y) = \ln(1+\exp(y)), \quad d = +\infty. \tag{3.8}$$

Directed Distances: Since for a convex function ϕ we have

$$(u-v)\phi'(v) \leq \phi(u) - \phi(v) \tag{3.9}$$

it is popular to define a directed distance function

$$d(u,v) = \phi(u) - \phi(v) + (v-u)\phi'(v) \geq 0 \tag{3.10}$$

measuring a distance between the points u and v. This can be used to construct our objective function f via

$$f_y(x) \equiv D(x,y) \equiv \int_\Omega d(x(t), y(t))\mathrm{d}t \tag{3.11}$$

For example, if $\phi(u)$ is the Boltzmann-Shannon entropy, then (3.10) becomes

$$d(u,v) = u\ln(u) - v\ln(v) + (v-u)(1+\ln(v)) = u\ln(u) - u\ln(v) + v - u \tag{3.12}$$

and minimizing $D(x(t), 1)$ is equivalent to minimizing the Boltzmann-Shannon entropy if one of the ψ_j is a constant function. If we know $y(t)$ and some ψ_j is constant then we can eliminate the linear piece in (3.12) and we get

$$f_y(x) = \int_\Omega x(t)\ln[x(t)/y(t)]\mathrm{d}t. \tag{3.13}$$

This is known as the *cross entropy* and is used to incorporate prior knowledge (the function y) into the reconstruction process.

3.2 Solution Forms

We shall assume throughout that the data generating functions $\psi_k(t) = t^{k-1}$, that is, we are solving the Hausdorff moment problem. For the more general case, simply substitute ψ_k where ever t^{k-1} appears.

These entropies are different in their behavior and in the form of solutions. For example, the L^2 norm gives as a solution

$$\hat{x}(t) = \left(\sum_{k=1}^{n} \lambda_k t^{k-1}\right)_+, \tag{3.14}$$

a truncated polynomial. In many cases the generated solution will be zero on a large set of positive measure, and if the goal is to reconstruct a function that can be zero, then this might be a positive attribute. However, this function is not smooth in general.

The Boltzmann-Shannon entropy is smooth, analytic in our case, but since the form of the solution is

$$\hat{x}(t) = \exp\left(\sum_{k=1}^{n} \lambda_k t^{k-1}\right) \tag{3.15}$$

the solution will never be exactly zero. We shall see that this entropy does a good job of approximating zero valued functions though.

Burg's entropy gives a solution of the form

$$\hat{x}(t) = -\left(\sum_{k=1}^{n} \lambda_k t^{k-1}\right)^{-1} \tag{3.16}$$

when the solution exists. The problem with this entropy for more general A is that ϕ^* is not everywhere finite and solutions are not guaranteed to exist in L^1, even if $(\phi_+^*)'(A^T\hat{\lambda}) \in L^1$. In our setting solutions are ensured by Theorem 2.1. Also, this function can be much more difficult to compute with numerically in cases where it is trying to reconstruct a function that is zero on a large set.

The Fermi–Dirac entropy has solutions of the form

$$\hat{x}(t) = \frac{\exp\left(\sum_{k=1}^{n} \lambda_k t^{k-1}\right)}{1 + \exp\left(\sum_{k=1}^{n} \lambda_k t^{k-1}\right)} \tag{3.17}$$

and consequently the range of solutions lie in $(0, 1)$. This entropy can be modified to give a smooth solution whose range is between two given functions α and β. While the other entropies can be forced to make solutions lie between such functions by adding the constraint $\alpha \leq x \leq \beta$, this produces a smooth function. For a more detailed discussion see [7] and [9].

The directed distances approach is a general method to create new objective functions incorporating prior knowledge. In image reconstruction this can be used to force a picture to be as close to an idealized picture as possible while still retaining the correct data, the authors of [14] discuss this idea in more detail. For the cross entropy (3.13) the conjugate is

$$f_y^*(z) = \int_\Omega y(t)\exp(z(t) - 1)\mathrm{d}t \tag{3.18}$$

and so the form of the solution is

$$\hat{x}(t) = y(t)\exp\left(\sum_{k=1}^{n} \lambda_k t^{k-1} - 1\right). \tag{3.19}$$

So the solution to the cross entropy problem is functionally proportional to the Boltzmann-Shannon solution.

4. Numerical Issues

In this section we discuss our implementation of the theory previously presented, with some simplifications that appear. In the first subsection we compare discretizing the primal problem and using the associated dual, with simply discretizing the continuous dual problem.

4.1 Discretization

Recall we are essentially trying to find solutions to

$$\inf_x \left\{ \int_\Omega \phi_+(x(t))\mathrm{d}t : \int_\Omega x(t)\psi_k(t)\mathrm{d}t = b_k, \quad k = 1, \ldots, n \right\} \tag{4.1}$$

by solving the dual problem

$$\max_\lambda \left\{ \langle b, \lambda \rangle - \int_\Omega \phi_+^* \left(\sum_{i=1}^n \lambda_i \psi_i(t) \right) \right\} \tag{4.2}$$

where from now on we assume $\Omega \subset I\!\!R$ for simplicity. One might think that we should discretize the integrals in (4.1), the primal problem, and thus reduce the problem to one in finite dimensions. We shall show that with a fixed integration scheme it makes no difference whether we discretize (4.1) and solve the related discrete dual problem, or discretize the continuous dual (4.2).

Theorem 4.2 *Let $\{w_i\}_{i=1}^M$ and $\{t_i\}_{i=1}^M$ define an integration scheme via*

$$\int_\Omega g(t)\mathrm{d}t \approx \sum_{i=1}^M g(t_i) w_i. \tag{4.3}$$

Then the (formal) dual of the discretization of (4.1) is the same as the discretization of (4.2).

Proof: If we discretize the primal problem we get

$$\inf_x \left\{ \sum_{i=1}^M \phi_+(x(t_i)) w_i : \sum_{i=1}^M x(t_i)\psi_j(t_i) w_i = b_j, \quad j = 1, \ldots, n \right\}. \tag{4.4}$$

Let $\tilde{x} \in I\!R^M$ denote the vector $\tilde{x}_i = x(t_i)$, let

$$\tilde{f}(\tilde{x}) = \begin{cases} \sum_{i=1}^{M} \phi(\tilde{x}_i) w_i & \tilde{x} \geq 0 \\ +\infty & \text{else.} \end{cases} \tag{4.5}$$

The associated dual problem is

$$\max_{\lambda \in I\!R^n} \left\{ \langle b, \lambda \rangle - \tilde{f}^*(\tilde{A}^T \lambda) \right\} \tag{4.6}$$

where

$$\tilde{A} = \begin{bmatrix} \psi_{11} w_1 & \psi_{12} w_2 & \cdots & \psi_{1M} w_M \\ \vdots & \vdots & \ddots & \vdots \\ \psi_{n1} w_1 & \psi_{n2} w_2 & \cdots & \psi_{nM} w_M \end{bmatrix} \tag{4.7}$$

with $\psi_{ij} = \psi_i(t_j)$. Then for $\lambda \in I\!R^n$ we have

$$\tilde{A}^T \lambda = \begin{pmatrix} w_1 \sum_{i=1}^{M} \psi_{i1} \lambda_i \\ \vdots \\ w_M \sum_{i=1}^{M} \psi_{iM} \lambda_i \end{pmatrix}. \tag{4.8}$$

It is easy to check that for $g, h : I\!R \to (-\infty, +\infty]$ and $w \in I\!R$ if $g(x) = wh(x)$ then $g^*(y) = wh^*(y/w)$. Using this we can calculate $\tilde{f}^*$. Let $y \in I\!R^M$. Then by definition

$$\begin{aligned} \tilde{f}^*(y) &= \sup_{z \in I\!R^M} \left\{ \langle y, z \rangle - \tilde{f}(z) \right\} && (4.9) \\ &= \sup_{z \in I\!R^M_+} \left\{ \langle y, z \rangle - \sum_{i=1}^{M} \phi(z_i) w_i \right\} && (4.10) \\ &= \sup_{z \in I\!R^M_+} \left\{ \sum_{i=1}^{M} (y_i z_i - \phi(z_i) w_i) \right\} && (4.11) \\ &= \sum_{i=1}^{M} \sup_{z \in I\!R^M_+} \{ y_i z_i - \phi(z_i) w_i \} && (4.12) \\ &= \sum_{i=1}^{M} w_i \phi_+^*(y_i / w_i) && (4.13) \end{aligned}$$

and so the discrete dual becomes

$$\max_{\lambda \in I\!R^n} \left\{ \langle b, \lambda \rangle - \sum_{j=1}^{M} \phi_+^* \left(\sum_{i=1}^{n} \psi_{ij} w_j \lambda_i / w_j \right) w_j \right\} \tag{4.14}$$

$$= \max_{\lambda \in I\!R^n} \left\{ \langle b, \lambda \rangle - \sum_{j=1}^{M} \phi_+^* \left(\sum_{i=1}^{n} \psi_{ij} \lambda_i \right) w_j \right\}. \tag{4.15}$$

Now consider a direct discretization of the continuous dual problem (4.2):

$$\max_{\lambda} \left\{ \langle b, \lambda \rangle - \sum_{j=1}^{M} \phi_+^* (A^T \lambda(t_j)) \psi_j(t_j) w_j \right\} \tag{4.16}$$

where $(A^T\lambda)(t) = \sum_{i=1}^{n} \lambda_i \psi_i(t)$ and so

$$(A^T\lambda)(t_j) = \sum_{i=1}^{n} \lambda_i \psi_i(t_j) = \sum_{i=1}^{n} \lambda_i \psi_{ij}. \tag{4.17}$$

and we see the two dual problems are equivalent. $\square$

We believe it is usually best to follow a general philosophy of not discretizing until absolutely necessary, and this theorem justifies doing just that. There are definitely several reasons one wouldn't want to discretize the primal first. For example, if we are using an adaptive numerical integration scheme, it makes no sense to set the nodes based on the integrals in the primal, when in fact it is the dual problem that gets solved. Also, once an integration scheme is imposed by a discretization of the primal, the scheme is fixed in the dual. In our implementation we use a fixed node Gaussian quadrature rule, and by discretizing the dual problem we can precompute the values ψ_{ij} once and for all in our iterative scheme to maximize the dual.

It is an interesting comment to note that, in some applications, like image reconstruction from projections, characterizing the solutions to the infinite dimensional primal problem in L^1 leads to a natural optimal discretization for the problem [21].

Another issue that arises is whether as we improve the integration scheme by using more nodes, generating the vectors λ^k, we actually converge to the vector $\hat{\lambda}$ that solves the continuous dual problem. We use the notions of epi-convergence of a sequence of functions from [30], which in our case is implied by pointwise convergence (see Corollary 4 in [30]).

Theorem 4.3 *Let I_k be an integration scheme defined by*

$$I_k(g) = \sum_{j=1}^{M_k} g(t_j^k) w_j^k \tag{4.18}$$

and assume $I_k(g) \to \int_\Omega g$ for all continuous functions g. Assume the ψ_i's are continuous and linearly independent, ϕ_+^* is continuous on $\mathbb{R}$ and let

$$\lambda^k = \operatorname{argmax}\left\{\langle b, \lambda\rangle - I_k(\phi_+^*(\tilde{A}_k^T \lambda)\right\} \tag{4.19}$$

where $\tilde{A}_k$ is defined as $\tilde{A}$ was. Then $\lambda^k \to \hat{\lambda}$ where $\hat{\lambda}$ solves the continuous dual problem.

Proof: The conditions given assure us that the solutions λ^k and $\hat{\lambda}$ are unique (see [4]). If we let

$$g_k(\lambda) = I_k(\phi_+^*(\tilde{A}_k^T \lambda)) - \langle b, \lambda\rangle \tag{4.20}$$

$$g(\lambda) = \int_\Omega \phi_+^*(A^T \lambda))\mathrm{d}t - \langle b, \lambda\rangle \tag{4.21}$$

so that $g_k : \mathbb{R}^n \to \mathbb{R}$ is convex and by assumption $g_k(\lambda) \to g(\lambda)$ for each $\lambda \in \mathbb{R}^n$. By [30] Corollary 7, this implies $\lambda^k \to \hat{\lambda}$. □

4.2 Extendibility

In order for the duality theory to hold we require that b lie in the relative interior of the extendible vectors, E, where

$$E = \left\{c \in \mathbb{R}^n : c = \int_\Omega \Psi(t) d\nu(t), \quad \text{some } \nu \geq 0, \quad \nu \in M(\Omega)\right\}, \tag{4.22}$$

$M(\Omega)$ is the dual of $C(\Omega)$, and $\Psi(t) = [\psi_1(t), \ldots, \psi_n(t)]^T$. The equivalence of our constraint qualification, (2.31), and this condition is given in [5]. In this discussion, we assume that the functions ψ_i are continuous.

Let

$$P = \{\nu \in M(\Omega) : \nu \geq 0, \quad \nu(\Omega) = 1\}. \tag{4.23}$$

Since P is a weak-$*$ compact convex set, it is generated by the extreme points ν_t, the point mass measures on Ω, by the Krein Milman Theorem[10]. Let $T : M(\Omega) \to \mathbb{R}^n$ by

$$T(\nu) = \int_\Omega \Psi(t) d\nu(t). \tag{4.24}$$

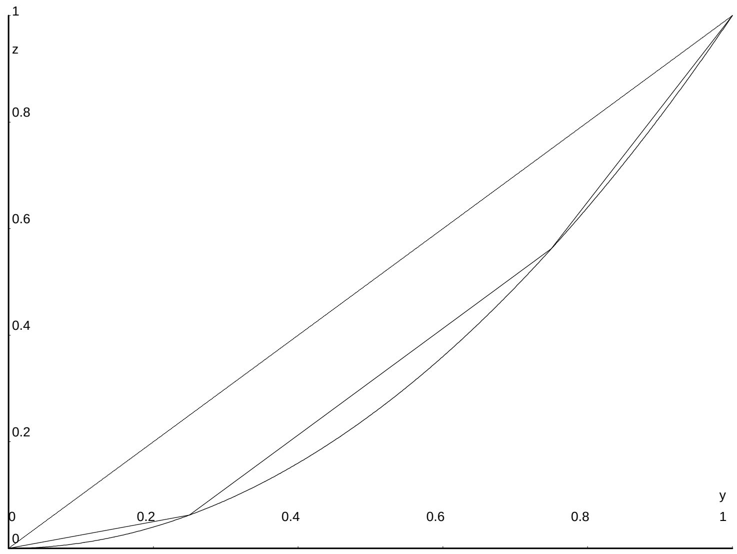


Figure 1: The extendible vectors

Then $T(P) = S$ where S is

$$S = \left\{ c \in I\!R^n : c = \int_\Omega \Psi(t) d\nu(t), \quad \text{some } \nu \geq 0, \quad \nu(\Omega) = 1 \right\}, \tag{4.25}$$

and S is a compact convex set that generates the cone E. Conway [10] contains a discussion of these results. Thus, S is generated by the set $\Psi(t)$ for $t \in \Omega$.

Assuming $1 \equiv \psi_1(t)$, b is in the interior of E if and only if b/b_1, the vector obtained by dividing b componentwise by b_1, which is nonzero by the constraint qualification, is in the relative interior of S.

When we employ a discrete integration technique, the above analysis still holds, provided b/b_1 is in the relative interior of the convex hull, E_n, of $\Psi(t_i)$, $i = 1, \ldots, M$ where the t_i's are the nodes of our integration scheme. Simple examples show that although b/b_1 is in $\mathrm{int}(S)$ it may not be in the interior of E_n.

For example, let $\psi_i = t^{i-1}$, $i = 1, 2, 3$ and $\Omega = [0, 1]$. In Figure 1, we are looking down the x-axis at the curve $[1, t, t^2]$. The region bounded by this curve and $z = y$, $x = 1$ is S. The region bounded by $z = y$, $x = 1$ and the other three polygonal lines is E_n, where the nodes are $t = 0, \frac{1}{4}, \frac{3}{4}, 1$. The point $(1, \frac{1}{2}, \frac{1}{4} + \epsilon)$ for small ϵ, is an example of a point in the interior of S but not even in E_n. It is clear from the figure, that a fairly dense set of nodes generates most of the convex set S

In fact, for an increasing sequence of convex sets $C_n \subset I\!R^n$, we have [2]

$$\bigcup (\mathrm{int} C_n) = \mathrm{int}\left(\bigcup C_n\right). \tag{4.26}$$

Suppose E_n is the nth discretization of E and we have

$$\bigcup E_n = E, \qquad E_{n-1} \subset E_n. \tag{4.27}$$

Then from (4.26) we see that $b \in \mathrm{int}(E)$ if and only if b is eventually in $\mathrm{int}(E_n)$. So, we can say that the discrete CQ is eventually satisfied if it is satisfied in the continuous problem.

There exists algorithms for determining if b/b_1 is in E_n, see [24]. However, these algorithms are computationally intensive, and in practice we simply check for convergence in our algorithm.

4.3 Implementation Details

To solve (4.1) we form the dual (4.2), discretize the integral using Gaussian quadrature with a fixed number of nodes and solve the resulting problem with Newton's method [12]. The reason we choose a fixed rather than an adaptive integration scheme is two fold. First, we don't want the integration to slow down the routine, which is likely with piecewise continuous data in an adaptive scheme, and second we can then precompute the ψ_{ij}'s. We use $\Omega = [0, 1]$, though we could use any interval in the real line.

For some of the popular entropies, the Newton step can be simplified. Denote the Hessian matrix of the dual functional (2.24) at step k as H_k, then for the L^2 norm the Newton step is

$$\lambda_{k+1} = -H_k^{-1} b \tag{4.28}$$

This can be seen by showing that

$$\begin{aligned} H_k &= \left(\int_{\{t: A^T\lambda_k(t)>0\}} t^{i+j-2}\mathrm{d}t\right)_{ij=1}^n && (4.29) \\ \nabla G(\lambda_k) &= \left(b_j - \int_0^1 (A^T\lambda_k)_+ t^{j-1}\mathrm{d}t\right)_{j=1}^n && (4.30) \end{aligned}$$

and $-H_k\lambda_k = \nabla G(\lambda_k)$ so that Newton's method reduces as claimed.

Similar reductions can be made for the Boltzmann-Shannon, Burg and L^p norm entropies. We also implemented a guarded line search Newton's method which give global convergence in most cases, but we found this mostly unnecessary as pure Newton's method appears to work quite well in most cases.

For simplicity of implementation we used the symbolic language Maple for our experiments. One advantage of such a language is the ease with which we can experiment with different underlying functions $x^*(t)$ and we can easily change entropies since Maple will automatically compute Fenchel conjugates.

In most runs we used 55 nodes in the integration scheme and worked with 15 digits of accuracy, though with Maple either of these can be changed easily.

Our measure of convergence was the gradient of the dual functional; that is, when

$$||\nabla G(\lambda_k)||_\infty < \epsilon \tag{4.31}$$

we stop. This is equivalent to stopping when the reconstructed function is of the correct form for the solution and is within ϵ of having the correct data, see equation (2.25).

In our implementation, the user gives the entropy type, the underlying function x^*, the number of moments to work with, and the maximum number of Newton steps allowed. The program computes the vector b and solves the associated dual problem. From the solution vector λ we can construct the solution function $\hat{x}$.

For the Burg entropy we added a check to make sure the Newton step didn't push us out of the region where $A^T\lambda < 0$. This is part of the difficulty of working with the Burg entropy. For the Fermi-Dirac entropy we found that the choice of initial guess is very important. We used

$$\lambda_0 = \left(\ln\left|\frac{b_1}{1-b_1}\right|, 0, \ldots, 0\right)^T. \tag{4.32}$$

5. Numerical Comparisons

5.1 A Brief Description of the Comparison Criterion

In this section, we see how well we can reconstruct a function given a small number of its moments. We apply several tests to see how well each construction performed. The first group of tests measure the error from the known function generating the moments. That is, we introduce various measures of the error from our ideal function in each of our estimates. We measure this error in the L^2 norm, the L^∞ norm and what we call the convergence in measure metric. These each give some indication of how well each estimate reconstructed the known function x^* generating the data. Another reason for computing such estimates is to provide what is called an extension of the original moments. That is, once one knows the optimal estimate, one can estimate unknown moments by computing

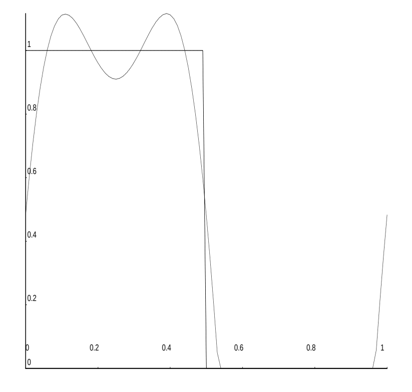
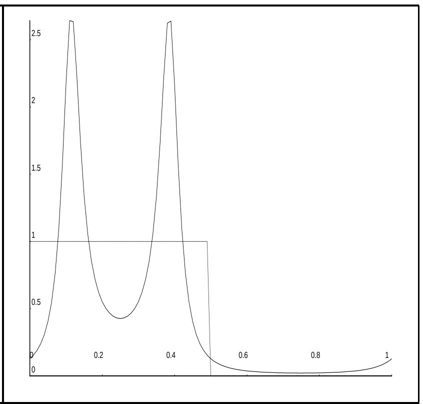
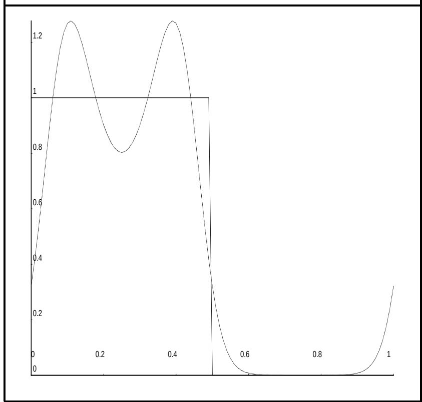
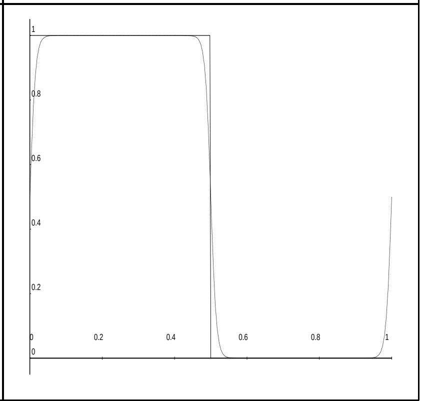
	
L^2 Reconstruction	Burg Reconstruction
	
B-S Reconstruction	F-D Reconstruction

Figure 2: $\chi_{[0,1/2]}$ with 5 trig moments

moments of our estimate. We see how well each of our estimates extend our original moments by testing to see how close these new moments are to the actual corresponding moments of our known function. The next three tests are what we call the homefield advantage tests. In these tests we know our optimal estimate minimizes some functional. We test to see how well our other estimates do in minimizing this same functional. The last test of this sort we impose is a simple subjective visual test. That is, which estimate looks the best.

5.2 Ideal Band Pass Function

	L^2	Bg	B-S	F-D
iters	8	9	9	8
stop	5.9×10^{-4}	8.8×10^{-5}	2.5×10^{-7}	1.9×10^{-3}
rel-stop	1.2×10^{-3}	1.8×10^{-4}	5.2×10^{-7}	3.7×10^{-3}
moment	0.041	0.33	0.094	0.0088
L^2-err	.13	.52	.17	.11
L^∞-err	.52	1.6	.68	.50
meas-err	.057	.19	.093	.019
L^2	.69	.86	.70	.70
Bg	∞	1.8	2.8	6.6
B-S	-.53	-.45	-.54	-.53

Table 1: Various Comparisons, $x^* = \chi_{[0,1/2]}$

Our first test case is the characteristic function of the interval $[0, 1/2]$ where Ω is the interval $[0, 1]$:

$$x^*(t) = \begin{cases} 1 & t \in [0, 1/2] \\ 0 & t \in (1/2, 1]. \end{cases} \tag{5.1}$$

Our basis functions in this example are the five orthonormal trigonometric functions, 1, $\sqrt{2}\sin(\pi t)$, $\sqrt{2}\cos(\pi t)$, $\sqrt{2}\sin(2\pi t)$ and $\sqrt{2}\cos(2\pi t)$. The solution forms given in § 3.2 apply to this case as well, with the trigonometric functions substituted for the polynomials t^{i-1}. In Figure 2 we compare the reconstructions with the actual true function that generated the first five moments. Table 1 contains the following information. Across the table is the list of our five estimates. Down the table are respectively the number of iterations used in the Newton method to compute our solution (iters), how small the residual, the gradient of the dual functional G, equivalently, the error in the moments of the the reconstruction versus the original function, was for our given moments (stop), the relative error (rel-stop) which is the norm of the final gradient over the norm of the original moment vector b,

$$\text{rel-stop} = \frac{||b - A(\phi^*)'(A^T\lambda_n)||_\infty}{||b||_\infty}, \tag{5.2}$$

the largest relative error made in our estimate $\hat{x}$ of the next five moments (moment),

$$\text{moment} \;=\; \frac{\max\limits_{i=n+1,n+5}\left\{\left|\int(\hat{x}-x^*)\psi_i\right|\right\}}{\max\limits_{i=n+1,n+5}\left\{\left|\int x^*\psi_i\right|\right\}}, \tag{5.3}$$

the L^2 norm of the difference between our estimate and the ideal band pass function (L^2-err),

$$L^2\text{-err} \;=\; \sqrt{\int(\hat{x}-x^*)^2}, \tag{5.4}$$

the L^∞ norm of the difference between our estimate and the known band pass function (L^∞-err),

$$L^\infty\text{-err} \;=\; \max_{t\in\Omega}|\hat{x}(t)-x^*(t)| \tag{5.5}$$

and lastly the meas-err. This last quantity needs some explanation. In [13] it was shown that as the number of basis functions goes to infinity, the estimates we are discussing all approach the function generating the moments in measure. In [6] it is shown that a sequence f_n converges to f in measure if and only if $\rho(f_n - f) \to 0$ where

$$\rho(f) = \int_\Omega \frac{|f(t)|}{1+|f(t)|}dt. \tag{5.6}$$

Thus this last quantity is $\rho(f_n - f)$ where f_n is our estimate, and f is the ideal band pass function. Thus this last quantity is a rough measure of how close our estimate is to the true function in measure.

Table 1 also contains our so-called homefield advantage test. The last three rows contain the corresponding measures of each of our entropies. Thus in the L^2 row second column of the table is found the size of the L^2 norm of the Burg estimate. As expected, the smallest numbers occur down the main diagonal, that is, the function reconstructed by minimizing a functional should have that measure smaller than the other reconstructions.

We don't show the Fermi-Dirac measure as most reconstructions exceed one and thus the F-D measure is usually $+\infty$.

Remarks: Figure 2 is the basis of our visual test. The Burg estimate is to us the least pleasing. It has the largest visual errors although it doesn't have the large errors (due to the periodicity of our basis functions) at the right end of the interval. The Fermi-Dirac estimate is the most pleasing as it is very close to the ideal band pass function. The L^2 and Boltzmann-Shannon estimates look very similar, but the L^2 estimate seems closer to

	L^2	Bg	B-S	F-D
iters	15	10	17	7
stop	6.5×10^{-3}	1.3×10^{-3}	2.7×10^{-5}	2.3×10^{-3}
rel-stop	3.3×10^{-2}	6.5×10^{-3}	1.4×10^{-4}	1.2×10^{-2}
moment	9.1×10^{-5}	4.3×10^{-3}	1.1×10^{-4}	2.7×10^{-3}
L^2-err	.094	.16	.084	.10
L^∞-err	1.2	4.9	8.1	.42
meas-err	.047	.085	.050	.21
L^2	.35	.38	.37	.35
Bg	∞	3.2	35.5	6.7
B-S	-.32	-.30	-.31	-.32

Table 2: Various Comparisons, $x^* = \cos(10^x)^{10}$

the true function. Another estimate similar to the Fermi-Dirac estimate can be achieved by imposing measurable bounds on the various entropies (see [9] and [20]).

Table 1 shows fairly fast convergence and reasonable residual errors in the given moments for each of the methods. Also, each method gives reasonable estimates to the next five moments of the given function. As expected from Figure 2, we see the Burg estimate is the furthest off in each of the last three measures. Table 1 also shows, as expected, that each estimate performs best on the functional it optimizes. Also, as expected the Burg functional is infinite on the best L^2 estimate, since this estimate is zero on a set of positive measure, although the ideal band pass function is itself zero on a set of positive measure. Overall, it seems that the Boltzmann-Shannon estimate gives a reasonable value in each category.

5.3 Locating Spikes

Another application of function reconstruction is to locate spikes in the function that generated the moments. To see how each of our estimates performs this task we generate ten moments using the basis functions t^{i-1}, $i = 1, \ldots 10$, and our ideal function generating the data is given by $f(x) = \cos(10^x)^{10}$. This function attains a maximum value of one three times on the interval $[0, 1]$, and is analytic on $[0, 1]$ but nearly zero on intervals there.

Figure 3 is the basis of our visual test in this case. The Burg reconstruction locates the positions of the spikes quite well, but over estimates the first maximum value. The

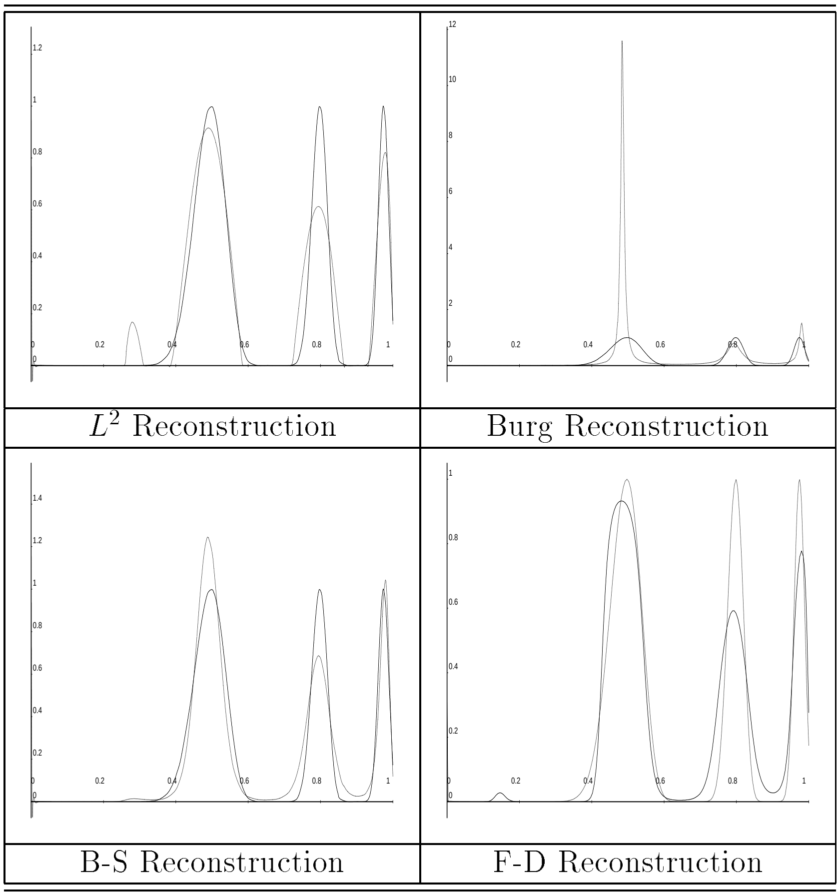
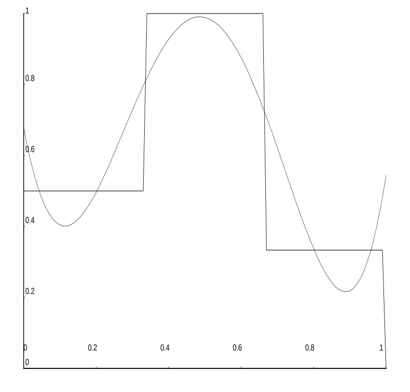
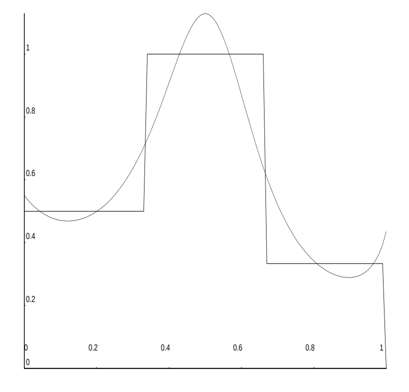


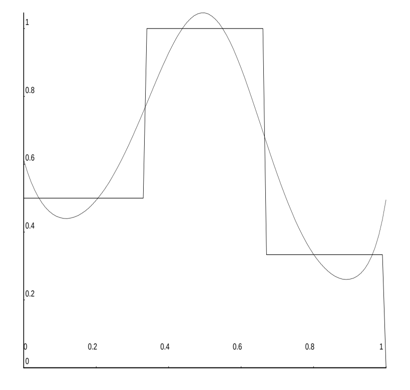
Figure 3: $\cos(10^x)^{10}$ with 10 Hausdorff moments



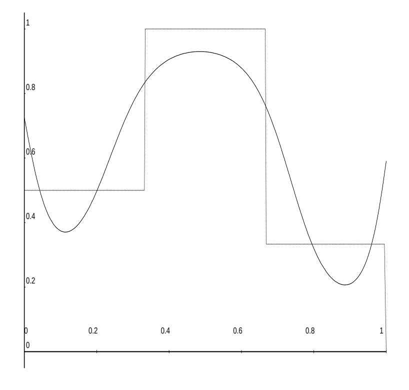
L^2 Reconstruction



Burg Reconstruction



B-S Reconstruction



F-D Reconstruction

Figure 4: $\frac{1}{2}\chi_{[0,1/3]} + \chi_{[1/3,2/3]} + \frac{1}{3}\chi_{[2/3,1]}$ with 6 Hausdorff moments

L^2 and Boltzmann-Shannon estimates are similar, but the L^2 estimate has a small false maximum to the left of the first true maximum. In this example, it appears that the Fermi-Dirac reconstruction is the most accurate.

Table 2 shows a relatively poor estimate for the next five moments for the Fermi-Dirac estimate. The L^∞ error shows the overestimate for each of the approximations except the Fermi-Dirac estimate which is forced to lie between 0 and 1 and therefore does a better job of staying close to the underlying function.

5.4 A Non-zero step function

	L^2	Bg	B-S	F-D
iters	1	6	5	4
stop	3×10^{-15}	8.6×10^{-6}	1.1×10^{-5}	3.1×10^{-3}
rel-stop	4.9×10^{-15}	1.4×10^{-5}	1.9×10^{-5}	5.1×10^{-3}
moment	1.5×10^{-2}	5.0×10^{-3}	9.6×10^{-3}	1.6×10^{0}
L^2-err	.13	.12	.12	.15
L^∞-err	.37	.36	.33	.42
meas-err	.089	.074	.076	.10
L^2	.66	.66	.66	.66
Bg	.59	.58	.58	.60
B-S	-.86	-.86	-.86	-.86

Table 3: Various Comparisons, $x^* = \frac{1}{2}\chi_{[0,1/3]} + \chi_{[1/3,2/3]} + \frac{1}{3}\chi_{[2/3,1]}$

Since the Burg estimate by construction does not reconstruct a function that is zero on a set of positive measure, we consider the case of a step function with three different positive values as our ideal function. We generate 6 moments using a polynomial basis.

Figure 4 shows that the estimates look quite similar to each other. The Boltzmann-Shannon and Burg estimates slightly overestimate the middle step, while the L^2 and Fermi-Dirac estimates underestimate the middle.

Table 3 again shows good agreement between all the estimates except the Fermi-Dirac estimate. The homefield test also shows good agreement of all the estimates.

5.5 A Function that Rapidly Decreases

We take the function $\exp(-100x^4)$ and generate five moments with the polynomial basis functions. This function is never zero, but it does decrease to zero rapidly.

Figure 5 shows a very good agreement with the Boltzmann-Shannon estimate. This is to be expected at the form of the Boltzmann-Shannon estimate is the same form as the ideal function in this case. We see a good agreement with the Fermi-Dirac function as well as the L^2 solution. The Burg reconstruction again has problems because the values of the function are close to zero.

Table 4 shows good agreement between the the estimates with the Boltzmann-Shannon estimate doing the best for reasons already discussed. The homefield test shows reasonable agreement between the estimates except for the Burg functional since the values of the

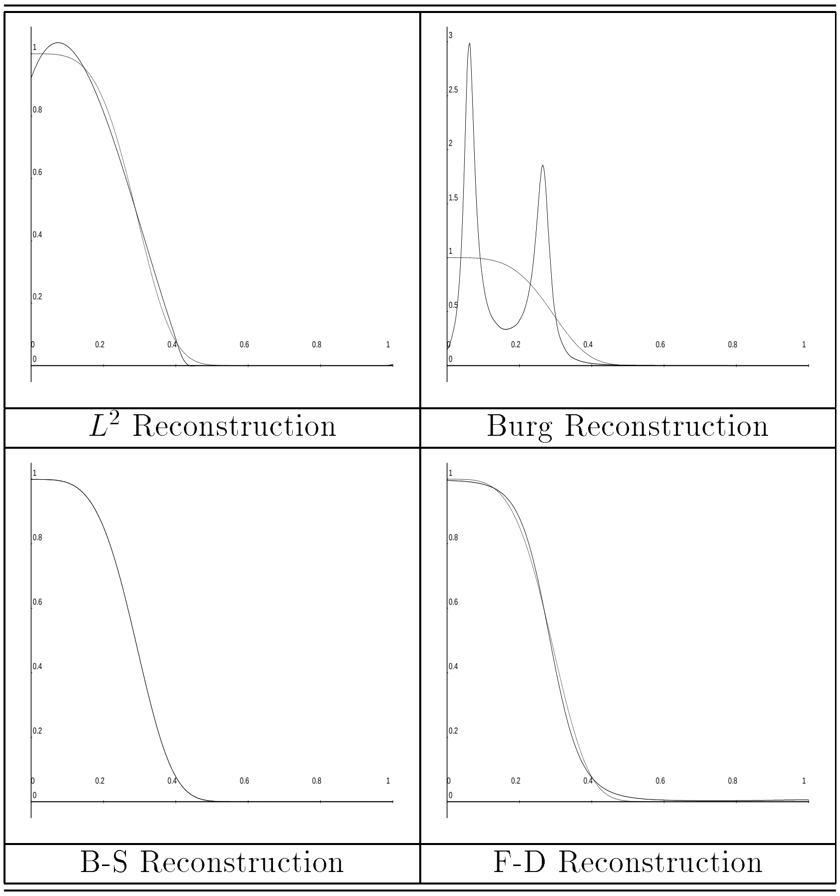


Figure 5: $\exp(-100x^4)$ with 5 Hausdorff moments

	L^2	Bg	B-S	F-D
iters	12	16	15	6
stop	1.4×10^{-6}	2.1×10^{-5}	8.6×10^{-6}	1.9×10^{-3}
rel-stop	4.9×10^{-6}	7.5×10^{-5}	2.9×10^{-4}	6.6×10^{-3}
moment	1.3×10^{-5}	1.7×10^{-4}	1.7×10^{-7}	7.6×10^{-4}
L^2-err	2.0×10^{-2}	4.0×10^{-1}	1.1×10^{-4}	1.4×10^{-2}
L^∞-err	7.8×10^{-2}	2.0	5.5×10^{-4}	4.3×10^{-2}
meas-err	1.2×10^{-2}	.11	6.4×10^{-5}	1.0×10^{-2}
L^2	.49	.63	.49	.49
Bg	∞	4.7	19.8	3.3
B-S	-.36	-.27	-.36	-.37

Table 4: Various Comparisons, $x^* = \exp(-100x^4)$

function are so near zero.

6. Conclusions

We must reiterate that the problem we are addressing is the reconstruction of a function given *accurate* moments. Indeed, if we add noise to the data our method of solving becomes very unstable: the condition numbers of the Hessian matrices are very large, particularly if we use the Hausdorff moments. The trigonometric moments are much more stable in this respect, suggesting that the choice of basis is quite important to the stability of Newton's method in this context. We investigate this problem in detail in a later paper.

We have seen that the various entropies give reconstructions that look different, but from a quantitative standpoint as measured above are all very similar. We have found no measure which makes any reconstruction from one entropy stand out as the best. It can be convincingly argued, however, that Burg's entropy is the least appealling due both to the lack of solid theory backing it (solutions may not even exist, for example) and the distinctly poor visual reconstructions of densities with significant zero sections. In contrast, the Boltzmann-Shannon entropy possesses the compactness needed to guarantee solutions exist in L^1, is easy to implement and the computations are stable particularly in comparison with Burg's. The Fermi-Dirac entropy generally does quite good, but the restriction that the underlying function's range is known may be unrealistic in certain applications. We have found that in the case that the range is known, the Fermi-Dirac produces more visually pleasing solutions than truncated L^2 solutions for example. The Fermi-Dirac reconstruction also performs well in some problems with very few moments.

Based on the observation that no single entropy stands out in every situation, we recommend using the Boltzmann-Shannon or L^2 entropies, for there we have the theoretical backing and the computational ease that makes these entropies more useful than the others.

References

[1] C. Auyeung and R. M. Mersereau. A dual approach to signal restoration. *IEEE International Conference on Acoustics, Speech and Signal Processing, Glasgow*, pages 1326–1329, May 1989.

[2] J. M. Borwein. Direct theorems in semi-infinite programming. *Math Programming*, (21):301–318, 1981.

[3] J. M. Borwein and W. Z Huang. A heuristic method for polynomial moment problems with Boltzmann–Shannon entropy. Preprint.

[4] J. M. Borwein and A. S. Lewis. Duality relationships for entropy–like minimization problems. *SIAM Journal of Control Theory and Optimization*, 29(2):325–338, March 1991.

[5] J. M. Borwein and A. S. Lewis. Partially-finite convex programming, parts I and II. *Mathematical Programming*, pages 15–48,49–83, 1992.

[6] J. M. Borwein and A. S. Lewis. A survey of convergence results for maximum entropy methods. In *Proceedings of MaxEnt 92*. Kluwe Publishers, August 1992. University of Waterloo Research Report COOR 92–32.

[7] J. M. Borwein, A. S. Lewis, and M. A. Limber. Entropy minimization with lattice bounds. Technical report, University of Waterloo, 1992. CORR 92-05.

[8] J. P. Burg. Maximum entropy spectral analysis. *Paper presented at 37th meeting of the Society of Exploration Geophysicists, Oklahoma City*, 1967.

[9] R. Cole and R. K. Goodrich. L_p spectral analysis with an L_∞ upper bound. *Journal of Optimization Theory and Applications*, to appear.

[10] J. Conway. *A Course In Functional Analysis*. Springer-Verlag, New York, New York, 1985.

[11] A. Decarreau, D. Hilhorst, C. Lemaréchal, and J. Navaza. Dual methods in entropy maximization. Application to some problems in crystallography. *SIAM J. Optimization*, 2(2):173–197, May 1992.

[12] J. E. Dennis, Jr. and R. B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall, Inc., New Jersey, 1983.

[13] N. Dunford and J. T. Schwartz. *Linear Operators, Part 1: General Theory*. John Wiley & Sons, 1988.

[14] N. J. Dusaussoy and I. E. Abdou. The extended MENT algorithm: a maximum entropy type algorithm using prior knowledge for computerized tomography. *IEEE Transactions on Signal Processing*, 39(5):1164–1180, May 1991.

[15] R. K. Goodrich and A. Steinhardt. L_2 spectral estimation. *SIAM J. Appl. Math.*, 46:417–426, June 1986.

[16] R. K. Goodrich, A. Steinhardt, and R. Roberts. Spectral estimation via minimum energy correlation extension. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-33(6):1509–1515, December 1985.

[17] E. T. Jaynes. Information theory and statistical mechanics. *The Physical Review*, 106, 108:620–630, 171–190, 1957.

[18] R. W. Johnson and J. E. Shore. Which is the better entropy expression for speech processing: $-s \log s$ or $\log s$? *IEEE Trans on Acoustics, Speech and Signal Processing*, February 1984.

[19] L. K. Jones and V. Trutzer. Computationally feasible high-resolution minimum-distance procedures which extend the maximum-entropy method. *Inverse Problems*, pages 749–766, 1989.

[20] M. A. Limber and R. K. Goodrich. Quasi interiors, Lagrange multipliers and L^p spectral estimation with lattice bounds. *Journal of Optimization Theory and Applications*, 78(1), July 1993.

[21] M. A. Limber, T. A. Manteuffel, S. F. McCormick, and D. S. Sholl. Optimal resolution in maximum entropy image reconstruction from projections with multigrid

acceleration. In *Proceedings of the Sixth Annual Copper Mountain Conference on Multigrid Methods*, 1993.

[22] D. G. Luenberger. *Optimization by Vector Space Methods*. John Wiley & Sons, Inc., New York, New York, 1969.

[23] J. Navaza. The use of non-local constraints in maximum-entropy electron density reconstruction. *Acta Cryst.*, 42:212–223, 1986.

[24] F. P. Preparata and M. I. Shamos. *Computational Geometry, and Introduction*. Springer–Verlag, New York, New York, 1985.

[25] M. L. Reis and N. C. Roberty. Maximum entropy algorithms for image reconstruction from projections. *Inverse Problems*, pages 623–644, 1992.

[26] R. T. Rockafellar. *Conjugate Duality and Optimization*. SIAM, Philadelphia, PA., 1974.

[27] C. E. Shannon and W. Weaver. *The mathematical theory of communication*. University of Illinois Press, Urbana, Illinois, 1963.

[28] A. B. Templeman. Entropy-based optimization methods for engineering design. Preprint.

[29] B. Turkington and N. Whitaker. Statistical equilibrim solutions to the Euler equations. Preprint.

[30] R. J. B. Wets. Convergence of convex functions, variational inequalities and convex optimization problems. In Cottle, Giannessi, and Lions, editors, *Variational Inequalities and Complementarity Problems*, New York, New York, 1979. J. Wiley.