Maximum Entropy-type Methods
and (Non-)Convex Programming

Prepared for
AMS Special Session on
Nonsmooth Analysis in
Variational and Imaging Problems

January 7–8, 2005

Jonathan M. Borwein, Canada Research Chair

Faculty of Computer Science, Halifax, NS Canada

URL: www.cs.dal.ca/~jborwein

Revised: January 2, 2005
“I feel so strongly about the wrongness of reading a lecture that my language may seem immoderate. … The spoken word and the written word are quite different arts. … I feel that to collect an audience and then read one's material is like inviting a friend to go for a walk and asking him not to mind if you go alongside him in your car.”

(Sir Lawrence Bragg)
“If my teachers had begun by telling me that mathematics was pure play with presuppositions, and wholly in the air, I might have become a good mathematician. But they were overworked drudges, and I was largely inattentive, and inclined lazily to attribute to incapacity in myself or to a literary temperament that dullness which perhaps was due simply to lack of initiation.”

(George Santayana)


I shall discuss in “tutorial mode” the formalization of inverse problems such as signal recovery and option pricing as (convex and non-convex) optimization problems over the infinite dimensional space of signals. I shall touch on* the following:

1. The impact of the choice of “entropy” (e.g., Boltzmann-Shannon, Burg entropy, Fisher information) on the well-posedness of the problem and the form of the solution.

2. Convex programming duality: what it is and what it buys you.

3. Algorithmic consequences.


† Related papers at www.cecm.sfu.ca/preprints/

*More is an unrealistic task!
THE PROBLEM

• Many applied problems reduce to “best” solving (underdetermined) systems of linear (or non-linear) equations $Ax = b$, where $b \in \mathbb{R}^n$, and the unknown $x$ lies in some appropriate function space. Discretization reduces this to a finite dimensional setting where $A$ is now a $m \times n$ matrix.

◊ In many cases, I believe it is better to address the problem in its function space home, discretizing only as necessary for computation.

• Thus, the problem often is how do we estimate $x$ from a finite number of its 'moments'? This is typically an underdetermined linear inversion problem where the unknown is most naturally a function, not a vector in $\mathbb{R}^m$. 
• Consider, extrapolating an autocorrelation function $R(t)$ given several sample measurements.

○ The Fourier transform $S(z)$ of the autocorrelation function is the power spectrum of the data. Fourier moments of the power spectrum are the same as samples of the autocorrelation function, so by computing several values of $R(t)$ directly from the data, we are in essence computing moments of $S(z)$.

• If we compute a finite number of moments of $S$, we can then estimate $S$ from these moments, and we may compute more moments from the estimate $\hat{S}$ by direct numerical integration, thereby affording an extrapolation of $R$, without directly computing $R$ from the potentially noisy data.
THE ENTROPY APPROACH

• Following (BLi) I sketch a maximum entropy approach to underdetermined systems where the unknown, \( x \), is a function, typically living in a *Hilbert space*, or more general space of functions. This technique picks a “best” representative from the infinite set of *feasible* functions (functions that possess the same \( n \) moments as the sampled function) by minimizing an integral functional, \( f \), of the unknown.

◊ The approach finds applications including: acoustics, constrained spline fitting, image reconstruction, option pricing, multidimensional NMR, tomography, statistical moment fitting, and time series analysis.

• However, the derivations and mathematics are fraught with subtle errors. I will discuss some of the difficulties inherent in infinite dimensional calculus, and provide a simple theoretical algorithm for correctly deriving maximum entropy-type solutions.
WHAT is...

**Boltzmann (1844-1906)**

**Shannon (1916-2001)**
WHAT is ENTROPY?

Despite the narrative force that the concept of entropy appears to evoke in everyday writing, in scientific writing entropy remains a thermodynamic quantity and a mathematical formula that numerically quantifies disorder. When the American scientist Claude Shannon found that the mathematical formula of Boltzmann defined a useful quantity in information theory, he hesitated to name this newly discovered quantity entropy because of its philosophical baggage. The mathematician John Von Neumann encouraged Shannon to go ahead with the name entropy, however, since “no one knows what entropy is, so in a debate you will always have the advantage.”

- **19C**: Boltzmann—thermodynamic disorder
- **20C**: Shannon—information uncertainty
- **21C**: JMB—potentials with superlinear growth
Information theoretic characterizations abound. A nice one is:

**Theorem 1** \( H(\overrightarrow{p}) = -\sum_{k=1}^{N} p_k \log p_k \) is the unique continuous function (up to a positive scalar multiple) on finite probabilities such that

**I. Uncertainty grows:**

\[
H \left( \frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n} \right)
\]

increases with \( n \).

**II. Subordinate choices are respected:** for distributions \( \overrightarrow{p_1} \) and \( \overrightarrow{p_2} \) and \( 0 < p < 1 \),

\[
H (p \overrightarrow{p_1}, (1-p) \overrightarrow{p_2}) = p H(\overrightarrow{p_1}) + (1-p) H(\overrightarrow{p_2})
\]
Let $X$ be our function space, typically Hilbert space $L^2(\Omega)$, or the function space $L^1(\Omega)$.

For $p \geq 1$,

$$L^p(\Omega) = \left\{ x \text{ measurable} : \int_{\Omega} |x(t)|^p dt < \infty \right\}.$$ 

It is well known that $L^2(\Omega)$ is a Hilbert space with inner product $\langle x, y \rangle = \int_{\Omega} x(t)y(t) dt$.

A bounded linear map $A : X \to \mathbb{R}^n$ is set by

$$(Ax)_i = \int x(t)a_i(t) dt$$

for $i = 1, \ldots, n$ and $a_i \in X^*$ the ‘dual’ of $X$ ($L^2$ in the Hilbert case, $L^\infty$ in the $L^1$ case).
To pick a solution from the infinitude of possibilities, we may freely define “best”. The most common approach is to find the minimum norm solution*, by solving the Gram system

\[ A A^T \lambda = b. \]

The solution is then \( \hat{x} = A^T \lambda. \) This recaptures all of Fourier analysis!

This actually solved the following variational problem:

\[
\min \left\{ \int_\Omega x(t)^2 dt : Ax = b \; x \in X \right\}. 
\]

*Even in the (realistic) infeasible case.
• We generalize the norm with a strictly convex functional $f$ as in

$$\min \{ f(x) : Ax = b, \ x \in X \} , \quad (P)$$

where $f$ is what we call, an entropy functional, $f : X \to (-\infty, +\infty]$. Here we suppose $f$ is a strictly convex integral functional* of the form

$$f(x) = \int_{\Omega} \phi(x(t))dt.$$ 

The functional $f$ can be used to include other constraints†. For example,

$$f(x) = \begin{cases} \int_0^1 x(t)^2 dt & \text{if } x \geq 0 \\ +\infty & \text{else} \end{cases}$$

is the constrained $L^2$ norm functional (‘positive energy’), used in constrained spline fitting.

*Essentially $\phi''(t) > 0$.
†Including nonnegativity, by appropriate use of $+\infty$. 

13
Two popular choices for $f$ are the Boltzmann-Shannon entropy (in image processing)

$$f(x) = \int x \ln x,$$

and the Burg entropy (in time series analysis),

$$f(x) = -\int \ln x.$$

Both implicitly impose nonnegativity (positivity in Burg case) constraint.

There has been much information-theoretic debate about which entropy is best. This is more theology than science!

More recently the use of Fisher Information

$$f(x, x') = \int \frac{x'^2}{2x}$$

has become more usual as it penalizes large derivatives; and can be argued for physically.
WHAT CAN GO WRONG?

• Consider solving $Ax = b$, where, $b \in \mathbb{R}^n$ and $x \in L^2[0, 1]$. Assume further that $A$ is a continuous linear map, hence represented as above.

• As $L^2$ is infinite dimensional, and $\mathbb{R}^n$ is finite dimensional, the null space of $A$ is infinite dimensional: if there are any solutions to $Ax = b$, there is an infinite number. We pick our solution to minimize the functional

$$f(x) = \int \phi(x(t)) \, dt$$

[$\phi(x(t), x'(t))$ in Fisher-like cases (BN1,BN2)].

• We introduce the Lagrangian

$$L(x, \lambda) = \int_0^1 \phi(x(t)) \, dt + \sum_{i=1}^n \lambda_i \left( b_i - \langle x, a_i \rangle \right),$$

and the associated dual problem

$$\max_{\lambda \in \mathbb{R}^n} \min_{x \in X} \{ L(x, \lambda) \}. \quad (D)$$
• So we formally have a “dual pair” (BL1)

$$\min \left\{ f(x) : Ax = b, \ x \in X \right\}, \quad (P)$$

and

$$\max \min_{\lambda \in \mathbb{R}^n} \min_{x \in X} \{L(x, \lambda)\}. \quad (D)$$

• Moreover, for the solutions \( \hat{x} \) to \((P)\), \( \hat{\lambda} \) to \((D)\), the derivative (w.r.t. \( x \)) of \( L(x, \hat{\lambda}) \) should be zero, since \( L(\hat{x}, \hat{\lambda}) \leq L(x, \hat{\lambda}), \forall x \). This implies

$$\hat{x}(t) = (\phi')^{-1} \left( \sum_{i=1}^{n} \hat{\lambda}_i a_i(t) \right)$$

$$= (\phi')^{-1} (A^T \hat{\lambda}) .$$

• This allows us to reconstruct the primal solution (qualitatively and quantitatively) from a presumptively easier dual computation.
There are two major problems with this approach.*

1. **The assumption that a solution \( \hat{x} \) exists.** For example, consider the problem

\[
\inf_{x \in L^1[0,1]} \left\{ \int_0^1 x(t)dt : \int_0^1 tx(t)dt = 1, x \geq 0 \right\}.
\]

\(\diamond\) The optimal value is not attained. Similarly, existence can fail for the Burg entropy with trig moments. Additional conditions on \( \phi \) are needed to insure solutions exist\(\dagger\) (see BL2).

2. **The assumption that the Lagrangian is differentiable.** In the above example, \( f \) is \( +\infty \) for every function \( x \) negative on a set of positive measure. This implies the Lagrangian is \( +\infty \) on a dense subset of \( L^1 \), the set of functions *not* nonnegative a.e.. The Lagrangian is nowhere continuous, much less differentiable.

* A third, the existence of \( \hat{\lambda} \), is less difficult to surmount.
\(\dagger\) The solution is actually the *absolutely continuous part of a measure* in \( C(\Omega) \).*
One approach to circumvent the differentiability problem, is to pose the problem in $L^\infty(\Omega)$, or in $C(\Omega)$, the space of essentially bounded, or continuous, functions. However, in these spaces, even with additional side qualifications, we are not necessarily assured solutions to $(P)$ exist.

In (BL2), there is an example of a problem where $\Omega \subset \mathbb{R}^3$, the moments are the 4 Fourier coefficients, and the entropy is Burg’s, yet no solutions exist for certain feasible data values.

Another example, Minerbo poses the problem of tomographic reconstruction in $C(\Omega)$ with the Boltzmann-Shannon entropy. However, there the functions $a_i$ are characteristic functions of strips across $\Omega$, and the solution is piecewise constant, not continuous.
We state a theorem that guarantees that the form of solution found in the above faulty derivation \( \hat{x} = (\phi')^{-1}(A^T\hat{\lambda}) \) is, in fact, correct. A full derivation is given in (BL2).

- We introduce the *Fenchel (Legendre) conjugate* (see BL1) of a function \( \phi : IR \rightarrow (-\infty, +\infty] \):

  \[
  \phi^*(u) = \sup_{v \in IR} \{uv - \phi(v)\}.
  \]

- Often this can be computed explicitly, using Newtonian calculus. Thus,

  \[
  \phi(v) = v \log v, -\log v \text{ and } v^2/2
  \]

  yield

  \[
  \phi^*(u) = \exp(u - 1), -1 - \log(-u) \text{ and } u^2/2
  \]

  respectively.

- The Fisher case is similarly explicit.
The Hoch and Stern information measure, or neg-entropy, is defined in complex $n$–space by

$$H(z) = \sum_{j=1}^{n} h(z_j/b),$$

where $h$ is convex and given (for scaling $b$) by:

$$h(z) \triangleq |z| \ln \left(|z| + \sqrt{1 + |z|^2}\right) - \sqrt{1 + |z|^2}$$

for quantum theoretic (NMR) reasons.

- Recall the Fenchel-Legendre conjugate

$$f^*(y) := \sup_x \langle y, x \rangle - f(x).$$

- Our symbolic convex analysis package (stored at www.cecm.sfu.ca/projects/CCA/) produced:

$$h^*(z) = \cosh(|z|)$$

- Compare the Shannon entropy:

$$(z \ln z - z)^* = \exp(z).$$
COERCIVITY AND DUALITY

• We say $\phi$ possess regular growth if either $d = \infty$, or $d < \infty$ and $k > 0$, where $d = \lim_{u \to \infty} \phi(u)/u$ and $k = \lim_{v \uparrow d} (d - v)(\phi^*)'(v)$.

• The domain of a convex function is $\text{dom}(\phi) = \{u : \phi(u) < +\infty\}$; $\phi$ is proper if $\text{dom}(\phi) \neq \emptyset$. Let $i = \inf \text{dom}(\phi)$ and $\sigma = \sup \text{dom}(\phi)$.

• Our constraint qualification†, $(CQ)$, reads

$$\exists \bar{x} \in L^1(\Omega), \text{ such that } A\bar{x} = b, \quad f(\bar{x}) \in \mathbb{R}, \quad i < \bar{x} < \sigma \text{ a.e.}$$

\diamond In many cases, $(CQ)$ in fact reduces to feasibility, and trivially holds.

• In this language, the dual problem for $(P)$ is

$$\sup \left\{ \langle b, \lambda \rangle - \int_\Omega \phi^*(A^T\lambda(t))dt \right\}. \quad (D)$$

*-log does nor possess regular growth; $v \to v \ln v$ does.

†Slater’s condition fails; this is what guarantees dual solutions exist.
Theorem 1 (BL2) Suppose $\Omega$ is a finite interval, $\mu$ is Lebesgue measure, each $a_k$ continuously differentiable (or just locally Lipschitz) and $\phi$ is proper, strictly convex with regular growth. Suppose $(CQ)$ holds and also

(1) $\exists \tau \in \mathbb{R}^n$ such that $\sum_{i=1}^{n} \tau_i a_i(t) < d \ \forall t \in [a, b],$

then the unique solution to $(P)$ is given by

(2) $\hat{x}(t) = (\phi^*)'(\sum_{i=1}^{n} \tilde{\lambda}_i a_i(t))$

where $\tilde{\lambda}$ is any solution to dual problem $(D)$.

- This theorem generalizes to cover the case $\Omega \subset \mathbb{R}^n$, and more elaborately in Fisher-like cases. The results can be found in (BL2, BN1).
• What Theorem 1 means in practice is that the form of the maximum entropy solution can be legitimated simply by validating the easily checked conditions of Theorem 1.

• Also, any solution to $Ax = b$ of the form in (2) is automatically a solution to $(P)$. Thus, finding solutions to $(P)$ is equivalent to solving the nonlinear system of equations

$$
\langle (\phi^*)'(A^T \lambda), a_i \rangle = b_i, \quad i = 1, \ldots, n
$$

for $\lambda \in \mathbb{R}^n$: a finite dimensional system of nonlinear equations.

One can then apply a standard ‘industrial strength’ nonlinear equation solver, like Newton’s method, to this system, to find the optimal $\lambda$.

• In many cases, $\boxed{ (\phi')^{-1} = (\phi^*)' }$ and so the incorrectly derived solution agrees with the ‘honest’ solution. Importantly, we may tailor $(\phi')^{-1}$ to our needs.
Note that discretization is only needed to compute the terms in (3). Indeed, these integrals can sometimes be computed exactly (in some tomography and option estimation problems, [BCM]). This is what we gain by *not discretizing* too early. By waiting to see what form the dual problem takes, one can customize one’s integration scheme to the problem at hand.

Even when this is not the case one can often use the shape of the dual solution to fashion very efficient heuristic reconstructions that avoid any iterative steps (see BN2).
• The *MomEnt* project at CECM, provides teaching code ([www.cecm.sfu.ca/interfaces/](http://www.cecm.sfu.ca/interfaces/), presently being upgraded) implementing the entropic reconstructions described above.

◊ Types of moments, entropies and problem dimension are easily varied, and the code allows several methods of solving the dual problem, including Newton’s method, quasi-Newton methods (BFGS, DFP), the method of conjugate gradients, and the Barzilai-Borwein gradient line-search free method.

• It also allows one to explore the effect of adding noise or of allowing relaxations of the constraints.
• The positive $L^2$, Boltzmann-Shannon and Burg entropy reconstruction of the characteristic function of $[0, 1/2]$ using 10 algebraic moments ($b_i = \int_{0}^{1/2} t^{i-1} dt$) on $\Omega = [0, 1]$.

• Solution is $\hat{x}(t) = (\phi^*)'(\sum_{i=1}^{n} \tilde{\lambda}_i t^{i-1})$.

Burg over-oscillates since $(\phi^*)'(t) = 1/t$. 

26
THE NON-CONVEX CASE

• In general non-convex optimization is a much less satisfactory field. We can usually hope only to find critical points \( f'(x) = 0 \) or local minima. Thus problem specific heuristics dominate.

• **Crystallography**: We of course wish to estimate \( x \) in \( L^2(\mathbb{R}^n) \)\(^*\). Then the modulus \( c = |\hat{x}| \) is known (\( \hat{x} \) is the Fourier transform of \( x \)).\(^\dagger\)

Now \( \{y : |\hat{y}| = c\} \), is not convex. So the issue is to find \( x \) given \( c \) and other convex information. An appropriate optimization problem extending the previous one is

\[
\min \{ f(x) : Ax = b, \|Mx\| = c, \ x \in X \}, \ \ (NP)
\]

where \( M \) models the modular constraint, and \( f \) is as in Theorem 1.

\(^*\)Here \( n = 2 \) for images, 3 for holographic imaging, etc.

\(^\dagger\)Observation of the modulus of the diffracted image in crystallography. Similarly, for optical abberation cor-

rection.
A CAUTION

• My collaborator Patrick Combettes is expert on various optimization perspectives on cognates to \((NP)\) and related feasibility problems.

◊ Most methods rely on a two-stage (easy convex, hard non-convex) decoupling schema—the following from Decarreau et al. (D). They suggest solving

\[
\min \{ f(x) : Ax = y, \|B_k y\| = b_k, \ x \in X \},
\]

\((NP^*)\)

where \(\|B_k y\| = b_k, k \in K\) encodes the hard modular constraints.

• They write down formal first-order Kuhn-Tucker conditions for a relaxed form of \((NP^*)\). The easy constraints are treated by Theorem 1. I am obscure, largely because the results were largely negative:
• They applied these ideas to a prostaglandin molecule (25 atoms), using quasi-Newton (which could fail to find a local min), truncated Newton (better) and trust-region (best) numerical schemes.

◊ They observe that the “reconstructions were often mediocre” and highly dependent on the amount of prior information – a small proportion of unknown phases to be satisfactory.

“Conclusion: It is fair to say that the entropy approach has limited efficiency, in the sense that it requires a good deal of information, especially concerning the phases. Other methods are wanted when this information is not available.”

• Thus, I offer my presentation largely to illustrate the difficulties.
### REFERENCES


