List of Figures

Date: August 9, 2011.

2010 *Mathematics Subject Classification.* Primary 45Q05; Secondary 49N45 90C25 49N15 90C90.

Key words and phrases. Entropy optimization, Fenchel duality theorem, inverse problems, feasibility problems, alternating projection and reflection methods.

Research partially supported by ARC Grant #DP1093769.

MAXIMUM ENTROPY AND FEASIBILITY METHODS FOR CONVEX AND NONCONVEX INVERSE PROBLEMS

JONATHAN M. BORWEIN

Abstract. We discuss informally two approaches to solving convex and nonconvex feasibility problems — via entropy optimization and via algebraic iterative methods. We shall highlight the advantages and disadvantages of each and give various related applications and limiting-examples. While some of the results are very classical, they are not as well known to practitioners as they should be. A key role is played by the Fenchel conjugate.

1. Introduction

> *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.* William Lawrence Bragg (Nobel crystallographer, 1890-1971)

We shall discuss in a 'tutorial mode'[1] the formalization of **inverse problems** such as signal recovery, phase retrieval and option pricing: first as (convex and nonconvex) **optimization problems** and second as **feasibility problems** — each over the infinite dimensional space of signals. We shall touch on the following:[2]

(1) The impact of the choice of "entropy" (e.g., Boltzmann-Shannon entropy, Burg entropy, Fisher information, etc.) 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: for both design and implementation.
(4) Non-convex extensions and feasibility problems: life is hard. Entropy optimization, used directly, does not have much to offer. But sometimes we observe that more works than we yet understand why it should.

[1] A companion lecture is at http://www.carma.newcastle.edu.au/~jb616/inverse.pdf.

[2] More is an unrealistic task; all details may be found in the references!

2. The General Problem

The infinite we shall do right away. The finite may take a little longer.
Stanislav Ulam (1909-1984)[3]

Many applied problems, and some rather pure ones [26], reduce to 'best' solving (under-determined) systems of **linear** (or non-linear) equations:

$$\boxed{\text{Find } x \text{ such that} A(x) = b,}$$

where $b \in I\!R^n$, 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 most cases, I believe it is better to address the problem in its function space home, discretizing only as necessary for numerical computation. One is by then more aware of the shape of the solutions and can be guided by the analysis analysis.

Thus, the problem often is *how do we estimate x from a finite number of its 'moments'?* This is typically an **under-determined inverse problem** (linear or non-linear) where the unknown is most naturally a function, not a vector in $I\!R^m$.

Example 1 (Robust autocorrelation)**.** Consider, extrapolating an *autocorrelation function* from given sample measurements:

$$R(t) := \frac{E\left[(X_s - \mu)(X_{t+s} - \mu)\right]}{\sigma}$$

The *Wiener-Khintchine theorem* says that the Fourier moments of the power spectrum $S(\sigma)$ are samples of the autocorrelation function, so values of $R(t)$ computed directly from the data yields *moments* of $S(\sigma)$.

$$R(t) = \int_R e^{2\pi it\sigma} S(\sigma) d\sigma \quad \textit{and} \quad S(\sigma) = \int_R e^{-2\pi it\sigma} R(t) dt.$$

Hence, we may compute a *finite* number of moments of S, and use them to make an estimate $\hat{S}$ of S. We may then *estimate more moments* from $\hat{S}$ by direct numerical integration. So we dually *extrapolate* R. This avoids having to compute R directly from potentially noisy (unstable) larger data series. ♢

[3] In D. MacHale, *Comic Sections* (Dublin 1993).

3. Part I: The Entropy Approach

Following [25], I now sketch a maximum entropy approach to under-determined systems where the unknown, x, is a function, typically living in a *Hilbert space*, or a more general space of functions. For Hilbert space theory an excellent new reference is [5].

> The entropy technique picks a 'best' representative from the infinite set of *feasible signals* (functions that possess the same n moments as the sampled function) by minimizing an (integral) functional, $f(x)$, of the unknown x.

The approach finds applications in myriad fields including (to my personal knowledge):

> Acoustics, actuarial science, astronomy, biochemistry, compressed sensing, constrained spline fitting, engineering, finance and risk, image reconstruction, inverse scattering, optics, option pricing, multidimensional NMR (MRI), quantum physics, statistical moment fitting, time series analysis, and tomography ... (Many thousands of papers).

Some of these are described in the examples throughout this paper.

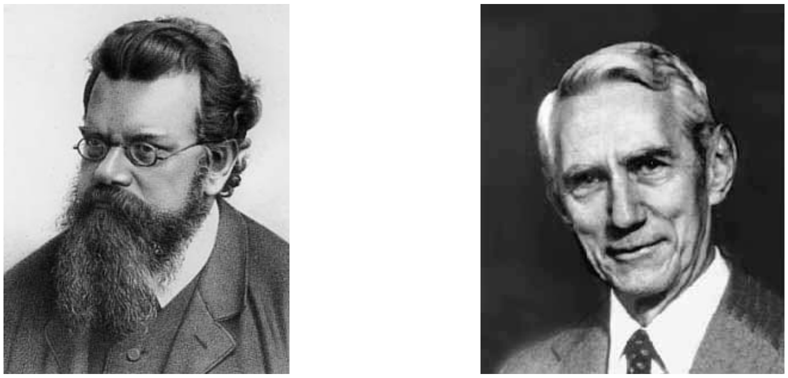
However, the derivations and mathematics are fraught with subtle — and less subtle — errors. I will next discuss some of the difficulties inherent in infinite dimensional calculus, and provide a simple theoretical algorithm for correctly deriving maximum entropy-type solutions.

4. 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."*[4]

- In the **19C** for Ludwig Boltzmann entropy was a measure of thermodynamic *disorder*;
- In the **20C** for Claude Shannon it had become information *uncertainty*;

[4] This possibly apocryphal anecdote is taken from *The American Heritage Book of English Usage*, p. 158.



(a) Boltzmann (b) Shannon

Figure 1. Boltzmann (1844-1906) and Shannon (1916-2001).

- In the **21C**: for my collaborators and others, entropies have become barrier functions (potentials, merit functions) — often with *superlinear growth*.

Information theoretic characterizations abound. A nice example is:

Theorem 1 (Entropy characterization)**.** [45] *Up to a positive multiple,*

$$H(\overrightarrow{p}) := -\sum_{k=1}^{N} p_k \log p_k$$

is the unique continuous function on finite probabilities such that:

[I.] Uncertainty grows:

$$H\left(\overbrace{\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n}}^{n}\right)$$

increases with n.

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

$$H\left(p\,\overrightarrow{p_1}, (1-p)\,\overrightarrow{p_2}\right) = p\,H(\overrightarrow{p_1}) + (1-p)\,H(\overrightarrow{p_2}).$$

5. Entropies for Us

Let X be our *function space*, typically the Hilbert space $L^2(\Omega)$ on a reasonable set Ω, or often more appropriately the function space $L^1(\Omega)$ (or a Sobolev space) where as always for $+\infty \geq p \geq 1$,

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

and we assume for simplicity that the measure is finite. (The 'infinite horizon' case with infinite measure is often more challenging and sometimes the corresponding results are false or unproven.) We also recall that $L^2(\Omega)$ is a Hilbert space with *inner product*

$$\langle x, y\rangle := \int_\Omega x(t)y(t)dt,$$

(with variations in Sobolev space).

A *bounded (continuous) linear map* $A : X \to I\!R^n$ is then fully determined 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^∞ in the L^1 case, L^q in the L^p case where $1/p + 1/q = 1)$.

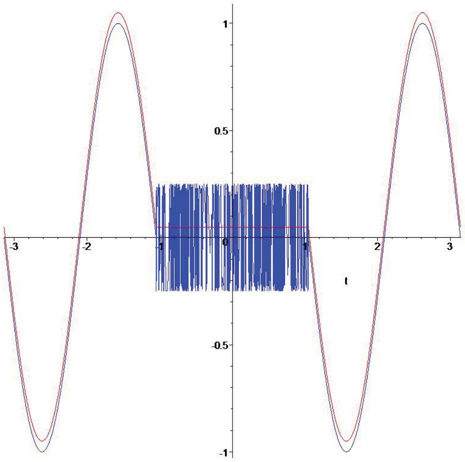


FIGURE 2. Lebesgue's continuous function with divergent Fourier series at zero.

To pick a solution from the infinitude of possibilities, we may freely define "**best**".

Example 2 (Selecting a feasible signal)**.** The most common approach, both for pragmatic reasons — ease of computation — and for theoretic reasons is to find the *minimum norm solution* — even in the (realistic[5]) infeasible case, by solving the *Gram system*:

$$\boxed{\text{Find } \lambda \text{ such that } AA^T\lambda = b}\ .$$

[5] Given noise, modeling, measurement and numerical errors, the problem may well not be feasible in practice.

For any solution $\hat{\lambda}$, a primal solution is then $\hat{x} = A^T\hat{\lambda}$.

Elaborated, this recaptures all of *Fourier analysis*, e.g., understanding Lebesgue's example illustrated in Figure 2 and much deeper Fourier analysis!

As we shall confirm later in Section 10, the Gram system solves the following *variational problem*}:

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

So we see the traditional approach given a variational flavour. $\diamondsuit$

We generalize the norm with a *strictly convex functional* f as in

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

where f is what we shall call an *entropy functional*, $f : X \to (-\infty, +\infty]$. Here we suppose f is a strictly convex integral functional[6] of the form

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

The functional f can be used to include other constraints including non-negativity, as we shall see, by appropriate use of $+\infty$.

Example 3 (Various important entropies)**.** Some of the most useful entropies are described below. Full details are in [25, 24], see also [20].

(1) The constrained L^2-norm functional ('positive energy'),

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

is used in constrained *spline fitting* [42], and is implemented in various commercial packages.

(2) Entropy constructions abound: two useful classes follow.

- *Bregman* (based on convex subgradients $\phi(y) - \phi(x) - \phi'(x)(y - x)$); and
- *Csizar distances* (based on $x\phi(y/x)$).

Both model statistical *divergences* extending the *cross-entropy* or *Kullback-Leibler* divergence.

(3) Use of the *Fisher Information*, based on a Csizar distance,

$$f(x, x') := \int_\Omega \frac{x'(t)^2}{2x(t)} \, \mu(dt)$$

[6] This is ensured by the condition that $\phi''(t) > 0$.

which turns out to be *jointly convex* has become more usual as it *penalizes* large derivatives; and can be argued for physically ('hot' over past ten years) [24].

(4) Two popular choices for f are the (negative of) *Boltzmann-Shannon* entropy (in image processing),

$$f(x) := \int x \log x \ (-x) \, d\mu,$$

(changes *dramatically* with μ) and the (negative of) *Burg entropy* (in time series analysis and acoustics),

$$f(x) := -\int \log x \, d\mu.$$

The later includes the *log barrier* and *log det* functions from interior point theory. Both implicitly impose a nonnegativity constraint (positivity in Burg's non-superlinear case), and it is this ability to force the objective to assist in modeling the problem that gives the approach much of its power. $\Diamond$

There has been much information-theoretic debate about which entropy is best.

`This is more theology than science` !

6. What 'Works' Formally

Consider solving $Ax = b$, where, $b \in I\!\!R^n$ and $x \in L^2[0,1]$. Assume further that A is a continuous linear map, hence represented as above. As $L^2[0,1]$ is infinite dimensional, so is the *null space* $N(A)$. That is, if $Ax = b$ is solvable, it is underdetermined.

We pick our solution to *minimize*

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

(or $\phi(x(t), x'(t))$ in Fisher-like cases) [17, 19, 24]. We introduce the *Lagrangian*

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

and the associated *dual problem*

$$\max_{\lambda \in I\!\!R^n} \min_{x \in X} \{L(x,\lambda)\}. \tag{D}$$

So we formally have a "dual pair" (see [16, 52] and Section eight)

$$\min\{f(x) : Ax = b, \quad x \in X\} = \min_{x \in X} \max_{\lambda \in I\!\!R^n} \{L(x,\lambda)\}, \tag{P}$$

and its formal dual (D) above.

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

$$L(\hat{x}, \hat{\lambda}) \leq L(x, \hat{\lambda}),$$

$\forall\, x \in X$. As

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

this implies

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

Thus, we can now reconstruct the primal solution (qualitatively and quantitatively) from a presumptively easier dual computation.

7. An Interlude with George Dantzig

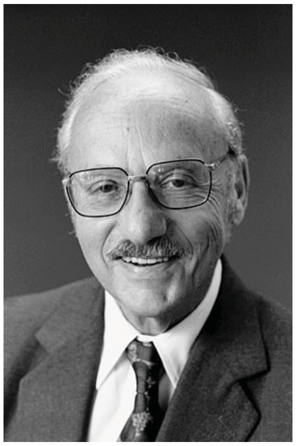


Figure 3. George Danzig (1914–2005).

> *"The term Dual is not new. But surprisingly the term Primal, introduced around 1954, is. It came about this way. W. Orchard-Hays, who is responsible for the first commercial grade L.P. software, said to me at RAND one day around 1954: 'We need a word that stands for the original problem of which this is the dual.' I, in turn, asked my father, Tobias Dantzig, mathematician and author, well known for his books popularizing the history of mathematics. He knew his Greek and*

> *Latin. Whenever I tried to bring up the subject of linear programming, Toby (as he was affectionately known) became bored and yawned. But on this occasion he did give the matter some thought and several days later suggested Primal as the natural antonym since both primal and dual derive from the Latin. It was Toby's one and only contribution to linear programming: his sole contribution unless, of course, you want to count the training he gave me in classical mathematics or his part in my conception."*
>
> A lovely story. I heard George recount this a few times and, when he came to the "conception" part, he always had a twinkle in his eyes. (Saul Gass, 2006)[7]

In a Sept 2006 *SIAM book review* about dictionaries [7], I asserted George assisted his father with his dictionary — for reasons I still believe but cannot reconstruct. This led to Saul Gass's letter above. I also called Lord Chesterfield, Lord Chesterton (*gulp*!). Donald Coxeter used to correct such errors in libraries during lecturing visits.

8. Pitfalls Abound

There are two major problems with this free-wheeling formal approach.

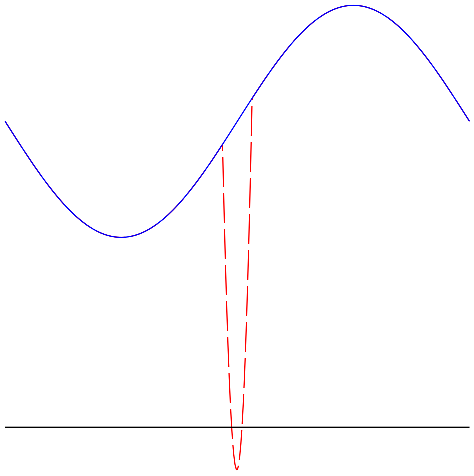


Figure 4. A 'spike' proving the nonnegative cone in L^p with $1 \leq p < \infty$ has empty interior.

[7] Letter from Saul Gass. Dantzig's reminiscence is quoted from [31].

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

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

Above, the optimal value is not attained. While this is a rather contrived example with no strict convexity, we will see that existence can fail for the Burg entropy with simple three-dimensional trigonometric moments. Additional conditions on ϕ are needed to insure solutions exist.[8]

(2) *The assumption that the Lagrangian is differentiable.* In the above problem, f is $+\infty$ for every x negative on a set of positive measure.

Thus, for $1 \leq p < +\infty$ the Lagrangian is $+\infty$ on a dense subset of L^1, the set of functions *not* nonnegative a.e. As Figure 4 illustrates, we can perturb a function in L^p (for $p < \infty$) and change its norm as little as we wish. Hence, the Lagrangian is *nowhere continuous*, much less differentiable.

(3) A third problem, the existence of $\hat{\lambda}$, is less difficult to surmount.

Understanding and fixing the problems. One way to assure continuity or differentiability of f is to work in some $L^\infty(\Omega)$, or $C(\Omega)$, using essentially bounded, or continuous, functions. Even with such unrealistic settings, solutions to (P) *may still not* exist, or may be unnatural. For example, Minerbo [48], posed tomographic reconstruction in $C(\Omega)$, with Shannon entropy. But, his moments are characteristic functions of strips across Ω, and the solution is piecewise constant.

Example 4 (Burg entropy failure of attainment)**.** Consider the following Burg entropy maximization in $L^1[T^3]$, where T is the circle, and variables (x, y, z):

$$\sup \int_{T^3} \log(w)\,dV \text{ subject to } \int_{T^3} w(x,y,z)\,dV \;= 0$$

with side constraints given by

$$\int_{T^3} w\cos(x)\,dV \;=\; \int_{T^3} w\cos(y)\,dV = \int_{T^3} w\cos(z)\,dV = \alpha.$$

For $1 > \alpha > \overline{\alpha}$, solutions only exist in $(L^\infty)^*$; indeed $\overline{\alpha}$ is a computable statistical mechanical quantity [13]. In contrast, for $0 < \alpha < \overline{\alpha}$ the problem attains its infimum in L^1. I challenge anyone to see a physical difference in the two cases. ♢

[8] The solution is actually the *absolutely continuous part of a measure* in $C(\Omega)^*$ [13].

9. Convex Analysis: an Advert

We shall now turn to a theorem that guarantees the form of solution found in the above faulty derivation by ruling out pathology such as in Example 4. That is, we shall legitimate:

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

A full derivation of what follows can be found in [13, 25] and an early summary in [18]. In finite dimensions the underlying convex analysis is described in [16, 24] and Terry Rockafellar's classic 1970 book [52] while [25, 24] describes the modern infinite dimensional theory. A highly recommended addition is Jean-Paul Penot's recent book [49].

Figure 5. Werner Fenchel (1905-1988).

We recall that the (effective) *domain* of a convex function is $\mathrm{dom}(\phi) = \{u : \phi(u) < +\infty\}$ and that ϕ is *proper* if $\mathrm{dom}(\phi) \neq \emptyset$. We may now introduce the *Fenchel (Legendre) conjugate* [16, 20] of an arbitrary function $\phi : I\!R \to (-\infty, +\infty]$:

$$\phi^*(u) = \sup_{v \in I\!R}\{uv - \phi(v)\},$$

and more generally for $f : X \to (-\infty, +\infty]$:

$$f^*(u) = \sup_{v \in X^*}\{\langle u, v\rangle - f(v)\}.$$

This is also called some variation of the Fenchel-Legendre-Moreau-Rockafellar conjugate. We also need the convex *subgradient* defined by

$$\partial f(x) := \{y \in X^* \colon f(z) \geq f(x) + \langle y, z - x\rangle, \forall z \in X\},$$

which is is single-valued if and only if the function is Gâteaux differentiable at x. Also for f convex, proper and closed $f = f^{**}$. The Fenchel conjugate is hiding whenever the convex *subgradient* or *subdifferential* appears since if f is closed, proper and convex then

$$y \in \partial f(x) \Leftrightarrow f(x) + f^*(y) = \langle y, x\rangle. \tag{1}$$

The Fenchel conjugate — prefigured by Legendre's differential equation work and indeed by the 'pole to polar' duality of projective geometry — is at the heart of the modern theory (and practice) of optimization — and of non-linear functional analysis. It plays the role for '+' and 'max' that the Fourier transform plays for product and convolution. Its basic properties are accessible and useful to upper level undergraduates. For example, the *Fenchel-Young inequality*

$$f(x) + f^*(y) \quad \geq \quad \langle x, y\rangle \tag{2}$$

with equality if and only if $y^* \in \partial f(x)$ (see (1) and Figure 9) recaptures the classical Young inequality

$$\frac{1}{p}|x|^p + \frac{1}{q}|y|^q \quad \geq \quad xy$$

for $1/p + 1/q = 1$ and $1 \leq p \leq +\infty$.

Introduced in 1949 by Fenchel (see Figure 5) and refined in the next few decades, especially by Hörmander, Rockafellar and Moreau, the Fenchel conjugate is now a ubiquitous tool for both theoretical and algorithmic reasons. Its power and elegance is currently being exploited in nonsmooth optimization [49, 25], mathematical economics [41, 44], electrical engineering [28], optimization and control theory [30], statistics [50], optimal design [51], variational approaches to PDEs [36] and mechanics [36], statistical mechanics (not surprisingly given the entropy connection), and Banach space renorming theory [38], monotone operator theory [24, 53]; and many other subjects [24].

Yet it is rarely mentioned in undergraduate analysis books (Davidson and Donsig [34] being one pleasant exception) and when during a recent conference presentation I asked sixty numerical optimization specialists, only a handful knew of it. This is in part because it can happily lurk under the hood as we see in Section 10. Indeed, in the late David Gale's lovely short 1967 paper in the *Review of Economic Studies* [41] neither conjugate nor subdifferential is explicitly identified and yet it was the source from which I first clearly understood both. As this and [44] help emphasize, mathematical economists played a large part in the birth of modern convex analysis.

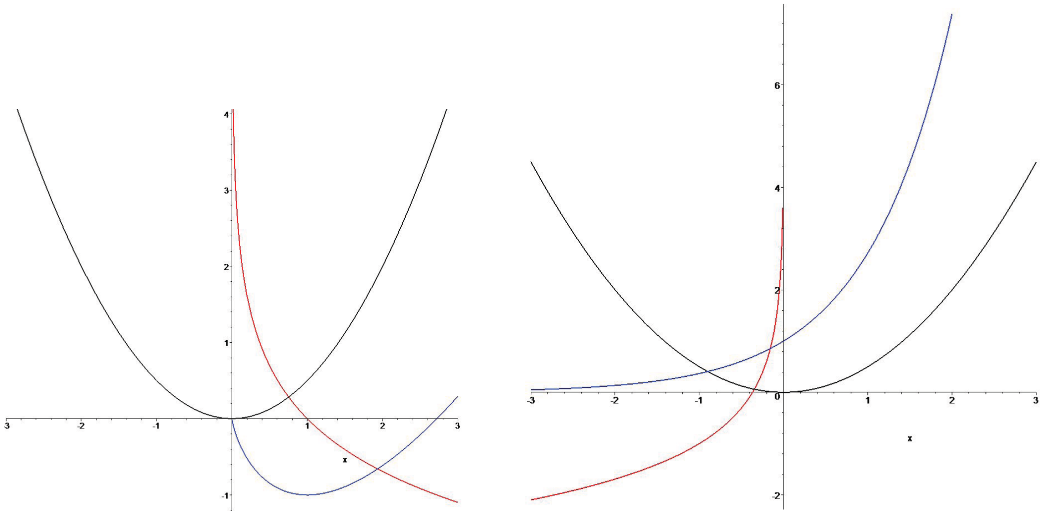
Often the conjugate can be (pre-)computed explicitly — using Newtonian calculus. For instance,

$$\phi(v) = v \log v - v, -\log v \text{ and } \frac{v^2}{2}$$

yield

$$\phi^*(u) = \exp(u), -1 - \log(-u) \text{ and } \frac{u^2}{2}$$

respectively. In the centre is the *log barrier* of interior point fame! The Fisher case is also explicit — via an integro-differential equation of Ricatti type [17, 19].



(a) $\phi(v) := v \log v - v, -\log v, v^2/2$

(b) $\phi^*(u) = \exp(u), -1 - \log(-u), u^2/2$

Figure 6. The three entropies and their conjugates.

Primals and Duals. We sketch our three core entropies and their conjugates in Figure 6. A more subtle conjugate pair is shown in Figure 7.

Example 5 (Conjugates & MRI)**.** 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:

$$\boxed{h(z) := |z| \log\left(|z| + \sqrt{1+|z|^2}\right) - \sqrt{1+|z|^2}}$$

for *quantum theoretic* (Magnetic Resonance Imaging (MRI) or (NMR)) reasons. (Here $|z|$ is the complex modulus.)

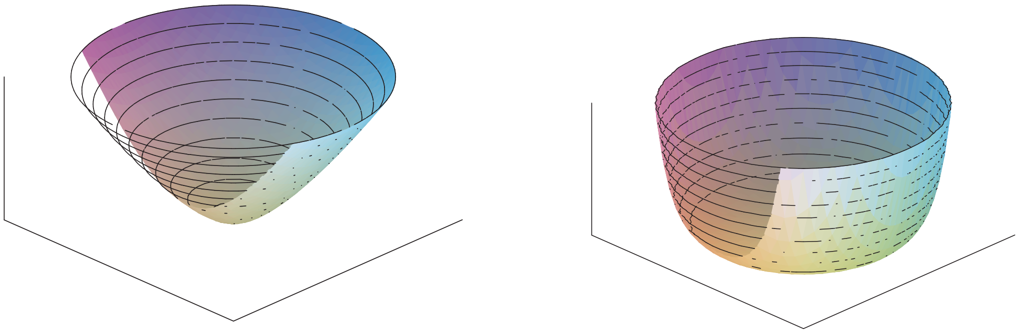
Our *symbolic convex analysis* package (see [10] and Chris Hamilton's Dalhousie thesis software package) produced:

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

Compare the *Shannon entropy*:

$$(|z|\log|z| - |z|)^* = \exp(|z|).$$

Remarkably, the former is smooth (being even) while the later is not. So physical requirements for the primal produces very nice mathematics — and efficient algorithms — in the dual [22]. ♢


(a) Hoch-Stern entropy

(b) cosh entropy

Figure 7. The MRI entropy and its conjugate.

We now turn to the star of this part of our work: the Fenchel duality theorem (1953) [40]. In its modern form this is:

Theorem 2 (Fenchel duality theorem (Utility grade))**.** *Given Banach spaces X and Y, suppose $f\colon X \to R \cup \{+\infty\}$ and $g\colon Y \to R \cup \{+\infty\}$ are convex while $A\colon X \to Y$ is linear and continuous. Then*

$$p := \inf_X f + g \circ A = \max_{Y^*} -g^*(-\cdot) - f^* \circ A^*,$$

if

$$\operatorname{int} A(\operatorname{dom} f) \cap \operatorname{dom} g \neq \emptyset,$$

(or if f, g are polyhedral*).*

We recall another important conjugacy which relates the *indicator function* ($\iota_C(x) :=$ 0 if $x \in C$ and $+\infty$ otherwise) to the *support function* $\sigma_C(x^*) := (\iota_C)^*(x^*) = \sup_{x\in C}\langle x^*, x\rangle$.

Example 6 (Specializations of Fenchel's duality theorem [16]). Three examples in Euclidean space suffice to show the power of Fenchel's result.

(1) The case with $A := I$ is *equivalent* to the analytic form of the Hahn-Banach theorem.

(2) Letting $g := \iota_{\{b\}}$ yields

$$p := \inf\{f(x) \colon Ax = b\}.$$

This specializes to a linear program if $f := \iota_{R_n^+} + c$.

(3) If $f := \iota_C, g := \sigma_D$ yields the von Neumann minimax theorem:

$$\inf_C \sup_D \langle Ax, y \rangle = \sup_D \inf_C \langle Ax, y \rangle.$$

Here C, D are appropriately closed and convex.

These specializations and more may be followed up in [16, 25, 24] or with a more applied bent in [20]. A recent complete application of case (2) with Shannon entropy is given in [12] where it is used to estimate rainfall. $\Diamond$

Using the concave conjugate: $g_* := -(-g)^*(-)$ we get a very fine symmetric formulation of the *Hahn-Banach sandwich theorem*: as shown in Figure 8. Unfortunately, optimizers like minimization while neo-classical economists like maximization so we seem doomed to lots of extra '-' signs.

$$\boxed{\inf_X f(x) - g(x) = \max_{X^*} g_*(x^*) - f^*(x^*)}$$

10. Coercivity and a Proof of Duality

We say ϕ possesses *regular growth* if either $d = \infty$, or $d < \infty$ and $k > 0$, where

$$d := \lim_{u \to \infty} \phi(u)/u, \quad k := \lim_{v \uparrow d} (d - v)(\phi^*)'(v).$$

Regular growth is a technically useful way of forcing the function to grow (that is, of coercivity). Let $\alpha := \inf \operatorname{dom}(\phi)$ and $\beta := \sup \operatorname{dom}(\phi)$.

To ensure dual solutions in function spaces, we need a *constraint qualification*[9], or (CQ), more flexible than Slater's condition because we are using constraint sets with empty interior (see Example 4). Our (CQ) reads:

$$\boxed{\begin{array}{c}\exists\, \overline{x} \in L^1(\Omega), \text{ such that } A\overline{x} = b, \\ f(\overline{x}) \in I\!R, \quad \alpha < \overline{x} < \beta \ \ a.e.\end{array}}$$

[9] Fenchel like other early researchers in nonlinear duality theory *missed* the need for a (CQ) in his 1951 *Princeton Notes*.

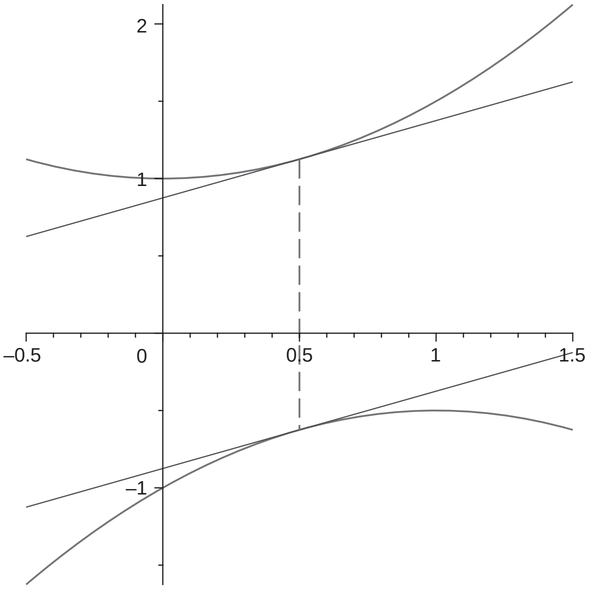


FIGURE 8. $f(x) := 1 + \frac{x^2}{2}$ and $g(x) := -\frac{1}{2} - \frac{(1-x)^2}{2}$. The least gap is at 1/2 with value 7/4.

In many cases, (CQ) *reduces to feasibility* — e.g., for spectral estimation or Hausdorff moment problems we have *pseudo-Haar* sets of moments, and so (CQ) trivially holds if the problem is feasible [14].

The *Fenchel dual problem* for (P) is now:

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

Theorem 3 (Solution of entropy problems [13])**.** *Let* Ω *be a finite interval,* μ *Lebesgue measure, each* a_k *continuously differentiable (or just locally Lipschitz) and* ϕ *proper, strictly convex with regular growth. Suppose* (CQ) *holds and also*[10]

$$\exists\, \tau \in I\!R^n \text{ such that } \sum_{i=1}^n \tau_i a_i(t) < d \quad \forall t \in [a,b], \tag{3}$$

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

$$\hat{x}(t) = \left(\phi^*)'(\sum_{i=1}^n \hat{\lambda}_i a_i(t)\right) \tag{4}$$

[10] This is trivial if $d = \infty$.

where $\hat{\lambda}$ is any solution to dual problem (D) (such $\hat{\lambda}$ must *exist).*

We have now obtained a powerful *functional reconstruction* for all $t \in \Omega$. This generalizes to cover $\Omega \subset I\!R^n$, certain unbounded cases, and more elaborately in Fisher-like cases [13, 17], etc. Indeed 'bogus' differentiation of a discontinuous function becomes the delicate conjugacy formula [24]:

$$\boxed{\left(\int_\Omega \phi\right)^* (x^*) = \int_\Omega \phi^*(x^*).}$$

Thus, the form of the max-ent solution can be legitimated by validating the *easily* checked conditions of Theorem 3.

Also, any solution to $Ax = b$ of the form in (4) is automatically a solution to (P). So solving (P) is equivalent to finding $\lambda \in I\!R^n$ with

$$\text{(5)} \qquad \boxed{\langle a_i, (\phi^*)'(A^T\lambda)\rangle = b_i, \quad i = 1, \ldots, n}$$

which is a *finite dimensional* set of non-linear equations. When $\phi(t) := t^2/2$ this recovers the Gram system of Example 2.

> One can now apply a standard 'industrial strength' nonlinear equation solver, based say on Newton's method, to this system, to find the optimal λ.

Let us emphasize that frequently $\boxed{(\phi')^{-1} = (\phi^*)'.}$ In which case, the 'dubious' formal solution and 'honest' rigorous solution agree. Importantly, we may tailor $(\phi')^{-1}$ to our needs:

- For Shannon entropy, the solution is strictly positive: $(\phi')^{-1} = \exp$.
- For positive energy, we can fit zero intervals: $(\phi')^{-1}(t) = t^+$.
- For Burg, we can locate the support well: $(\phi')^{-1}(t) = 1/t$.

These are excellent methods with relatively few moments (say 5 to 100). For larger problems, stability issues become more vexing. We note that in many medical imaging contexts [29], and other settings such as compressed sensing (as in Figure 12), it is the support that is being sought and it is unrealistic to expect the magnitude to be close.

Note that discretization is only needed to compute terms in evaluation of (5). Indeed, *these integrals can sometimes be computed exactly* (e.g., in some tomography problems [35] and option pricing problem as in Example 7). This is the gain of *not discretizing* early.

> By waiting to see the form of dual, one can customize ones 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 [19] and Wendy Huang's 1993 thesis or [11]).

Example 7 (Option asset pricing [9])**.** For *European call options* the precise problem we address is that of estimating the *density* $p(x)$ of the *price* x of a given *asset* at a set *future expiration time* T. As elsewhere our paradigm legitimates faulty arguments in the finance literature. Let b_i be the price of a *risk neutral European call option* for the given asset with *strike prices* $k_i, i = 1, 2, ..., m$. Following [1] and others, the problem can be formulated as a Shannon entropy optimization problem in exactly the form we have discussed:

$$\inf_X \left\{ \int_I p(x) \log p(x) - p(x)\, dt : \int_I x(t)\, dt = 1, \int_I a_i(x) p(x)\, dx = b_i \right\},$$

where the constraints — representing the price paid at a strike price — are 'hockey-sticks' of the form:

$$a_i(x) := \max\{0, x - k_i\},$$

while $I = [0, K]$ is an interval known to contain the range of x, which may just be I=$[0, \infty]$. In this case the dual can be computed *exactly* and leads to a relatively small and explicit dual problem or nonlinear equation to solve the problem:

$$\max_{\lambda \in \mathbb{R}^{m+1}} \left\{ \lambda_0 + \sum_{i=1}^m \lambda_i d_i - e^{\lambda_0} \sum_{j=1}^m e^{-\nu_j} \frac{\exp(k_{j+1}\mu_j) - \exp(k_j \mu_j)}{\mu_j} \right\},$$

in which $k_{m+1} := K, \nu_j := \sum_{i=1}^j \lambda_i k_i, \mu_j := \sum_{i=1}^j \lambda_i$. The primal is now easily reconstructed as above. For $K = \infty$ we obtain the primal optimal density as:

$$p(x) = e^{-\lambda_0 - \sum_{i=1}^m \lambda_i (x - k_i)^+}$$

where λ is the optimal dual value. Note that we can write

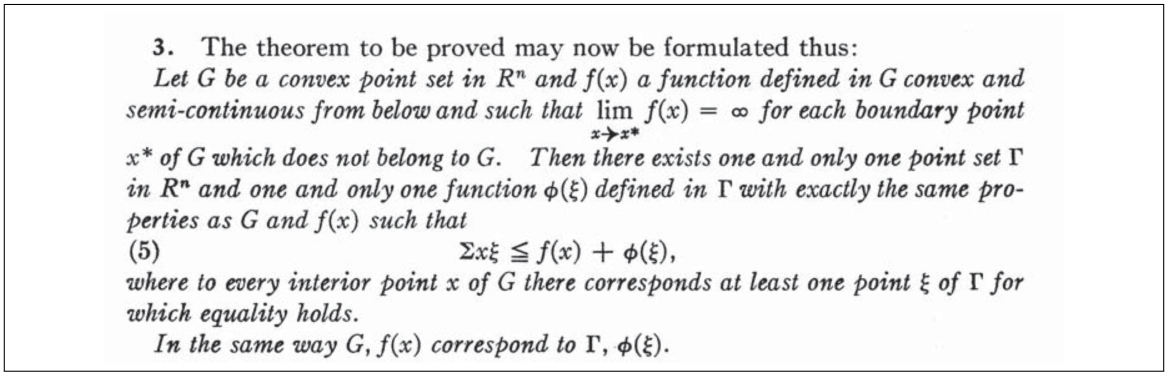
$$-\lambda_0 = \log \int_0^\infty \exp(\sum_{i=1}^m \lambda_i (x - k_i)^+)\, dx.$$

We have implemented this using realistic financial data and have been very pleased with the results [9] . These results have been taken quite a bit further by [27] for financially interesting boundary cases not covered by our basic theory, since we require strict convexity of the data. $\Diamond$

I might add that this example illustrates that the more nonlinear the optimization problem the more *wasteful, dangerous and misleading* it is to treat it purely formally or to discretise early.

11. From Fenchel's Acorn ...

The first publication of the Fenchel-Young inequality (2) was in the fist issue of the *Canadian Journal of Mathematics* [39]. The main result therein is reproduced in Figure 9. The first issue of the Canadian Journal was extraordinary. As the reproduction in (see Figure 10) of the title page — from the web version — makes clear, many great mathematicians wanted to help the launch.



3. The theorem to be proved may now be formulated thus:

Let G be a convex point set in R^n and $f(x)$ a function defined in G convex and semi-continuous from below and such that $\lim_{x \to x^*} f(x) = \infty$ for each boundary point x^* of G which does not belong to G. Then there exists one and only one point set Γ in R^n and one and only one function $\phi(\xi)$ defined in Γ with exactly the same properties as G and $f(x)$ such that

$$\Sigma x\xi \leqq f(x) + \phi(\xi), \tag{5}$$

where to every interior point x of G there corresponds at least one point ξ of Γ for which equality holds.

In the same way G, $f(x)$ correspond to Γ, $\phi(\xi)$.

Figure 9. Fenchel describing $\phi = f^*$ in 1949.

... **a modern oak emerges.** Theorem 3 works by relaxing the problem to $(L^1)^{**}$ — where solutions always exist — and using the Lebesgue decomposition theorem. Regular growth rules out a non-trivial singular part to the relaxed solution via analysis with the formula:

$$I_{\phi^{**}} = (I_\phi)^{**}|_X.$$

More generally, for Ω an interval, we can work with

$$I_\phi(x) := \int_\Omega \phi(x)\, d\mu$$

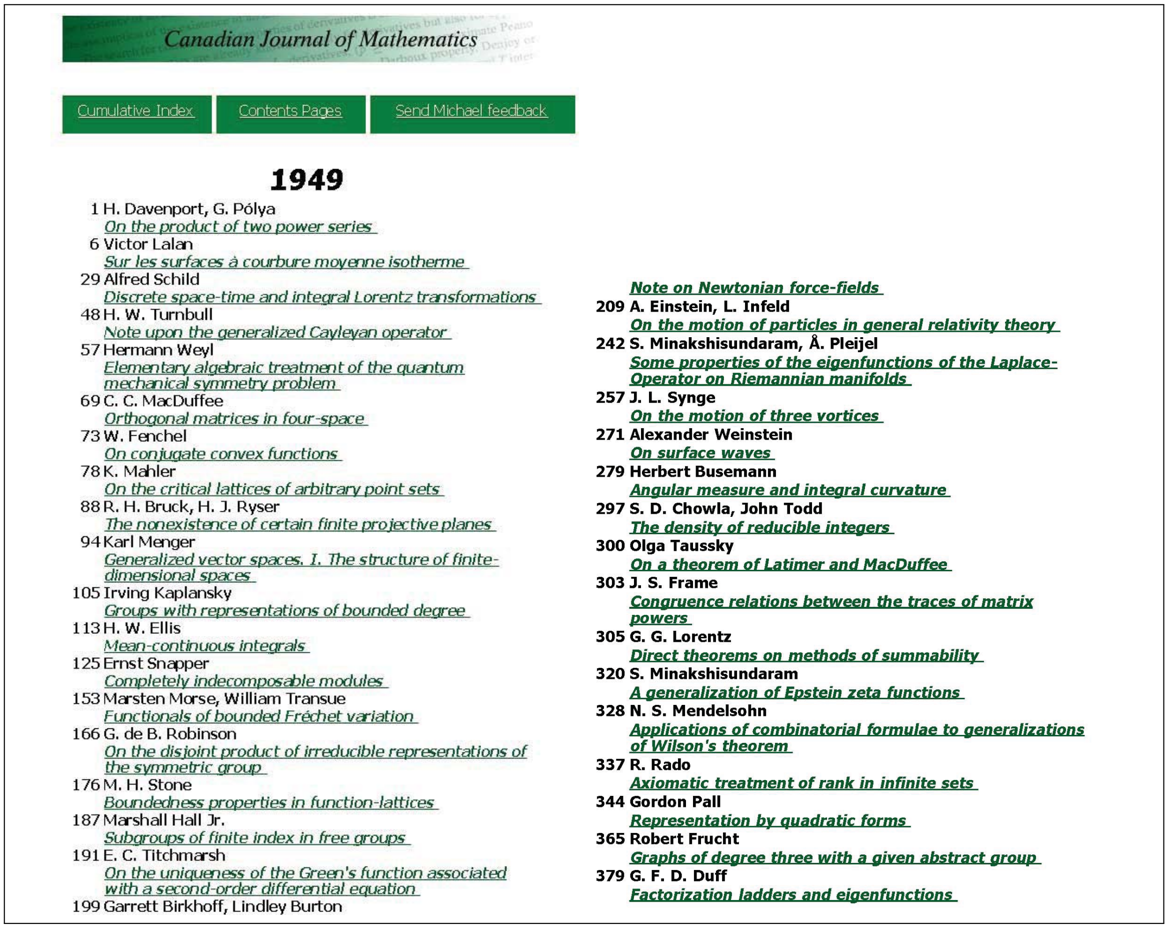
as a function on $L^1(\Omega)$.

Buried in Theorem 3 is the following notion, more-or-less dual to regular growth, of strong rotundity.

Strong rotundity. As in [15, 24], we say I_ϕ is *strongly rotund* (*very well posed*) if it is (i) strictly convex, with (ii) weakly compact lower level sets (the *Dunford-Pettis criterion* for weak compactness in L^1), and satisfies (iii) the *Kadec-Klee condition*:

$$\boxed{I_\phi(x_n) \to I_\phi(x), x_n \rightharpoonup x \Rightarrow x_n \to_1 x.}$$

Theorem 4 (Strong Rotundity Characterization [15])**.** *I_ϕ is strongly rotund as soon as ϕ^* is everywhere finite and differentiable on $\mathbb{R}$; and conversely if μ is not purely atomic.*



1949

1 H. Davenport, G. Pólya
On the product of two power series
6 Victor Lalan
Sur les surfaces à courbure moyenne isotherme
29 Alfred Schild
Discrete space-time and integral Lorentz transformations
48 H. W. Turnbull
Note upon the generalized Cayleyan operator
57 Hermann Weyl
Elementary algebraic treatment of the quantum mechanical symmetry problem
69 C. C. MacDuffee
Orthogonal matrices in four-space
73 W. Fenchel
On conjugate convex functions
78 K. Mahler
On the critical lattices of arbitrary point sets
88 R. H. Bruck, H. J. Ryser
The nonexistence of certain finite projective planes
94 Karl Menger
Generalized vector spaces. I. The structure of finite-dimensional spaces
105 Irving Kaplansky
Groups with representations of bounded degree
113 H. W. Ellis
Mean-continuous integrals
125 Ernst Snapper
Completely indecomposable modules
153 Marsten Morse, William Transue
Functionals of bounded Fréchet variation
166 G. de B. Robinson
On the disjoint product of irreducible representations of the symmetric group
176 M. H. Stone
Boundedness properties in function-lattices
187 Marshall Hall Jr.
Subgroups of finite index in free groups
191 E. C. Titchmarsh
On the uniqueness of the Green's function associated with a second-order differential equation
199 Garrett Birkhoff, Lindley Burton
Note on Newtonian force-fields
209 A. Einstein, L. Infeld
On the motion of particles in general relativity theory
242 S. Minakshisundaram, Å. Pleijel
Some properties of the eigenfunctions of the Laplace-Operator on Riemannian manifolds
257 J. L. Synge
On the motion of three vortices
271 Alexander Weinstein
On surface waves
279 Herbert Busemann
Angular measure and integral curvature
297 S. D. Chowla, John Todd
The density of reducible integers
300 Olga Taussky
On a theorem of Latimer and MacDuffee
303 J. S. Frame
Congruence relations between the traces of matrix powers
305 G. G. Lorentz
Direct theorems on methods of summability
320 S. Minakshisundaram
A generalization of Epstein zeta functions
328 N. S. Mendelsohn
Applications of combinatorial formulae to generalizations of Wilson's theorem
337 R. Rado
Axiomatic treatment of rank in infinite sets
344 Gordon Pall
Representation by quadratic forms
365 Robert Frucht
Graphs of degree three with a given abstract group
379 G. F. D. Duff
Factorization ladders and eigenfunctions

FIGURE 10. Contents of issue 1 of the first volume of Can. J. Math. in 1949.

What is quite fine about this result is that it allows for sophisticated functional analytic tools to be applied by practitioners who are not at home with their corpus. The requirement is only to check the global smoothness of ϕ^*. Thence we see that the Shannon entropy, the energy and positive energy yield a strongly rotund I_ϕ while the strictly convex Burg entropy does not. Correspondingly, the *Fermi-Dirac entropy* [21, 16, 24] sets

$$\phi(x) := x\log(x) + (1-x)\log(1-x),$$

which has dual

$$\phi^*(y) = \log(1-\exp(y))$$

and so is also strongly rotund.

This is excellent news, since we know that strongly rotund entropy optimization problems always attain their infima and are well-posed. The use of strong rotundity to establish the norm-convergence of moment estimates — as the number of moments increases — is also detailed in [15]. Moreover, as described in [15] a strongly rotund I_ϕ is by far the best *surrogate* for the properties of a reflexive norm on the non-reflexive Banach space L^1.

12. A Graphical Comparison

An old (circa 1997) interface: Moment+ (`www.cecm.sfu.ca/interfaces/`) provided code for entropic reconstructions as above. Moments (including wavelets), entropies and dimension are easily varied. It also allows for adding noise and relaxation of the constraints. Several methods of solving the dual are possible, including *Newton and quasi-Newton methods (BFGS, DFP), conjugate gradients, and the suddenly sexy Barzilai-Borwein line-search free method* [2].

Comparison of our three basic entropies. In Figure 11, we compare the positive L^2, Boltzmann-Shannon and Burg entropy reconstruction of the *characteristic function* $\chi(1/2, t)$ of $[0, 1/2]$ using ***10*** *algebraic moments* $b_i = \int_0^{1/2} t^{i-1}\, dt$ on $\Omega = [0, 1]$.

In each case the solution is

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

and so is a function of a genuine algebraic polynomial.

Note that Burg over-oscillates since $(\phi^*)'(t) = 1/t$. But is still often the 'best' solution as it finds the support of the signal well (and there is a closed form for Fourier moments!). We emphasize that the optimization problem we solved does not ask for the "best picture". For instance, a relaxation to $\|Ax - b\|_1 \le \varepsilon$ with $\varepsilon = .1^{10}$ will produce a solution that looks noting like $\chi(1/2, t)$.

13. Part II: The Non-convex Case

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

Example 8 (Molecular crystallography [32])**.** We wish to estimate x in $L^2(I\!R^n)$[11] and can suppose the modulus $c = |\hat{x}|$ is known (here $\hat{x}$ is the Fourier transform of x).[12]

[11] Here $n = 2$ for images, 3 for holographic imaging, etc.

[12] Observation of the modulus of the *diffracted image in crystallography*. Similarly, for *optical aberration correction.*

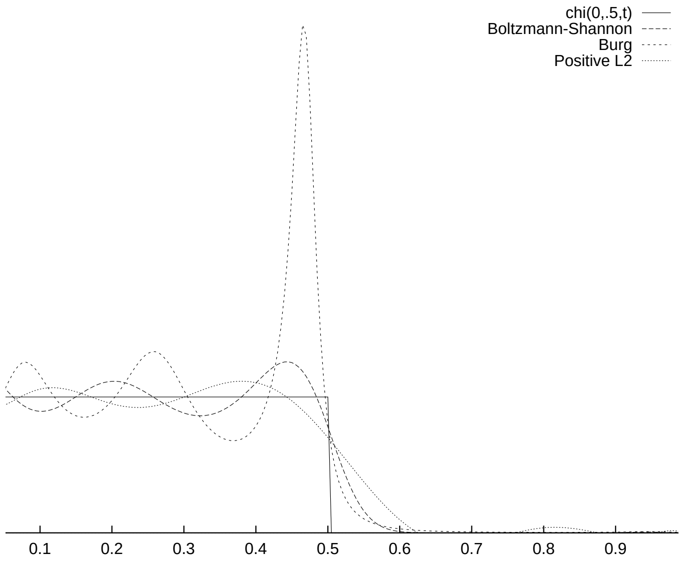


FIGURE 11. Comparison of three core entropies.

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

$$\min\{f(x) : Ax = b, \|Mx\| = c, \quad x \in X\}, \tag{NP}$$

where M models the modular constraint, and f is as in Theorem 3.

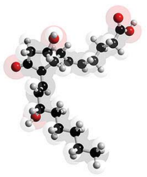
Most optimization methods rely on a *two-stage* (easy convex, hard non-convex) decoupling schema — the following is from Decarreau-Hilhorst-LeMaréchal-Navaza [32].

Now DHLN suggest solving

$$\min\{f(x) : Ax = y, \|B_k y\| = b_k, (k \in K) \ \ x \in X\}, \tag{NP^*}$$

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

They solve formal *first-order Kuhn-Tucker conditions* for a relaxed form of (NP^*). The easy constraints are treated by Thm. 3. I am obscure, mainly because the results were largely negative:



The authors applied these ideas to a prostaglandin molecule (25 atoms), with known structure, using quasi-Newton (which could fail to find a local min), truncated Newton (better) and trust-region (best) numerical schemes.

Ultimately DHL 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.*"

We had similar experiences with attempts at non-convex medical image reconstruction. But reporting negative results has been under-valued in optimization and in mathematics more generally. ◊

> *Another thing I must point out is that you cannot prove a vague theory wrong. ... Also, if the process of computing the consequences is indefinite, then with a little skill any experimental result can be made to look like the expected consequences.* Richard Feynman (1964 Nobel speech)

14. General Phase Reconstruction and Inversion

The basic setup [20] is as follows. (See also Examples 9 and 10 below.) We are given an *electromagnetic field:* $u : \mathbb{R}^2 \to \mathbb{C} \in L^2$ and *data:* these are *field intensities* for $m = 1, 2, \ldots, M$:

$$\psi_m : \mathbb{R}^2 \to \mathbb{R}_+ \in L^1 \cap L^2 \cap L^\infty.$$

We are also in possession of functions $\mathcal{F}_m : L^2 \to L^2$ (*modified Fourier Transforms*) for which we can measure the modulus (signal intensity)

$$|\mathcal{F}_m(u)| = \psi_m \quad \forall m = 1, 2, \ldots, M.$$

General Physical Inverse Problem. This now becomes:

> Given transforms $\mathcal{F}_m$ and *measured* field intensities ψ_m (for $m = 1, \ldots, M$), find a *robust estimate* of the underlying field function u satisfying
> $$|\mathcal{F}_m(u)| = \psi_m \quad \forall m = 1, 2, \ldots, M.$$

Example 9 (Some hope from Hubble)**.** The (human-ground) lens was mis-assembled by 1.33mm. The perfect back-up (computer -ground) lens stayed on earth!

NASA challenged ten teams to devise algorithmic fixes until the Hubble observatory (see Figure 13) could be visited and physically repaired. The winner was *Optical aberration correction*, using the *Misell algorithm*, a *method of alternating projections*, which works much better than it should — given that it is being applied to the following M-set *non-convex feasibility problem* examined more fully below:

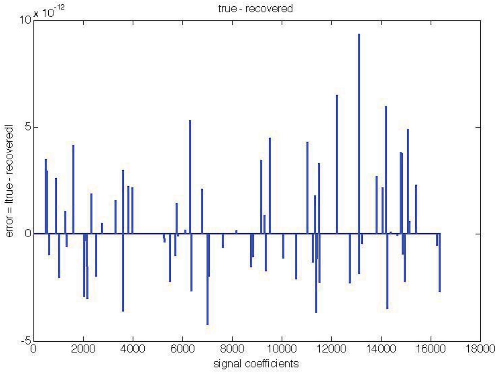


FIGURE 12. Compressed sensing reconstruction [21].

FIGURE 13. NASA's Hubble telescope.

PROBLEM. Find a member of a specification of

$$\Psi := \bigcap_{k=1}^{M} \{x : Ax = b, \|M_k x\| = c_k, \quad x \in X\}. \qquad (NCFP)$$

The Misell solution was so successful that such methods are now routinely applied to improve image quality (just as one might always add a Newton step to improve an algebraic numerical solution). It is worth asking whether there is hidden convexity structure to explain unanticipated such good behaviour?

Hubble has since been reborn twice and *exoplanet* discoveries have become quotidian. There were **228** listed at www.exoplanets.org in March 2009 and **432** a year later, **563** as of 22/6/11. Many more remain to be confirmed according to the new Kepler search http://kepler.nasa.gov/. Though, one might wonder how mathematically reliable are these determinations (of velocity, imaging, transiting, timing, micro-lensing, etc.)? ♢

15. Inverse Problems: Two Reconstruction Approaches

The two reconstruction approaches we wish to compare now are:

I. Error reduction of a *nonsmooth objective* (a freely chosen 'entropy' in the sense of Part I): In this case we attempt to solve for fixed $\beta_m > 0$

$$\text{minimize} \quad E(u) := \sum_{m=0}^{M} \frac{\beta_m}{2} \text{dist}^2(u, \mathbb{Q}_m) \quad \text{over } u \in L^2.$$

Needless-to-say, many variations on this theme are possible. Alternatively, we consider:

II. Solution of a *non-convex feasibility problem*: Given $\psi_m \neq 0$, let $\mathbb{Q}_0 \subset L^2$ be a convex set , and for $1 \leq m \leq M$

$$\mathbb{Q}_m := \left\{ u \in L^2 \colon \; |\mathcal{F}_m(u)| = \psi_m \; a.e. \right\} \quad \text{(nonconvex)}$$

we wish to find $u \in \bigcap_{m=0}^{M} \mathbb{Q}_m = \emptyset$.

This is often well solved via an *alternating projection method* or some variant: e.g., for two sets A and B, *repeatedly compute*

$$\boxed{x \mapsto P_B(x) =: y \mapsto P_A(y) =: x.}$$

For highly nonlinear and very large problems, it is unlikely that direct optimization methods will be implementable. It is in this arena particularly that iterative projection methods come to the fore. We return to such methods in the next section.

Example 10 (Inverse scattering [20])**.** We wish to determine the location and shape of buried objects (treasure or trash) from measurements of the *scattered field* after illuminating a region which has a known *incident field.* An example is shown in Figure 14.

Recent techniques determine if a point z is inside or outside of the scatterer by determining *solvability* of the linear integral equation:

$$\boxed{\mathcal{F} g_z \overset{?}{=} \varphi_z}$$

where $\mathcal{F} \colon X \to X$ is a compact linear operator constructed from the observed data, and $\varphi_z \in X$ is a known function parameterized by z [20]. Interestingly, $\mathcal{F}$ being a compact operator has *dense range*, but if z is on the exterior of the scatterer, then $\varphi_z \notin \text{Range}(\mathcal{F})$ (which has a Fenchel conjugate characterization).

Since $\mathcal{F}$ is compact, any numerical implementation to solve the above integral equation will need some *regularization scheme.* If *Tikhonov regularization* is used — in a restricted physical setting — the solution to the regularized integral equation, $g_{z,\alpha}$, has the behaviour $\|g_{z,\alpha}\| \to \infty$ as $\alpha \to 0$ *if and only if* z is a point outside the scatterer. An important open problem [20, 8] is to determine the behaviour of

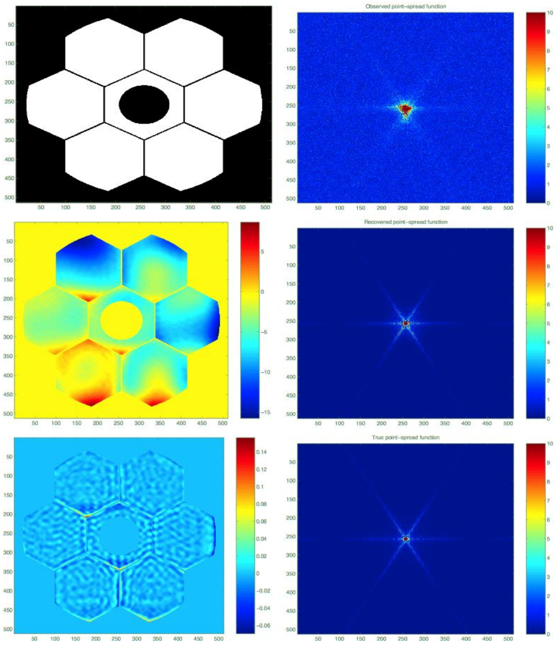


Figure 14. Reconstruction (via I). Top row: the data. Middle: reconstruction. Bottom: truth and error.

regularized solutions $g_{z,\alpha}$ under different regularization strategies. In other words, when can these techniques fail? $\Diamond$

To conclude this article we discuss further alternating projection algorithms (AP) [4, 5], and their extensions [5].

16. Alternating Projections and Reflections

> *A heavy warning used to be given [by lecturers] that pictures are not rigorous; this has never had its bluff called and has permanently frightened its victims into playing for safety. Some pictures, of course, are not rigorous, but I should say most are (and I use them whenever possible myself).* J.E. Littlewood (1885-1977)

The *alternating projection method* — discovered by many including Schwarz, Wiener, Von Neumann (as accessible good ideas often are) — is *fairly* well understood when all sets are convex. (See Figure 15.)

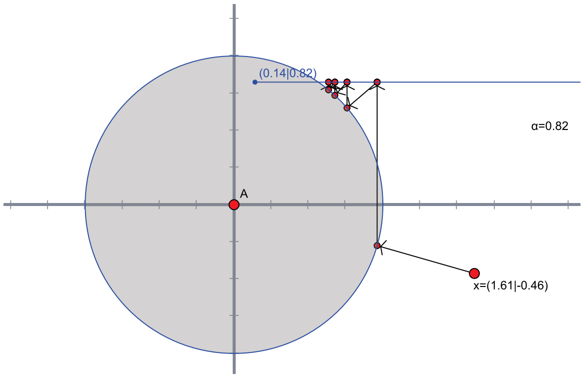


FIGURE 15. (AP) 'zig-zagging' to a point on the intersection of a sphere and line segment.

Theorem 5 (Alternating projections)**.** *If A, B are closed and convex subsets of Hilbert space and $A \cap B \neq \emptyset$ then for any $x_0 := x$*

$$x_n \mapsto P_B(x_n) =: y_n \mapsto P_A(y_n) =: x_{n+1}$$

converges weakly to a point in the intersection of A and B.

Norm convergence was shown by von Neumann (1933) for two subspaces, and the general result is due to Bregman (1965) [4, 5, 24]. It was only shown recently by Hundal (2002) that norm convergence can actually fail — but this phenomenon is only shown for an ingenious 'artificial' example with a hyperplane and an odd cone. This suggests the following.

Conjecture 1 (Norm convergence of realistic alternating projections models)**.** *If A has finite codimension, closed and affine (a translate of a vector subspace) and B is the nonnegative cone in $\ell^2(\mathbb{N})$, while $A \cap B \neq \emptyset$, then the alternating projection method is norm convergent.*

This conjecture, which covers a great many concrete cases, is known to be true when A is a closed hyperplane [3] but for codimension two it remains open.

Non-convex alternating projections methods can fail. Consider the alternating projection method to find the unique red point on the line-segment A (convex) and the circle B (non-convex). The method is 'myopic' in that it works locally. Consider what happens if B is replaced by its convex hull the sphere.

As suggested by Figure 16, starting on the line-segment outside the smaller circle, we converge to unique feasible solution. Starting inside that circle leads to a period-two locally 'least-distance' solution. It is possible to exhibit much more interesting periodic behaviour by perturbing the two cycle.

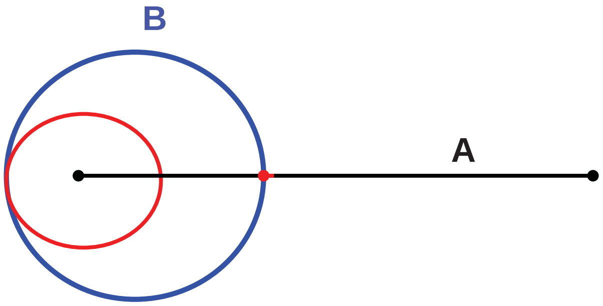


FIGURE 16. Projection method for circle B and segment A fails inside smaller circle.

This example and similar ones leads us to focus on the case where the convex set Q_0 is actually affine. This is certainly the case when $Q_0 := \{x\colon Lx = b\}$ for a finite rank bounded linear mapping L as defining the set A in Conjecture 1.

17. THE PROJECTION METHOD OF CHOICE

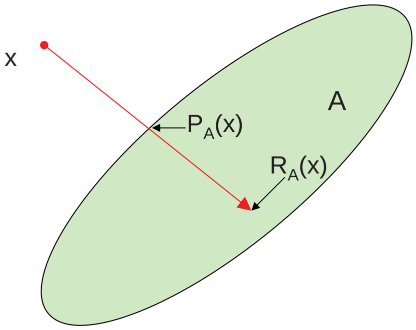


FIGURE 17. Projector $P_A(x)$ and reflector $R_A(x)$ for point x.

For optical abberation correction this is the classical *alternating projection* method:

$$x \mapsto P_A\left(P_B(x)\right).$$

For crystallography it is better to use (HIO) *over-relax and average*: *reflect* to $R_A(x) := 2\,P_A(x) - x$ and use

$$x \mapsto \frac{x + R_A\left(R_B(x)\right)}{2}$$

with the reflection illustrated in Figure 17.

Both reflection and projection on a closed convex set in Hilbert space are *non-expansive* mappings and it is this geometric property that ensures convergence of algorithms in the convex case.

Example 11 (Parallelization [4, 5].)**.** Both reflection and projection parallelize neatly to handle M sets in X by using

$$B := Q_1 \otimes Q_2 \otimes \cdots \otimes Q_M, \quad A := \{(x_1, x_2, \ldots, x_M) \colon x_1 = x_2 \cdots = x_m \in X\}.$$

Note that

$$R_B(x) = \prod_{i=1}^{M} R_{Q_i}(x_i), \quad R_A(x) = \prod_{i=1}^{M} \frac{x_1 + x_2 + \cdots x_M}{M}.$$

Put concisely the 'divide and conquer' algorithm we obtain from Douglas-Rachford is $x^0 := x \in X$ and

$$y^n = \frac{1}{M} \sum_{i=1}^{M} R_{Q_i}(x^n), \quad x^{n+1} = \frac{y^n + x^n}{2}. \tag{6}$$

Averaged projections is correspondingly

$$x^{n+1} = \frac{1}{M} \sum_{i=1}^{M} P_{Q_i}(x^n). \tag{7}$$

Of course many variants are possible and some are very useful [4, 5].

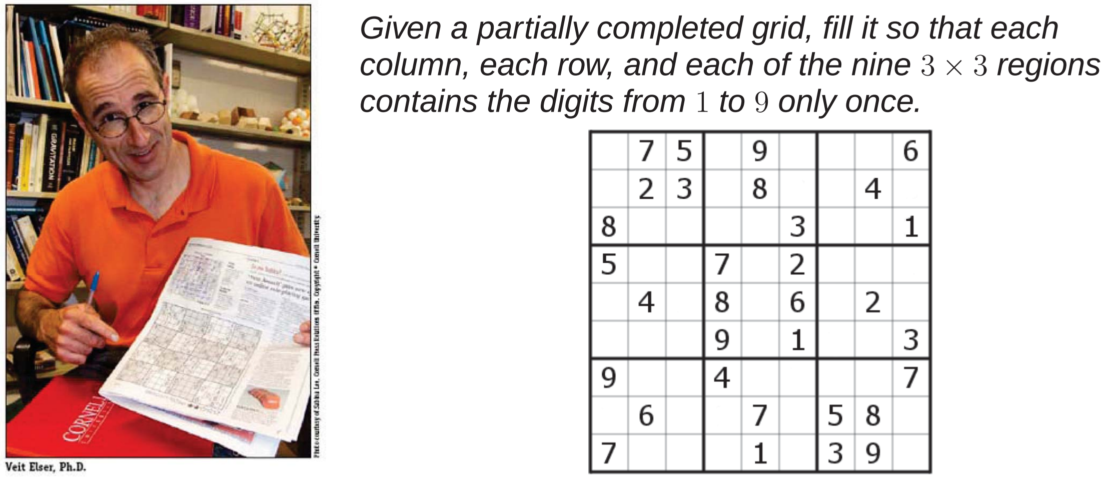
Thus, if many processors are available and the complexity of computing R_{Q_i} (P_{Q_i}) is fairly uniform the algorithm is well suited to a loosely couple cluster where a 'head-node' distributes the current estimate to the 'salve-nodes' which compute and return their assigned reflections (projections). Observe also that while B is non convex A is definitely affine. $\diamondsuit$

Both reflection and projection need new theory developed to understand the the non-convex case.

Names change when fields do. . . The optics community calls projection algorithms "*Iterative Transform Algorithms*". Hubble used *Misell's Algorithm*, which is just averaged projections. The best projection algorithm Russell Luke[13] found was *cyclic projections* (with no relaxation).

For the crystallography problem the best known method is called the *Hybrid Input-Output algorithm* in the optical setting. Bauschke-Combettes-Luke [6] showed HIO, *Lions-Mercier* (1979), *Douglas-Rachford* (1959), *Fienup* (1982), and *divide-and-concur* coincide. When $u(t) \geq 0$ is imposed, Fienup's method no longer coincides, and DR ('HPR') is still better.

[13] My former PDF, he was a *Hubble Graduate student*.


Veit Elser, Ph.D.
Photo courtesy of Sabina Lee, Cornell Press Relations Office. Copyright © Cornell University

Given a partially completed grid, fill it so that each column, each row, and each of the nine 3×3 regions contains the digits from 1 to 9 only once.

	7	5		9				6
	2	3		8			4	
8					3			1
5			7		2			
	4		8		6		2	
			9		1			3
9			4					7
	6			7		5	8	
7				1		3	9	

(a) Elser and a large puzzle

(b) Sudoku rules

FIGURE 18. Veit Elser and Sudoku.

Elser and Sudoku, Bauschke and Queens. Since **2006** Veit Elser [37, 43] at Cornell has had huge success (and good press) using *'divide-and-concur'* on very hard combinatorial optimization problems such as protein folding, sphere-packing, 3SAT, for Sudoku (posed in $\mathbb{R}^{2916}$), and more as in Figure 18. In **2008** Bauschke and Schaad likewise studied the classical *eight queens* chess problem (posed in $\mathbb{R}^{256}$) and in image-retrieval, both covered in *Science News* (2008) (See Figure 19.)

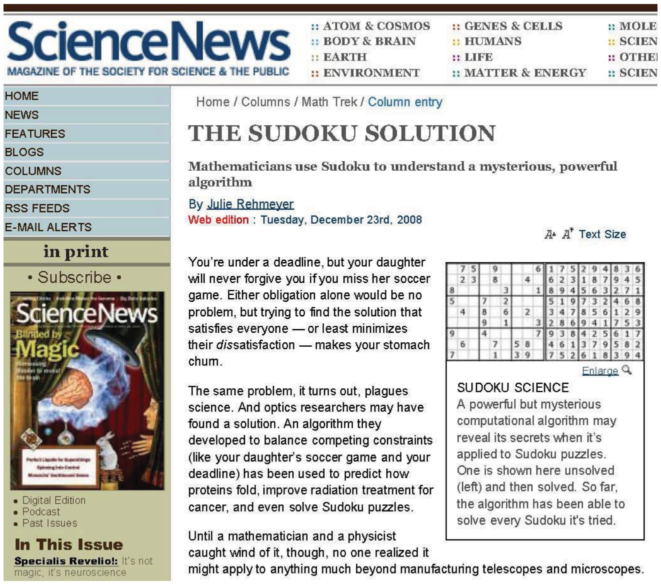
This success (are we observing convergence a.e.?) is not seen with alternating projections and cries out for explanation. Brailey Sims and I [23] have made some theoretical progress as we now indicate.

18. FINIS: DOUGLAS-RACHFORD ON THE SPHERE

The Douglas-Rachford iteration (DR) originated a half century ago as a heuristic algorithm to solve the heat equation [23]. Dynamics are already fascinating for B the unit circle and A the blue line at height $\alpha \geq 0$. Steps are determined by $x_{n+1} = T(x_n)$ where

$$T := \frac{I + R_A \circ R_B}{2}.$$

This can be thought of as the simplest realistic non-convex phase-reconstruction problem (and note that the set B is affine).



THE SUDOKU SOLUTION

Mathematicians use Sudoku to understand a mysterious, powerful algorithm

By Julie Rehmeyer
Web edition : Tuesday, December 23rd, 2008

You're under a deadline, but your daughter will never forgive you if you miss her soccer game. Either obligation alone would be no problem, but trying to find the solution that satisfies everyone — or least minimizes their *dis*satisfaction — makes your stomach churn.

The same problem, it turns out, plagues science. And optics researchers may have found a solution. An algorithm they developed to balance competing constraints (like your daughter's soccer game and your deadline) has been used to predict how proteins fold, improve radiation treatment for cancer, and even solve Sudoku puzzles.

Until a mathematician and a physicist caught wind of it, though, no one realized it might apply to anything much beyond manufacturing telescopes and microscopes.

	7	5		9				6
	2	3		8			4	
8					3			1
5			7		2			
	4		8		6		2	
			9		1			3
9			4					7
	6			7		5	8	
7				1		3	9	

1	7	5	2	9	4	8	3	6
6	2	3	1	8	7	9	4	5
8	9	4	5	6	3	2	7	1
5	1	9	7	3	2	4	6	8
3	4	7	8	5	6	1	2	9
2	8	6	9	4	1	7	5	3
9	3	8	4	2	5	6	1	7
4	6	1	3	7	9	5	8	2
7	5	2	6	1	8	3	9	4

SUDOKU SCIENCE
A powerful but mysterious computational algorithm may reveal its secrets when it's applied to Sudoku puzzles. One is shown here unsolved (left) and then solved. So far, the algorithm has been able to solve every Sudoku it's tried.

FIGURE 19. Bauschke and Schaad's work.

With θ_n denoting the argument this becomes the iteration

$$x_{n+1} := \cos\theta_n \tag{8}$$

$$y_{n+1} := y_n + \alpha - \sin\theta_n \tag{9}$$

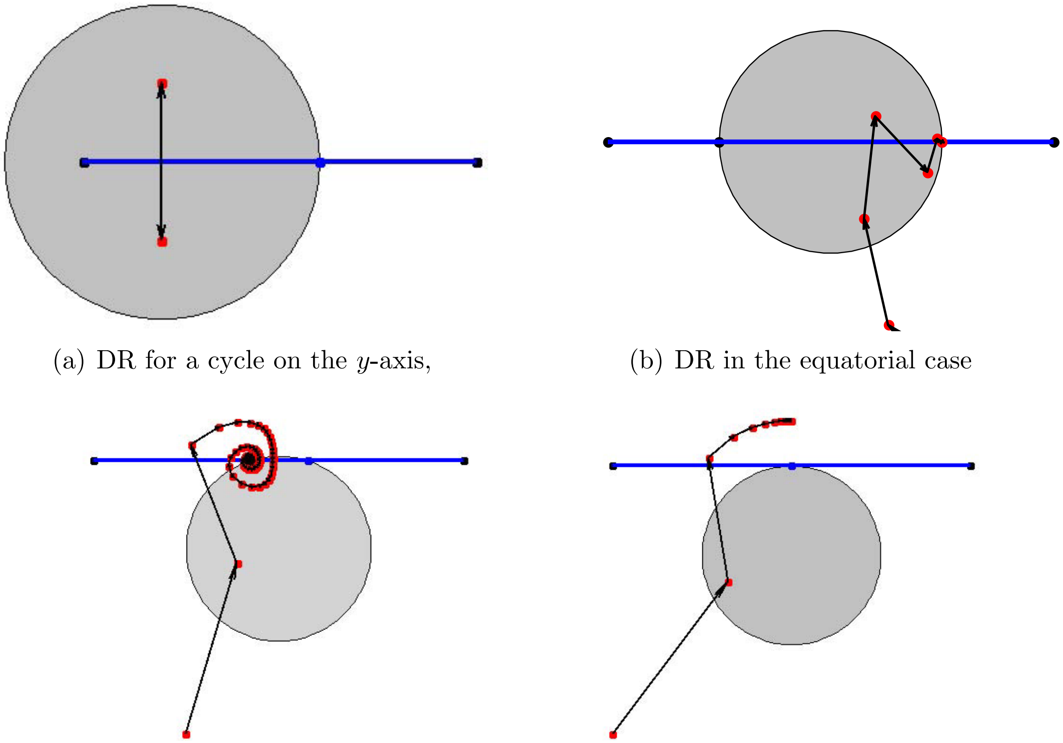
and we have convergence results **iff** we start off y-axis where 'chaos' provably rules. (See Figure 20(a) which shows a period two point on the y-axis. Figure 20(b) shows convergence in the case of a line through diametral points.)

For $0 < \alpha < 1$ we can prove we converge locally exponentially asymptotically, but we are sure that the result is true globally. For $\alpha > 1$ it is easy to show that $y_n \to \infty$, while for $\alpha = 0.95$ $(0 < \alpha < 1)$ and $\alpha = 1$ respectively we arrive at behaviour shown in the pictures of Figure 20(c) and Figure 20(d). These convergence results are explained in detail in [23]. They remain valid for a sphere and any affine manifold in Euclidean space [23].

Interactive geometry. Many of these Douglas-Rachford results were discovered by analyzing orbits of the iteration as dynamic and interactive objects. HTML versions of two *Cinderella* applets are available at:

(1) `www.carma.newcastle.edu.au/~jb616/composite.html`

(2) `www.carma.newcastle.edu.au/~jb616/expansion.html`



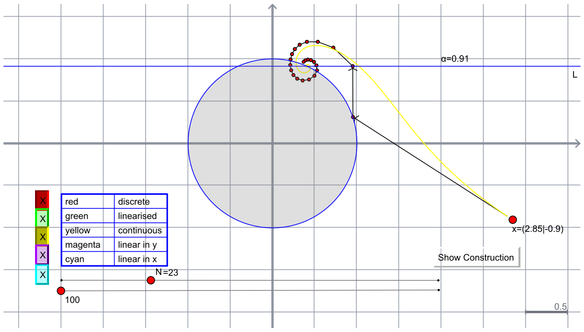
(a) DR for a cycle on the y-axis,

(b) DR in the equatorial case

(c) DR for a typical secant case

(d) DR in the tangent case

FIGURE 20. Douglas-Rachford cases.



X	red	discrete
X	green	linearised
X	yellow	continuous
X	magenta	linear in y
X	cyan	linear in x

FIGURE 21. A dynamic *Cinderella* applet studying trajectories of DR and variants.

The underlying *Cinderella* applets were built with Chris Maitland and their features are detailed in [23]. The first is illustrated in Figure 21 and the second in Figure 22. We invite the reader to play with them.

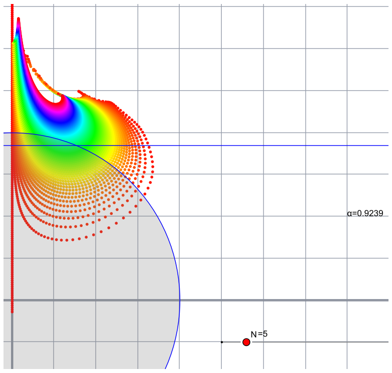


FIGURE 22. Applet studying thousands of trajectories simultaneously. Starting points are coloured by distance from the y-axis.

Acknowledgements. Thanks are due to many but especially to Adrian Lewis, Heinz Bauschke, Russell Luke and Brailey Sims who have been close collaborators on matters relating to this work over many years. Thanks are also due to Francisco Aragon for a careful reading of this manuscript.

References

[1] M. Avellaneda, "The minimum-entropy algorithm and related methods for calibrating asset-pricing models," *Proceedings of the International Congress of Mathematicians*, Vol III. Doc. Math., Berlin, 1998, 545–563.

[2] J. Barzilai and J. M. Borwein, "Two point step size gradient methods." *IMA J. Numer. Anal.*, **8** (1988), 141–148.

[3] H.H. Bauschke and J.M. Borwein, "On the convergence of von Neumann's alternating projection algorithm for two sets," Set-Valued Analysis, **1** (1993), 185–212.

[4] H.H. Bauschke and J.M. Borwein, "On projection algorithms for solving convex feasibility problems," *SIAM Review*, **38** (1996), 367–426.

[5] H.H. Bauschke and P.L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, CMS-Springer Books, 2011.

[6] H.H. Bauschke and P.L. Combettes, and D. Russell Luke, "Phase retrieval, error reduction algorithm, and Fienup variants: a view from convex optimization." *J. Opt. Soc Amer. A,* **19** (2002), 1334–1345.

[7] J.M. Borwein, "The Oxford Users' Guide to Mathematics", featured *SIAM REVIEW*, **48**:3 (2006), 585–594.

[8] J. Borwein, D. Bailey, N. Calkin, R. Girgensohn, R. Luke, and V. Moll, *Experimental Mathematics in Action*, A.K. Peters, 2007.

[9] J.M. Borwein, R. Choksi and P. Maréchal, "Probability distributions of assets inferred from option prices via the Principle of Maximum Entropy," *SIAMOpt*, **4** (2003), 464–478.

[10] J.M. Borwein and C. Hamilton, "Symbolic Convex Analysis: Algorithms and Examples," *Math Programming*, **116** (2009), 17–35.

[11] J.M. Borwein and W.Z. Huang, "A fast heuristic method for polynomial moment matching with Boltzmann-Shannon entropy," *SIAM J. Optimization*, **5** (1995), 68–99.

[12] Jonathan M. Borwein, Phil Howlett and Julia Piantadosi, "Copulas with Maximum Entropy." *Optimization Letters.* E-published, May 2011.

[13] J. M. Borwein and A. S. Lewis, "Duality relationships for entropy–like minimization problems," *SIAM Control and Optim.*, **29** (1991), 325–338.

[14] J.M. Borwein and A.S. Lewis, "Partially-finite convex programming, (I,II)," *Mathematical Programming*, Series B, **57**, (1992) 15–48 (49-83).

[15] J.M. Borwein and A.S. Lewis, "Strong rotundity and optimization," *SIAM J. Optimization*, **4** (1994), 146–158.

[16] J.M. Borwein and A.S Lewis, *Convex Analysis and Nonlinear Optimization* CMS-Springer, 2nd expanded edition, 2005.

[17] J.M. Borwein, A.S. Lewis, M.N. Limber and D. Noll, "Maximum entropy spectral analysis using first order information. Part 2,"*Numer. Math*, **69** (1995), 243–256.

[18] J.M. Borwein and M. Limber, "Under-determined moment problems: a case for convex analysis," *SIAMOpt*, Fall 1994.

[19] J. Borwein, M. Limber and D. Noll, "Fast heuristic methods for function reconstruction using derivative information," *App. Anal.*, **58** (1995), 241–261.

[20] J.M. Borwein and R.L. Luke, "Duality and Convex Programming," pp. 229–270 in *Handbook of Mathematical Methods in Imaging,* O. Scherzer (Ed.), Springer, 2010.

[21] Jonathan M. Borwein and D. Russell Luke, "Entropic Regularization of the ℓ_0 function." Chapter 5, pp. 65–91 *Fixed-Point Algorithms for Inverse Problems in Science and Engineering* in *Springer Optimization and Its Applications*. Galleys November 2010.

[22] J. M. Borwein, P. Maréchal and D. Naugler,"A convex dual approach to the computation of NMR complex spectra," *Mathematical Methods of Operations Research*, **51** (2000), 91–102.

[23] J.M. Borwein and B. Sims, "The Douglas-Rachford algorithm in the absence of convexity." Chapter 6, pp. 93–109 in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering* in *Springer Optimization and Its Applications*, 2011.

[24] J.M. Borwein and J.D. Vanderwerff, *Convex Functions*, Cambridge University Press, 2010.

[25] J.M. Borwein and Qiji Zhu, *Techniques of Variational Analysis*, CMS/Springer, 2005.

[26] Christopher J. Bose and Rua Murray, "Duality and the computation of approximate invariant densities for nonsingular transformations," *SIAM J. Optim*, **18** (2007), 691-707.

[27] Christopher J. Bose and Rua Murray, "Maximum entropy estimates for risk-neutral probability measures with non-strictly-convex data," preprint, 2011.

[28] S. Boyd, and L. Vandenberghe, *Convex Optimization*, 317, Cambridge University Pres, 2004.

[29] M.N. Limber, A. Celler, J.S. Barney, M.A. Limber, J.M. Borwein, "Direct Reconstruction of Functional Parameters for Dynamic SPECT," *IEEE Transactions on Nuclear Science*, **42** (1995), 1249–1256.

[30] F. H. Clarke, Yu. S. Ledyaev, R.J. Stern,and P. R. Wolenski, *Nonsmooth analysis and control theory*, Graduate Texts in Mathematics, **178**, Springer-Verlag, 1998.

[31] George Dantzig, "Reminiscences about the origins of linear programming, 1 and 2", *Oper. Res. Letters*, April 1982, p. 47.

[32] A. Decarreau, D. Hilhorst, C. LeMaréchal and J. Navaza, "Dual methods in entropy maximization. Application to some problems in crystallography," *SIAM J. Optim.* **2** (1992), 173–197.

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

[34] Kenneth Davidson and Allan P. Donsig, *Real Analysis and Applications* Springer Undergraduate Texts in Mathematics, 2008.

[35] 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** (1991), 1164–1180.

[36] I. Ekeland and R. Témam, *Convex analysis and variational problems.* Translated from the French. Corrected reprint of the 1976 English edition. Classics in Applied Mathematics, 28. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1999.

[37] V. Elser, V., I. Rankenburg, and P. Thibault, "Searching with iterated maps," *Proceedings of the National Academy of Sciences* **104** (2007), 418–423s

[38] M. Fabian, P. Habala, P. Hájek, V. Montesinos and V. Zizler, *Banach Space Theory*, CMS-Springer-Verlag, 2010.

[39] W. Fenchel, "On conjugate convex functions," *Canad. J. Math.* 1 (1949), 73–77.

[40] W. Fenchel, *Convex Cones, Sets and Funmctions* Lecture Notes, Princeton 1953.

[41] David Gale, "A geometric duality theorem with economic applications," *The Review of Economic Studies*, **34** (1967), 19—24.

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

[43] S. Gravel, S. and V. Elser, "Divide and concur: A general approach constraint satisfaction," preprint, 2008, `http://arxiv.org/abs/0801.0222v1`.

[44] A. Guerraggio, "The origins of quasi-concavity: a development between mathematics and economics," *Historia Math.* **31** (2004), 62–75.

[45] J. Havel, *Gamma: Exploring Euler's Constant*, Princeton University Press, 2003.

[46] 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.

[47] 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.

[48] G. Minerbo. "MENT: A maximum entropy algorithm for reconstructing a source from projection data." *Computer Graphics and Image Processing*, 10:48–68, 1979.

[49] J.-P. Penot, *Calculus without derivatives.* Book manuscript, May 2011.

[50] M. B. Priestly, *Non-linear and non-stationary time series analysis*, Academic Press, London, 1988.

[51] F. Pukelsheim, *Optimal design of experiments*, Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, 1993.

[52] R.T. Rockafellar, *Convex Analysis*, Princeton University Press, 1970.

[53] S. Simons, *From Hahn-Banach to Monotonicity*, Lecture Notes in Mathematics, **1693**, Springer-Verlag, 2008.

These and other references are available at http://docserver.carma.newcastle.edu.au and quotations are at http://carma.newcastle.edu.au/jon/quotations.html.

Laureate Professor and Director CARMA, University of Newcastle, Australia. Distinguished Professor, King Abdulaziz University, Jeddah, SA.

E-mail address: jonathan.borwein@newcastle.edu.au