Entropy and Projection Methods
for Convex and Nonconvex Inverse Problems

First prepared for

Technion Colloquium

Haifa May, 21 2012   Revised: 19/12/13

....

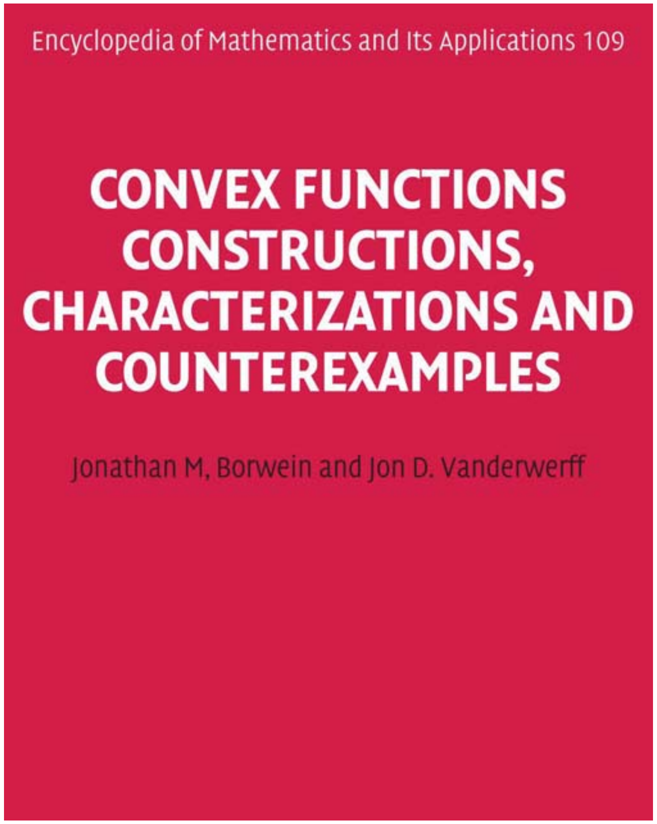
Jonathan M. Borwein, FRSC FAAAS FAA

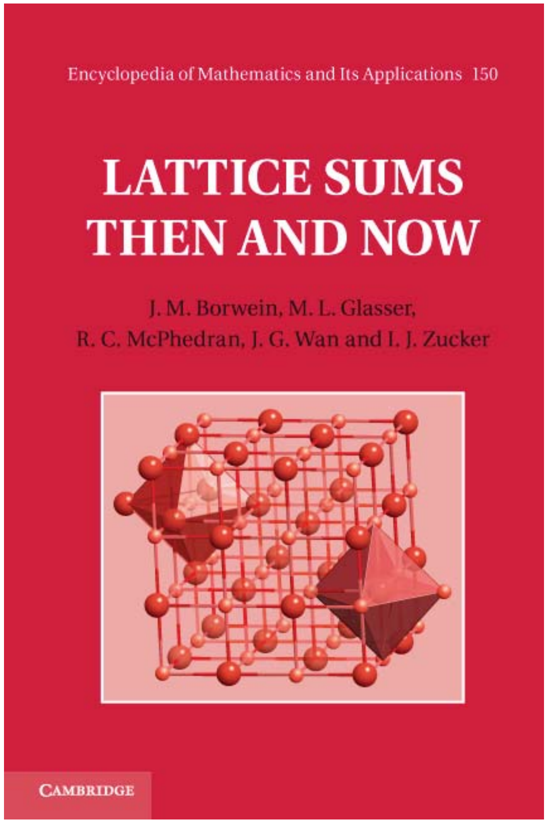
Laureate Professor and Director

School of Math and Phys Sciences, Univ of Newcastle, NSW

URL: www.carma.newcastle.edu.au/jon

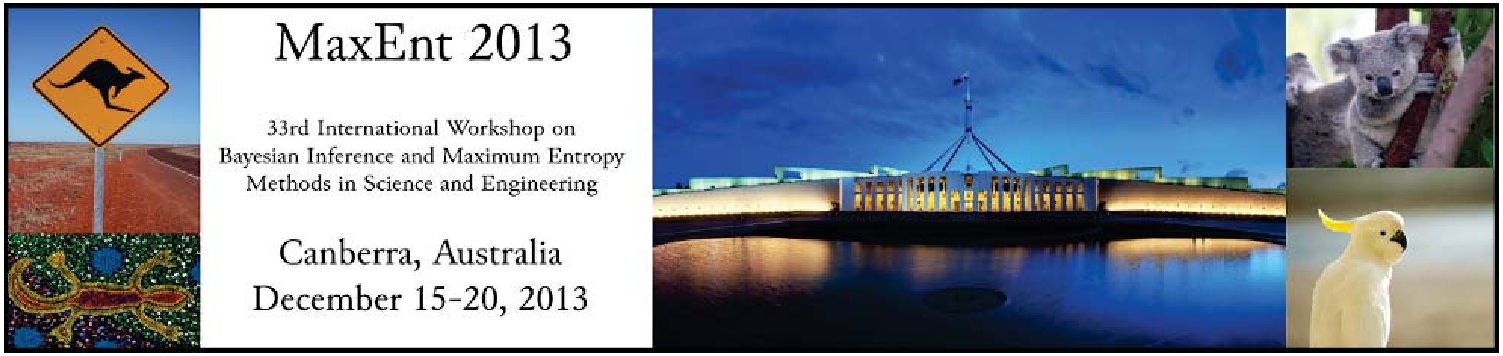
MY TWO MAIN RESEARCH FIELDS





Functional analytic optimization

Special functions and computation



The companion paper to this talk is:

J.M. Borwein, "Maximum entropy and feasibility methods for convex and non-convex inverse problems." *Optimization*, **61** (2012), 1–33.

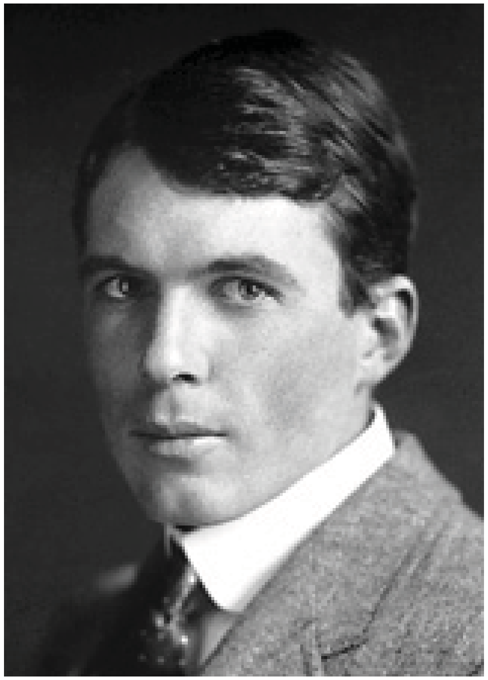
CARMA

I SHALL FOLLOW BRAGG

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

(1890-1971)

Nobel Crystallography

(Adelaide)

AND SANTAYANA

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**

In **Persons and Places**, 1945, 238–239.

FOUR 'FINE' BOOK REFERENCES:

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

BL1 J.M. Borwein and A.S Lewis, *Convex Analysis and Nonlinear Optimization*, CMS/Springer, 2nd expanded edition, 2005.

BLu 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 & 2015.

BV J.M. Borwein and J.D. Vanderwerff, *Convex Functions: Constructions, Characterizations and Counterexamples*, Cambridge Univ Press, 2010.

OUTLINE

I shall discuss in "tutorial mode" the formalization of **inverse problems** such as signal recovery and option pricing: first as (convex and non-convex) **optimization problems** and second as **feasibility problems**—each over the infinite dimensional space of signals. I shall touch on*:

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.

*More is an unrealistic task!

2. Convex programming duality:

– what it is and what it buys you.

3. Algorithmic consequences: for both design and implementation.

and *as time permits* (it won't)

4. Non-convex extensions & feasibility problems: life is hard. Entropy methods, used directly, have little to offer:

– sometimes (Hubble, protein reconstruction, Suduko, 3SAT, ...) more works than we know why it should.

- See also `http://carma.newcastle.edu.au/DRmethods/`

THE GENERAL PROBLEM

Many applied problems reduce to "**best**" solving (under-determined) systems of **linear** (or non-linear) equations:

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

where $\mathbf{b} \in \mathbb{R}^n$, and the unknown x lies in some appropriate function space.

> *The infinite we shall do right away. The finite may take a little longer*. **Stan Ulam**

- In D. MacHale, *Comic Sections* (Dublin 1993)

Discretisation 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. And guided by our 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 nonlinear) where the unknown is most naturally a function, not a vector in $\mathbb{R}^m$.

EXAMPLE 1. AUTOCORRELATION

- Consider, extrapolating an *autocorrelation function* from given sample measurements:

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

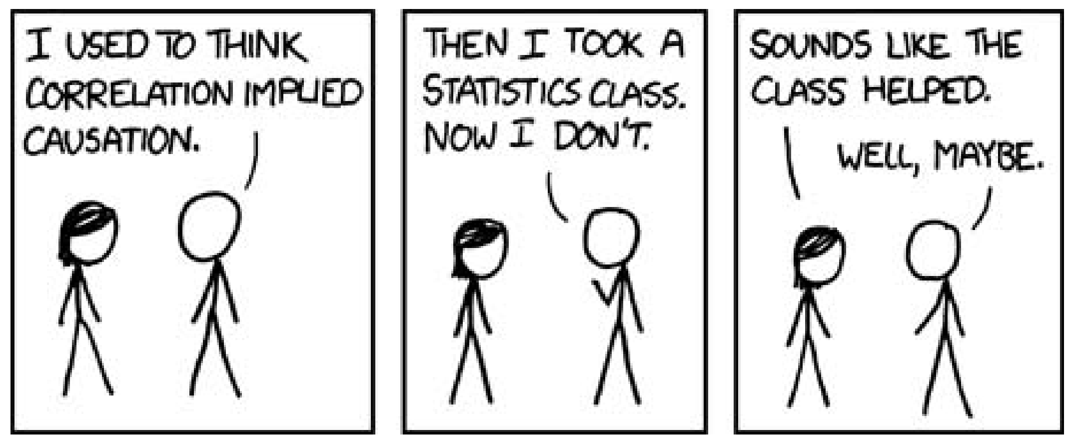
◇ (**Wiener-Khintchine**) Fourier moments of the power spectrum $S(\sigma)$ are samples of the autocorrelation function, so values of $R(t)$ computed directly from the data yields *moments* of $S(\sigma)$.

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

- Hence, we may compute a *finite* number of moments of S; use them to make estimate $\widehat{S}$ of S;

- We may then *estimate more moments* from $\widehat{S}$ by direct numerical integration. So we dually *extrapolate* R ...

- This avoids having to compute R directly from potentially noisy (unstable) larger data series.



PART ONE: THE ENTROPY APPROACH

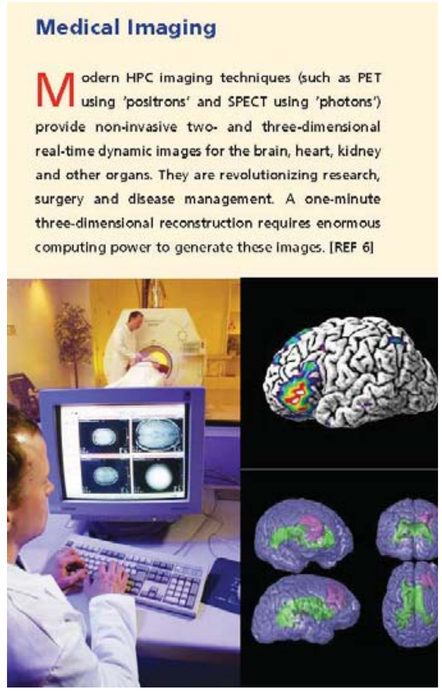
- Following [**BZ**] I sketch a maximum entropy approach to under-determined systems where the unknown, x, is a function, typically living in a *Hilbert space*, or more general space of functions.

  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, $f(x)$, of the unknown x.

◇ The approach finds applications in countless fields:

Including (to my personal knowledge) Acoustics, actuarial science, astronomy, biochemistry, compressed sensing, constrained spline fitting, engineering, finance, hydrology, image reconstruction, inverse scattering, multidimensional NMR (MRI), optics, option pricing, *philosophy*, tomography, statistical moment fitting, and time series analysis, ...

(Many thousands of papers)

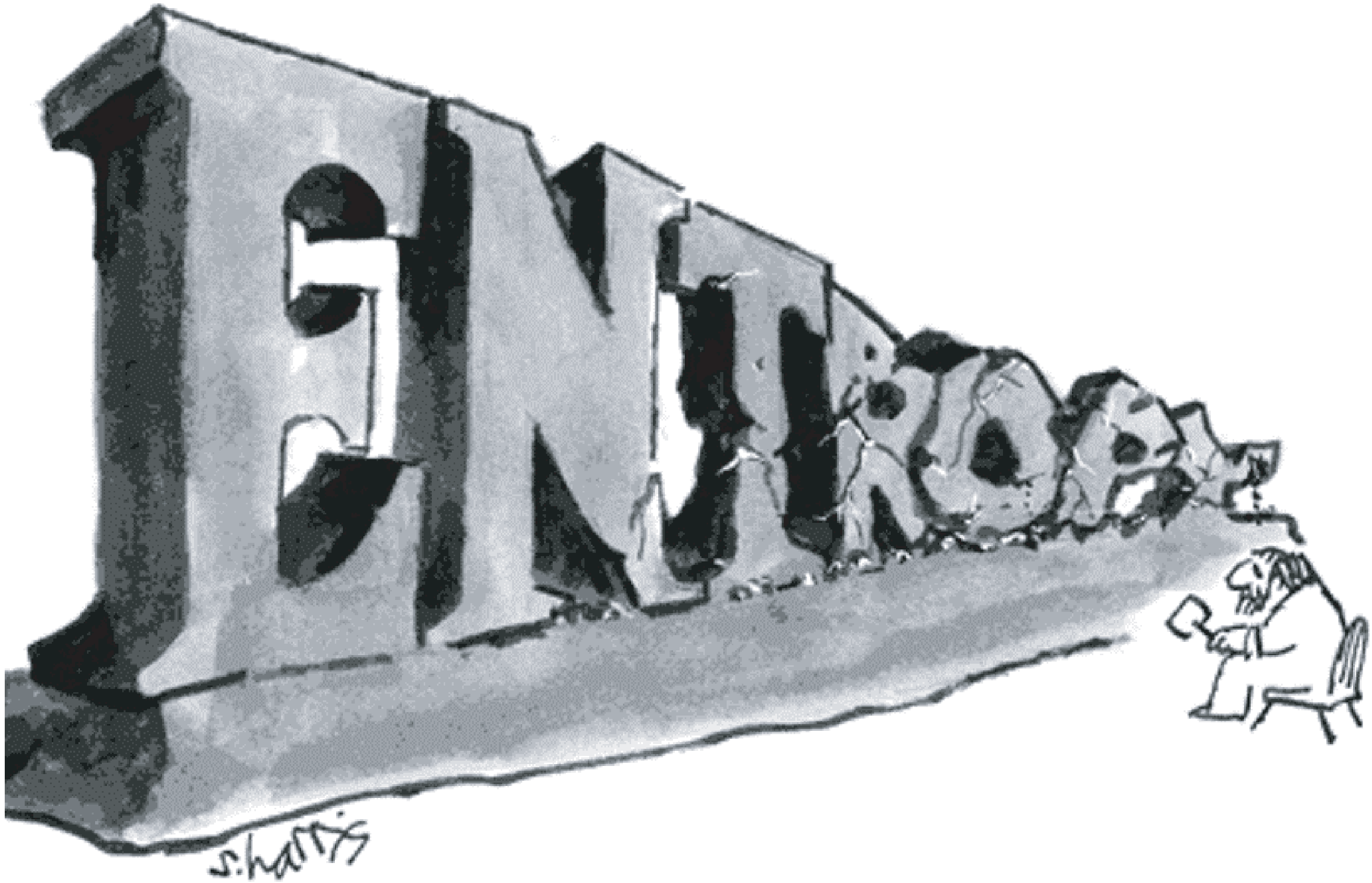


However, the derivations and mathematics are fraught with subtle — and less subtle — errors.

`www.carma.newcastle.edu.au`

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.

WHAT is



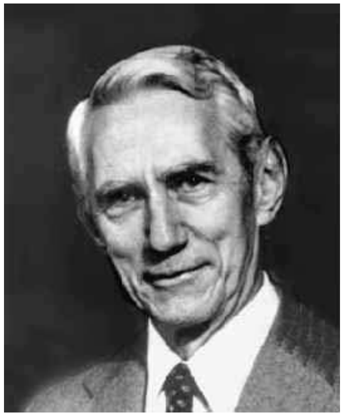
WHAT is ENTROPY?

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."

CHARACTERIZATIONS of ENTROPY

Boltzmann (1844-1906)



Shannon (1916-2001)

- **19C**: **Ludwig Boltzmann** — thermodynamic *disorder*
- **20C**: **Claude Shannon** — information *uncertainty*
- **21C**: **JMB** — potentials with *superlinear growth*

- Information theoretic characterizations abound. A nice example is:

Theorem. 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 $0 < p < 1$,

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

ENTROPIES FOR US

Let X be our *function space*, typically Hilbert space $L^2(\Omega)$, or the function space $L^1(\Omega)$ (or a Sobolev space).

◇ For $+\infty \geq p \geq 1$,

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

Recall that $L^2(\Omega)$ is a Hilbert space with *inner product*

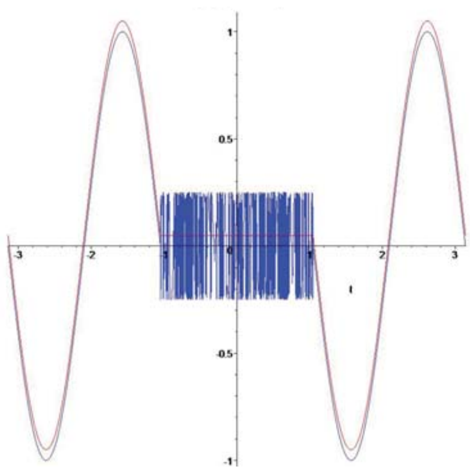
$$\langle x, y \rangle := \int_\Omega x(t)y(t)dt,$$

(with variations in Sobolev space).

A *bounded linear map* $A : X \to I\!R^n$ is 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).



Lebesgue's continuous function with divergent Fourier series at 0.

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

$\otimes$ The most common approach is to find the **minimum norm solution*** by solving the *Gram system*:

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

$\oplus$ The primal solution is then $\widehat{x} = A^T\lambda$. Elaborated, this recaptures all of *Fourier analysis*, e.g., Lebesgue's example!

- This solved the following *variational problem*:

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

*Even in the (realistic) infeasible case.

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

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

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

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

- The functional f can be used to include other constraints†.

*Essentially $\phi''(t) > 0$.

†Including nonnegativity, by appropriate use of $+\infty$.

For example, 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*.

- Entropy constructions abound: two useful classes follow.

  - *Bregman* (based on $\phi(y) - \phi(x) - \phi'(x)(y - x)$); and

  - *Csizar distances* (based on $x\phi(y/x)$)

- Both model statistical divergences.

Two popular choices—both discrete and continuous (differential)—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),

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

$\triangle$ Includes the *log barrier* and *log det* functions from interior point theory.

$\diamond$ *Both implicitly impose a nonnegativity constraint* (positivity in Burg's non-superlinear case).

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

This is more theology than science !

- Use of the Csizar distance based *Fisher Information*

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

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

WHAT 'WORKS' BUT CAN GO WRONG?

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

• As L^2 is infinite dimensional, so is $N(A)$.
That is, if $Ax = \mathbf{b}$ is solvable, it is under-determined.

We pick our solution to *minimize*

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

$\odot$ $\phi(x(t), x'(t))$ in Fisher-like cases [**BN1, BN2, BV10**].

- 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*

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

- So we formally have a "dual pair" (BL1)

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

and its dual

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

• Moreover, for the solutions $\widehat{x}$ to (P), $\widehat{\lambda}$ to (D), the derivative (w.r.t. x) of $L(x, \widehat{\lambda})$ should be zero, since

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

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

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

this implies

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

• We can now reconstruct the primal solution (qualitatively and quantitatively) from a presumptively easier dual computation.

A DANTZIG (1914-2005) ANECDOTE

"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."

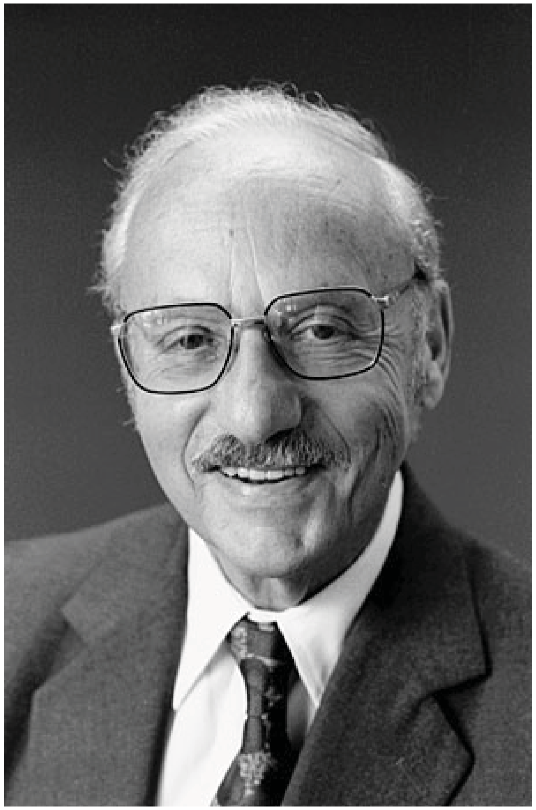
A lovely story. I heard George recount this a few times and, when he came to the "conception" part, he always had a twinkle in his eyes. (Saul Gass, 2006)

George wrote in "Reminiscences about the origins of linear programming," 1 and 2, *Oper. Res. Letters*, April 1982 (p. 47):

In a Sept 2006 *SIAM book review* about dictionaries[a], I asserted George assisted his father with his dictionary — for reasons I still believe but cannot reconstruct.

I also called Lord Chesterfield, Lord Chesterton (*gulp*!). Donald Coxeter used to correct such errors in libraries.

[a] *The Oxford Users' Guide to Mathematics*, Featured *SIAM REVIEW*, **48**:3 (2006), 585–594.



PITFALLS ABOUND

There are 2 major problems to this approach.

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

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

$\diamond$ The optimal value is not attained. As we will see, existence can fail for the Burg entropy with three-dim trig moments. Additional conditions on ϕ are needed to insure solutions exist.* [**BL2**]

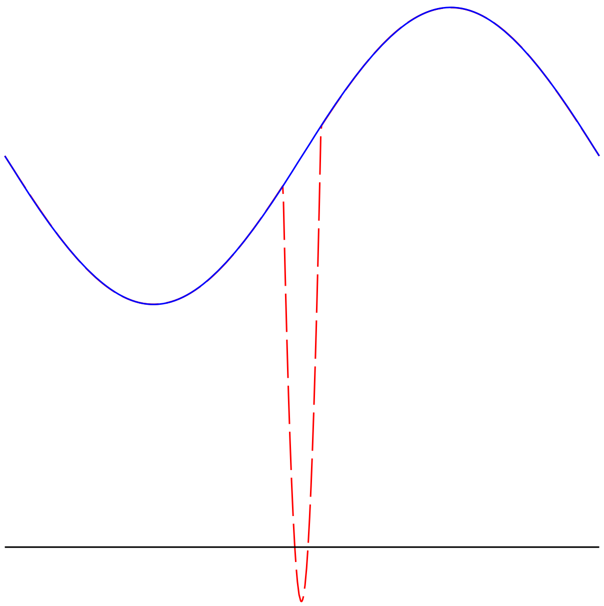
*The solution is actually the *absolutely continuous part of a measure* in $C(\Omega)^*$

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.

  $\dashrightarrow \quad \dashrightarrow \quad \dashrightarrow$

- The Lagrangian is *nowhere continuous*, much less differentiable.



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

FIXING THE PROBLEM

One way to get continuity/differentiability of f, is to:

- work in $L^\infty(\Omega)$, or $C(\Omega)$ using essentially bounded, or continuous, functions.

  But, even with such side qualifications, solutions to (P) *may still not* exist.

∇ Consider Burg entropy maximization in $L^1[T^3]$:

$$\mu := \sup \int_{T^3} \log(x)\,dV \ \text{ s.t. } \int_{T^3} x\,dV = 1 \qquad \text{and}$$

$$\int_{T^3} x\cos(a)\,dV = \int_{T^3} x\cos(b)\,dV = \int_{T^3} x\cos(c)\,dV = \alpha.$$

For $1 > \alpha > \overline{\alpha}$, sol'n is measure in $(L^\infty)^*$.

For $0 < \alpha < \overline{\alpha}$ sup is attained in L^1.

Value of $\overline{\alpha}$ is computable [**BL2**]. (Watson integral for face centered cubic lattice.)

We see continuous part of measure on screen.

Werner Fenchel (1905-1988)

- Minerbo, e.g., 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.

CONVEX ANALYSIS (AN ADVERT)

We will give a theorem that guarantees the form of solution found in the above faulty derivation

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

is, in fact, correct. (Full derivation in [**BL2**, **BZ**].)

- We introduce the *Fenchel (Legendre) conjugate* [**BL1**] of a function $\phi : \mathbb{R} \to (-\infty, +\infty]$:

$$\phi^*(u) = \sup_{v \in \mathbb{R}} \{uv - \phi(v)\}.$$

• Often this can be (pre-)computed explicitly
– using Newtonian calculus. Thus,

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

yield

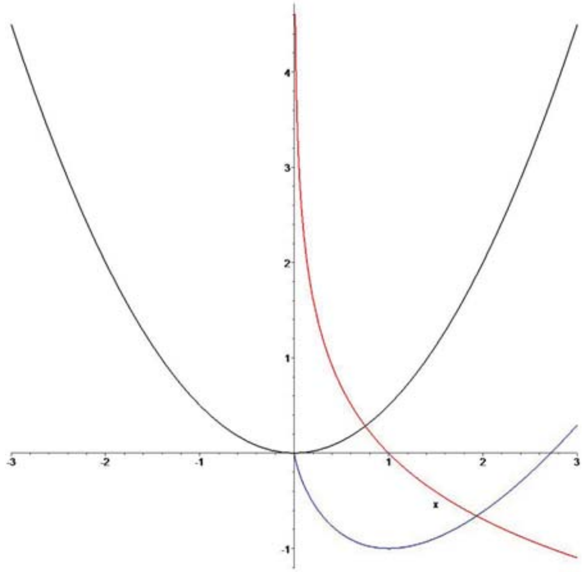
$$\phi^*(u) = \exp(u), -1 - \log(-u) \text{ and } u^2/2$$

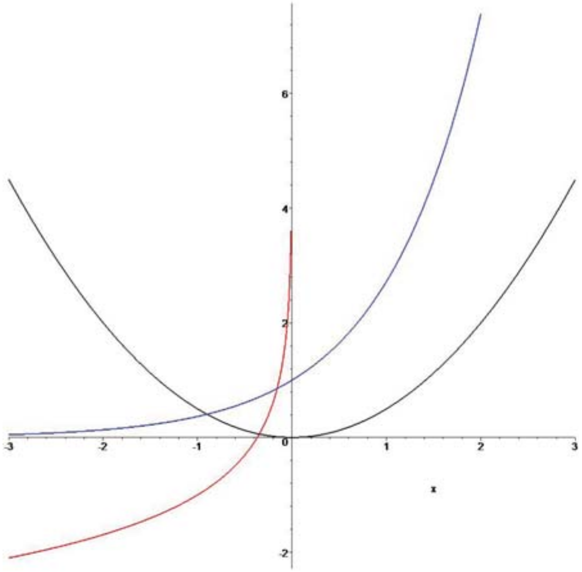
respectively. Red is the *log barrier* of interior point fame!

• The Fisher case is also explicit

— via an integro-differential equation.

PRIMALS AND DUALS





The three entropies below and their conjugates.

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

and

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

EXAMPLE 2. CONJUGATES & NMR

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:

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

for *quantum theoretic* (NMR) reasons.

- Recall the *Fenchel-Legendre conjugate*

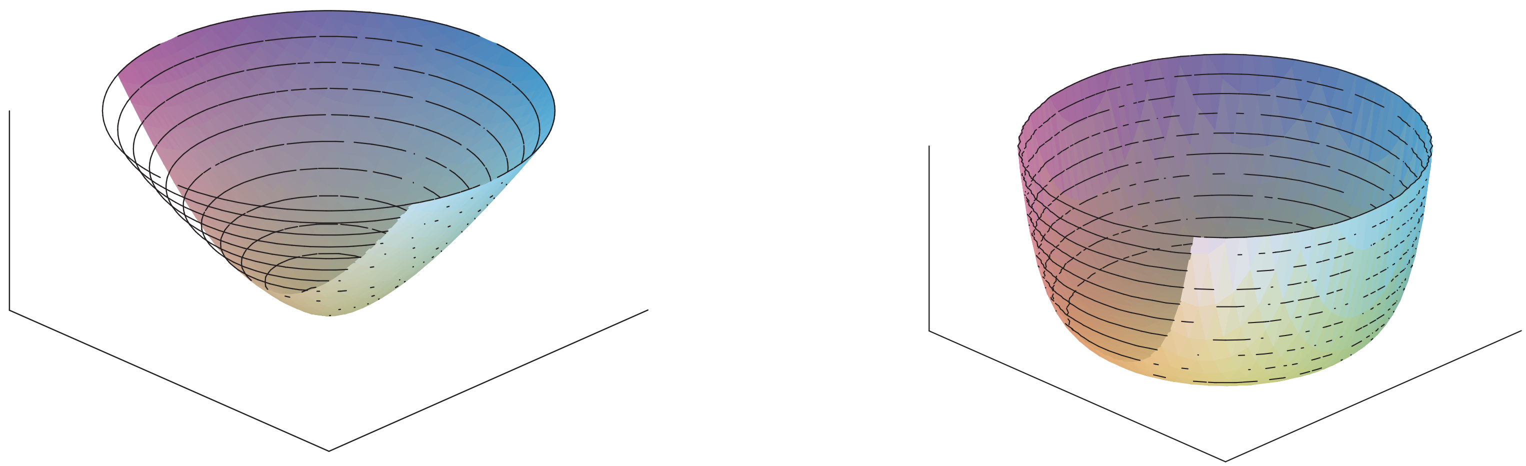
$$f^*(y) := \sup_x \langle y, x \rangle - f(x).$$

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

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

◇ Compare the *Shannon entropy*:

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



The NMR entropy and its conjugate.

http://carma.newcastle.edu.au/ConvexFunctions/links.html

FENCHEL DUALITY THEOREM (1951)

Theorem 1 (Utility Grade). *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. 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).*

- **indicator function** $\iota_C(x) := 0$ if $x \in C$ and $+\infty$ else.

- **support function** $\sigma_C(x^*) := (\iota_C)^*(x^*) = \sup_{x \in C} \langle x^*, x \rangle$.

EXAMPLES include:

(i) $A := I$ is equivalent to Hahn-Banach theorem.

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

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

– specializes to LP if $f := \iota_{R_n^+} + c$.

(iii) $f := \iota_C, g := \sigma_D$ yields minimax theorem:

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

FENCHEL DUALITY (SANDWICH)

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

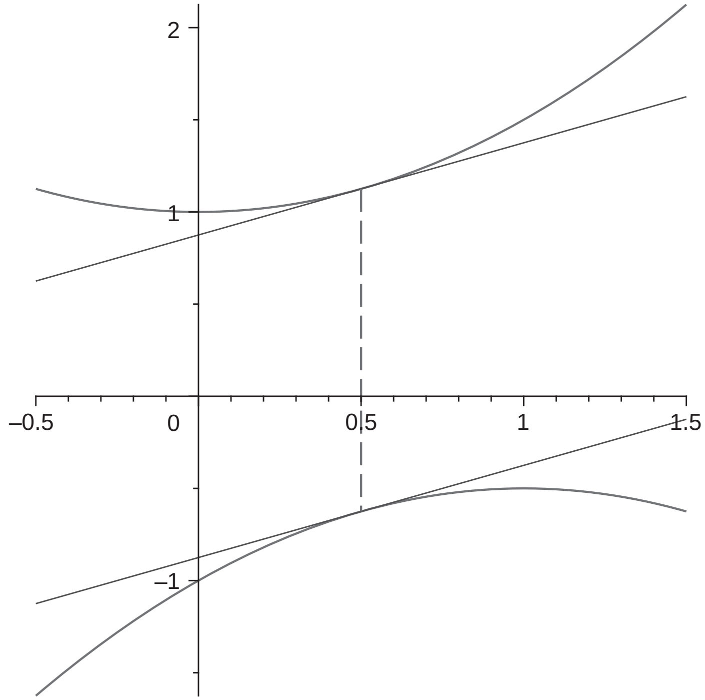


Figure 2.6 Fenchel duality (Theorem 2.3.4) illustrated for $x^2/2 + 1$ and $-(x-1)^2/2 - 1/2$. The minimum gap occurs at $1/2$ with value $7/4$.

- Using the concave conjugate: $g_* := -(-g)^*(-)$.

COERCIVITY AND PROOF OF DUALITY

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

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

Then $v \to v \log v$, $v \to v^2/2$ and the positive energy all have regular growth but -log does not.

- The *domain* of a convex function is

$$\text{dom}(\phi) = \{u : \phi(u) < +\infty\};$$

and ϕ is *proper* if $\text{dom}(\phi) \neq \emptyset$.

• Let $\imath := \inf \operatorname{dom}(\phi)$ and $\sigma := \sup \operatorname{dom}(\phi)$.

Our *constraint qualification*,* (CQ), reads:

$$\boxed{\begin{array}{c}\exists\, \overline{x} \in L^1(\Omega),\ \text{such that } A\overline{x} = \mathbf{b},\\ f(\overline{x}) \in \mathbb{R},\ \ \imath < \overline{x} < \sigma\ \ a.e.\end{array}}$$

◇ *In many cases, (CQ) reduces to feasibility*
– e.g., spectral estimation, and trivially holds.

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

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

*To ensure dual solutions. Standard Slater condition fails. Fenchel *missed* need for a (CQ) in his 1951 *Princeton Notes*.

Theorem 2 (BL2). *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**

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

then the unique solution to (P) is given by

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

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

*This is trivial if $d = \infty$.

♠ We have obtained a powerful *functional reconstruction* for all $t \in \Omega$.

- This generalises to cover $\Omega \subset \mathbb{R}^n$, and more elaborately in Fisher-like cases [**BL2**], [**BN1**], etc.

'Bogus' differentiation of a discontinuous function becomes the delicate conjugacy formula:

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

Thus, the form of the maxent solution can be legitimated by validating the *easily* checked conditions of Thm. 2.

♠ Also, any solution to $Ax = \mathbf{b}$ of the form in (2) is automatically a solution to (P).

So solving (P) is equivalent to finding $\lambda \in \mathbb{R}^n$ with

$$\boxed{\langle a_i, (\phi^*)'(A^T\lambda)\rangle = b_i, \quad i = 1, \ldots, n} \tag{3}$$

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

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

Often $(\phi')^{-1} = (\phi^*)'$

- So the 'dubious' solution and 'honest' 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 50 ...).

Note that discretization is only needed to compute terms in evaluation of (3).

Indeed, *these integrals can sometimes be computed exactly* (e.g., in some tomography and option estimation problems). This is the gain of *not discretizing* early.

> By waiting to see the form of dual, 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**] and Huang's 1993 thesis).

EXAMPLE 3. OPTION PRICING

For European option portfolio pricing the constraints are based on 'hockey-sticks' of the form:

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

- In this case the dual can be computed *exactly* and leads to a relatively small and explicit nonlinear equation to solve the problem (see [**BCM**]).

The more nonlinear the optimization problem the more *dangerous* it is to treat it purely formally.

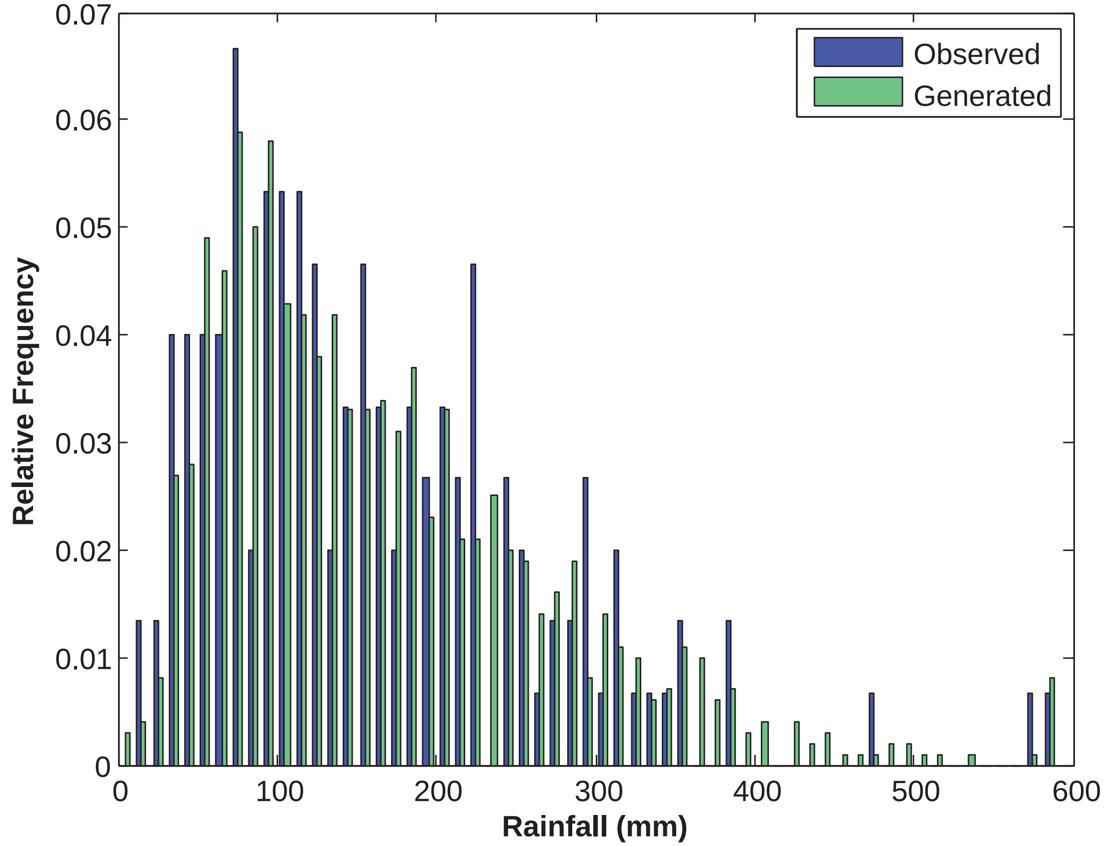
EXAMPLE 4. MODELLING RAINFALL

In **PHB, PHBH, 2012-2013** *checkerboard copulas* of maximum entropy were constructed to simulate monthly spring (and fall) rainfall at Sydney (and Kempsey)

- while preserving monthly correlations without back-fitting
  - and so to produce realistic variance in accumulated rainfall totals.

- Incomplete Gamma distributions were used for marginals
  - again justified by MaxEnt.

Accumulated rainfall totals over months Oct-Nov



Comparison of mean and variance for *observed accumulated totals*; generated accumulated totals using independent random variables (*Independent Model*) and copula of maximum entropy (*Correlated Model*)

	Mean (mm)	**Variance**
Observed Data	160.488	10830.299
Independent Model	161.705	8732.117
Correlated Model (Max Ent)	160.451	10769.729

– P-values for Kolmogorov-Smirnov goodness of fit: Observed versus generated **0.7637**.

- Normal copulas give similar (slightly worse?) results but are more costly computationally.

FROM FENCHEL'S ACORN ...

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, $\phi(\xi)$.

- in *Canad. J. Math*, volume **1**, #1.

1949

. . . a MODERN OAK

Theorem 2 works by relaxing the problem to $\left(L^1\right)^{**}$ — where solutions always exist — and using Lebesgue decomposition.

• Regular growth rules out a non-trivial singular part via analysis with the formula:

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

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

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

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

We say I_ϕ is *strongly rotund* (*very well posed*) if it is (i) strictly convex with (ii) weakly compact lower level sets (Dunford-Pettis) and (iii) *Kadec-Klee*:

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

Theorem 3 (BV)**.** *I_ϕ is strongly rotund as soon as ϕ^* is everywhere finite and differentiable on $\mathbb{R}$; and conversely when μ is not purely atomic.*

- Easy to check (holds for Shannon and energy but not Burg) and is the best surrogate for the properties of a reflexive norm on L^1.

$\mathcal{M}$om$\mathcal{E}$nt+

An old interface: $\mathcal{M}$om$\mathcal{E}$nt+ (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*.

COMPARISON OF ENTROPIES

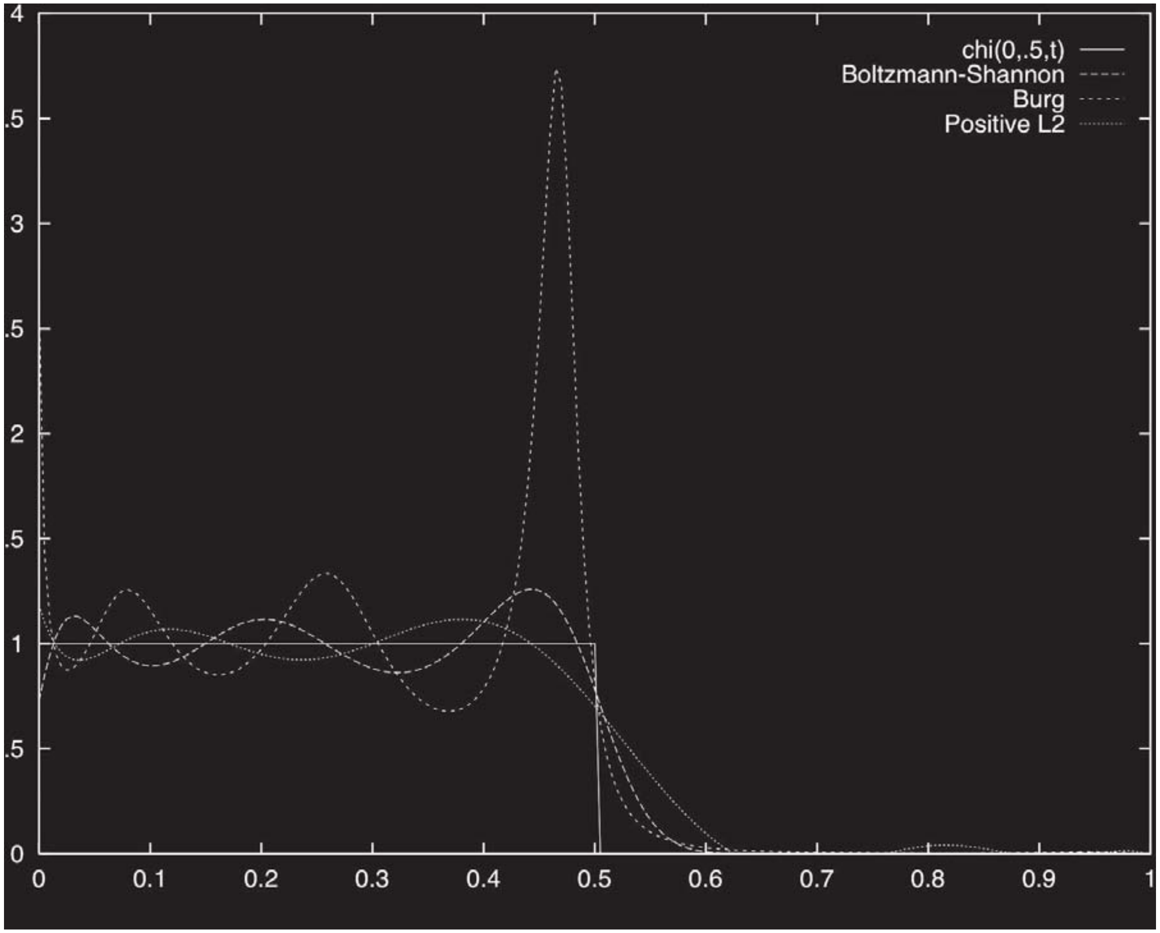
We compare the positive L^2, Boltzmann-Shannon and Burg entropy reconstruction of the **characteristic function** $\chi_{[0,1/2]}$ using **10** *algebraic moments*

$$b_i = \int_0^{1/2} t^{i-1}\, dt$$

on $\Omega = [0,1]$.

Burg over-oscillates since $(\phi^*)'(t) = 1/t$. But is still often the 'best' solution (with a closed form for Fourier moments)!

- Relaxation adds stability but degrades the reconstruction: a dance with ill-posedness.



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

PART TWO: THE NON-CONVEX CASE

For iterative methods as below, I recommend:

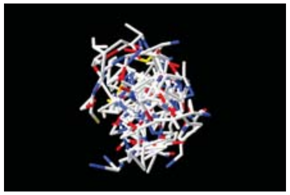
BaB H.H. Bauschke and J.M. Borwein, "On projection algorithms for solving convex feasibility problems," *SIAM Review*, **38** (1996), 367–426 (aging well with nearly 500 ISI cites).

BaC H.H. Bauschke and P.L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, CMS-Springer Books, 2012.

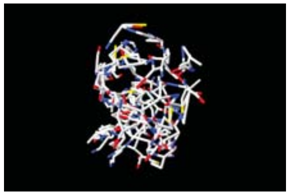
- In general, non-convex optimization is a much less satisfactory pursuit.

- We *can usually hope only to find critical points* ($f'(x) = 0$) or local minima.
  - Thus, **problem-specific heuristics** dominate:

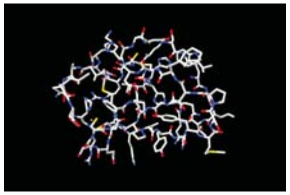
Douglas–Rachford method reconstruction:



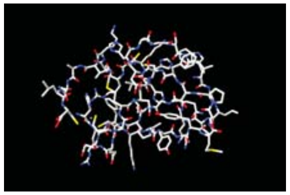
500 steps, -25 dB.



1,000 steps, -30 dB.

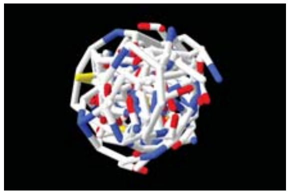


2,000 steps, -51 dB.

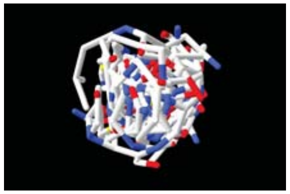


5,000 steps, -84 dB.

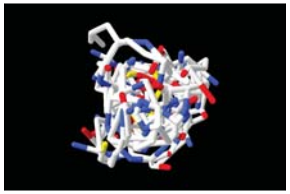
Alternating projection method reconstruction:



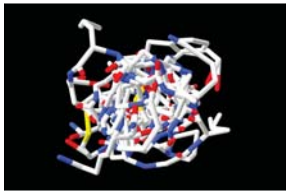
500 steps, -22 dB.



1,000 steps, -24 dB.



2,000 steps, -25 dB.



5,000 steps, -28 dB.

EXAMPLE 5. CRYSTALLOGRAPHY

We wish to estimate x in $L^2(\mathbb{R}^n)^*$ and can suppose the modulus $c = |\widehat{x}|$ is known (here $\widehat{x}$ is the Fourier transform of x).†

Now $\{y\colon |\widehat{y}| = c\}$, is not convex. So the issue is to find x given c and other convex information.

An appropriate problem extending the previous one is

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

where M models the modular constraint, and f is as in Theorem 2.

*Here $n = 2$ for images, $n = 3$ for holographic imaging, etc.

†Observation of the modulus of the diffracted image in crystallography. Similarly, for optical aberration correction.

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

They suggest solving

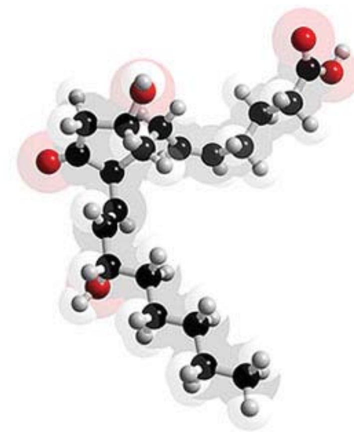
$$\min\{f(x) : Ax = y, \|B_k y\| = b_k, (k \in K) \;\; x \in X\}, \tag{NP^*}$$

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. 2.

I am obscure, mainly because the results were largely negative:

They 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.



- 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.*"

- I had similar experiences with non-convex medical image reconstruction.

  "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)**

GENERAL PHASE RECONSTRUCTION

The basic setup — more details follow.

- **Electromagnetic field:** $u : \mathbb{R}^2 \to \mathbb{C} \in L^2$

- **DATA:** Field intensities for $m = 1, 2, \ldots, M$:

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

- **MODEL:** Functions $\mathcal{F}_m : L^2 \to L^2$, are *modified Fourier Transforms*, for which we can measure the modulus (intensity)

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

ABSTRACT INVERSE PROBLEM

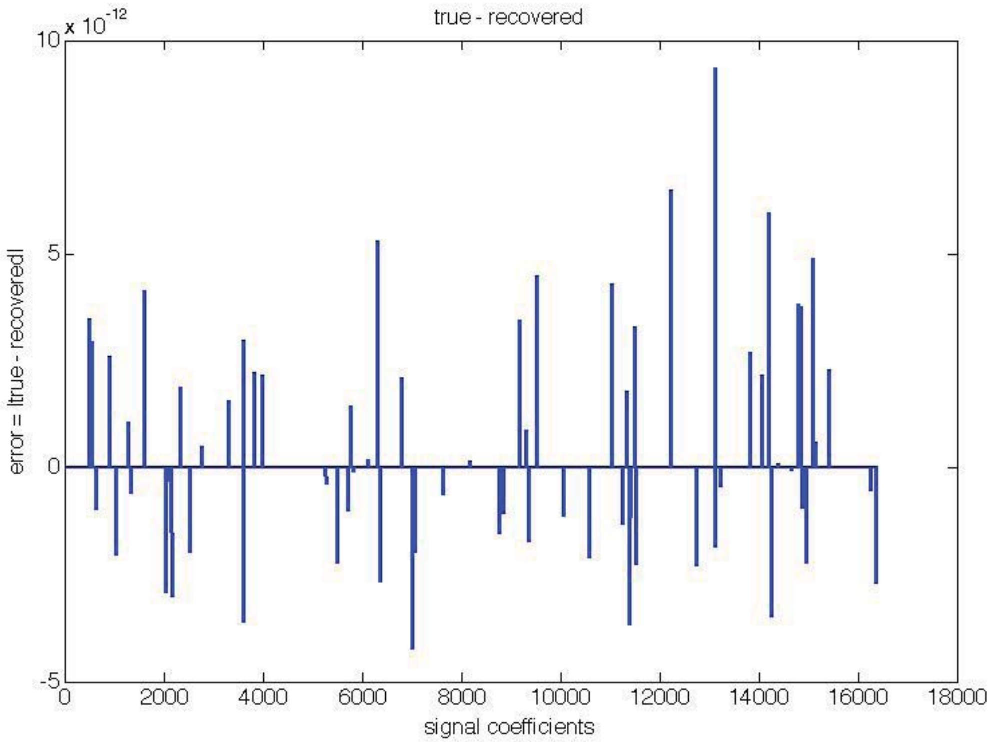
Given transforms

$$\mathcal{F}_m$$

and **measured** field intensities

$$\psi_m$$

(for $m = 1, \ldots, M$), find a **robust estimate** of the underlying field function u.



EXAMPLE 6. 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 asked 10 teams to devise algorithmic fixes.

- **Optical aberration correction**, using the *Misell algorithm*, a *method of alternating projections*, works much better than it should — given that it is being applied to:

PROBLEM. Find a member of a version of

$$\Psi := \bigcap_{k=1}^{M} \{x : Ax = b, \|M_k x\| = c_k, \;\; x \in X\}, \tag{NCFP}$$

which is a M-set **non-convex feasibility problem** as examined more below.

- Is there hidden convexity to explain good behaviour?

- Misell is now built in to home computer telescopes.

HUBBLE IS ALIVE AND KICKING

Hubble reveals most distant planets yet

Last Updated: Wednesday, October 4, 2006 | 7:21 PM ET
CBC News

Astronomers have discovered the farthest planets from Earth yet found, including one with a year as short as 10 hours — the fastest known.

Using the Hubble space telescope to peer deeply into the centre of the galaxy, the scientists found as many as 16 planetary candidates, they said at a news conference in Washington, D.C., on Wednesday.

The findings were published in the journal Nature.

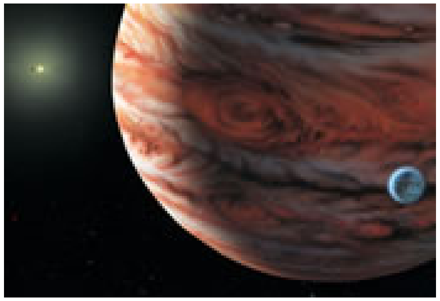
Looking into a part of the Milky Way known as the galactic bulge, 26,000 light years from Earth, Kailash Sahu and his team of astronomers confirmed they had found two planets, with at least seven more candidates that they said should be planets.

The bodies are about 10 times farther away from Earth than any planet previously detected.

A light year is the distance light travels in one year, or about 9.46 trillion kilometres.

- From *Nature* Oct 2006. Hubble was reborn twice and exoplanet discoveries have become quotidian.

- There were **228** listed at `exoplanets.org` in March 09 and **432** a year later, **563** as of 22/6/11 and **750** confirmed on 6/12/13. (More according to Kepler. There is an iPad Exoplanet app.)



- How reliable are these determinations (velocity, imaging, transiting, timing, micro-lensing)? The one above has been withdrawn!

5 Facts About Kepler (launch March 6)

-- Kepler is the world's first mission with the ability to find true Earth analogs -- planets that orbit stars like our sun in the "habitable zone." The habitable zone is the region around a star where the temperature is just right for water -- an essential ingredient for life as we know it -- to pool on a planet's surface.

-- By the end of Kepler's three-and-one-half-year mission, it will give us a good idea of how common or rare other Earths are in our Milky Way galaxy. This will be an important step in answering the age-old question: Are we alone?

-- Kepler detects planets by looking for periodic dips in the brightness of stars. Some planets pass in front of their stars as seen from our point of view on Earth; when they do, they cause their stars to dim slightly, an event Kepler can see.

-- Kepler has the largest camera ever launched into space, a 95-megapixel array of charge-coupled devices, or CCDs, as in everyday digital cameras.

-- Kepler's telescope is so powerful that, from its view up in space, it could see one person in a small town turning off a porch light at night.

NASA 05.03.2009

TWO RECONSTRUCTION APPROACHES

I. Error reduction of a nonsmooth objective (an 'entropy'): for fixed $\beta_m > 0$

⊙ we attempt to solve

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

– Many variations on this theme are possible.

II. Non-convex (in)feasibility problem: Given $\psi_m \neq 0$, define $\mathbb{Q}_0 \subset L^2$ convex, and

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

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

⊙ via an *alternating projection method*: e.g., for two sets A and B, **repeatedly compute**

$$x \to P_B(x) =: y \to P_A(y) =: x.$$

EXAMPLE 7. INVERSE SCATTERING

Central problem: determine the location and shape of buried objects from measurements of the *scattered field* after illuminating a region with a known *incident field.*

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

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

where $\mathcal{F} \to X$ is a compact linear operator constructed from the observed data, and $\varphi_z \in X$ is a known function parameterized by z [**BLu**].

- $\mathcal{F}$ 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

$$\|g_{z,\alpha}\| \to \infty \quad \text{as} \quad \alpha \to 0$$

  if and only if z is a point outside the scatterer.

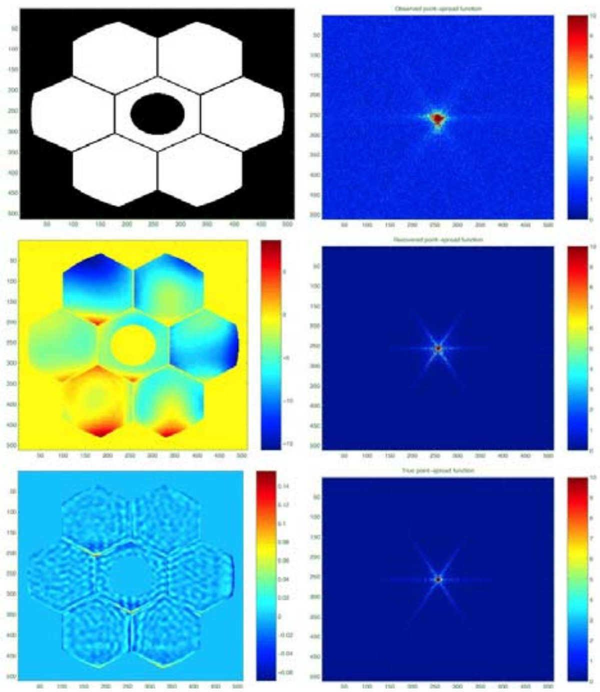
- **An important open problem** is to determine behavior of regularized solutions $g_{z,\alpha}$ under different regularization strategies.

  – In other words, when can these techniques fail? (Ongoing work with Russell Luke [**BLu**]: also in *Experimental Math in Action*, AKP, 2007.)

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***)*

A SAMPLE RECONSTRUCTION (via I)

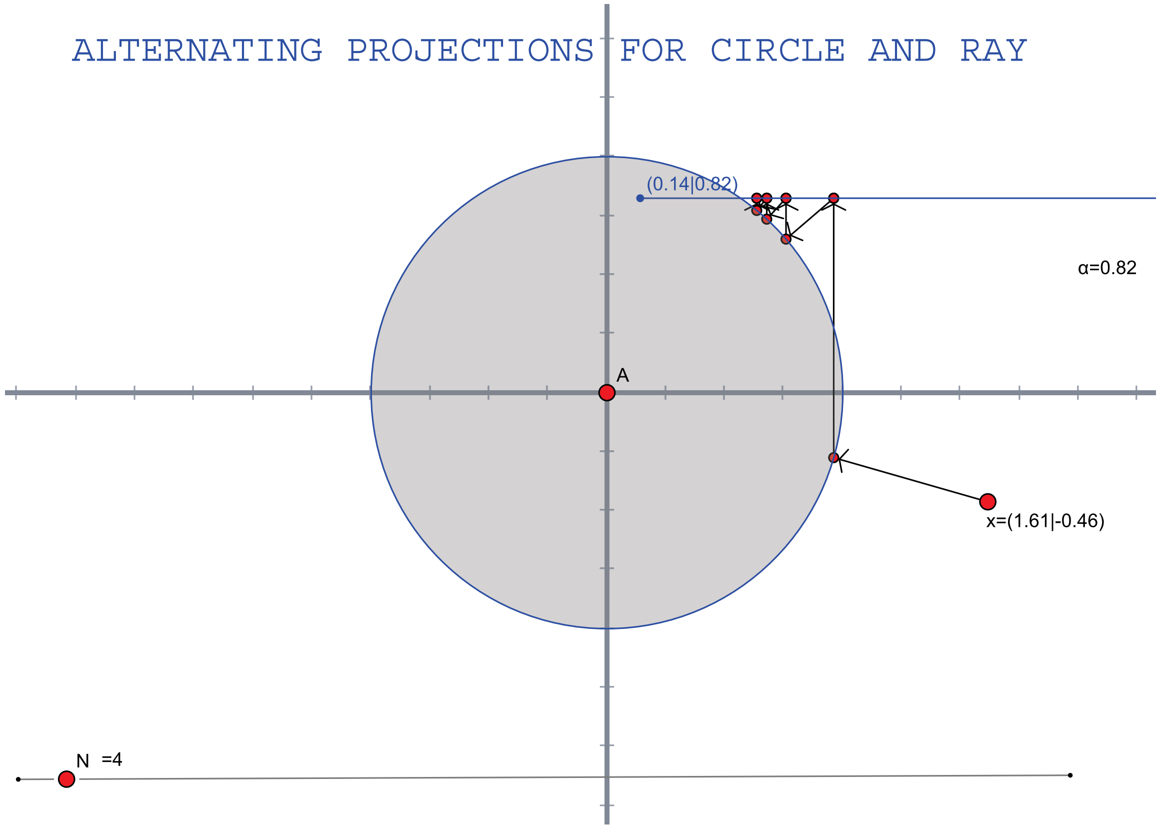
The object and its spectrum



Top row: data
Middle: reconstruction

Bottom: truth and error

ALTERNATING PROJECTIONS



The **alternating projection method** — discovered by Schwarz, Wiener, Von Neumann, ... — is *fairly* well understood when all sets are convex.

- If $A \cap B \neq \emptyset$ and A, B are closed convex then weak convergence (only 2002) is assured—von Neumann (1933) in norm for subspaces, Bregman (1965).

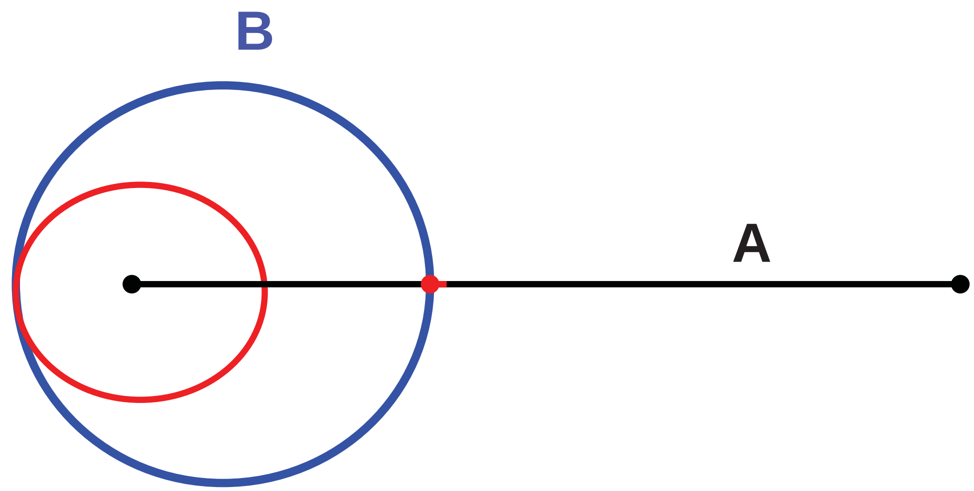
- First shown that norm convergence can fail by Hundal (2002) – but only for an 'artificial' example.

II: NON-CONVEX PROJECTION CAN FAIL

QUESTION. If A is finite codimension, closed and affine, B is the nonnegative cone in $\ell^2(N)$ and $A \cap B \neq \emptyset$, **is the method norm convergent**?

Consider the **alternating projection method** to find the unique red point on the **line-segment A** (convex) and the blue **circle B** (non-convex).

- The **method is 'myopic'**.

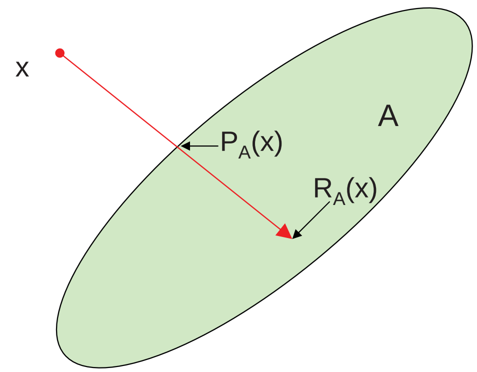


• Starting on line-segment outside *red circle*, we converge to unique feasible solution.

• Starting inside the red circle leads to a period-two locally ‘least-distance’ solution.

THE PROJECTION METHOD OF CHOICE

- For optical abberation correction this is the **alternating projection** method: $x \to P_A\,(P_B(x))$



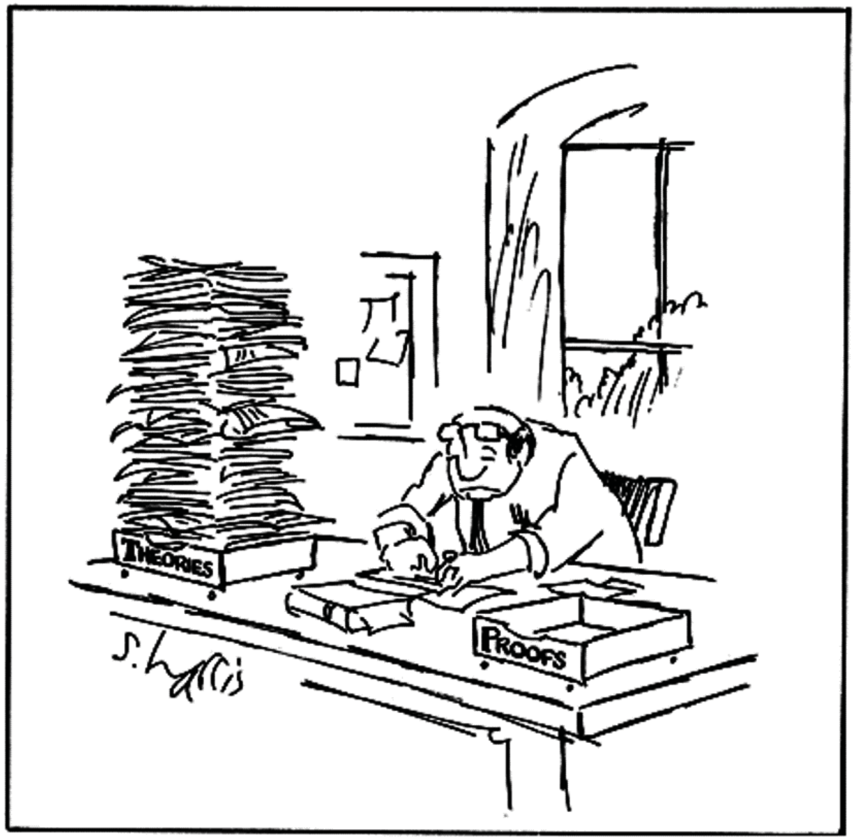
- For crystallography it is better to use (**HIO**) **over-relax and average**: *reflect* to $R_A(x) := 2\,P_A(x) - x$ and use

$$x \to \frac{x + R_A\,(R_B(x))}{2}$$

Both parallelize neatly: $A :=$diag, $B := \prod_i B_i$.

Both are non-expansive *in the convex case*.

Both need new theory *in 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 Luke* 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.

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

Bauschke-Combettes-Luke (JMAA, 2004) showed HIO, *Lions-Mercier* (1979), *Douglas-Rachford* (1959), *Feinup* (1982), and *divide-and-concur* coincide.

- When $u(t) \geq 0$ is imposed, Feinup's method no longer coincides, and DR ('HPR') is still better.

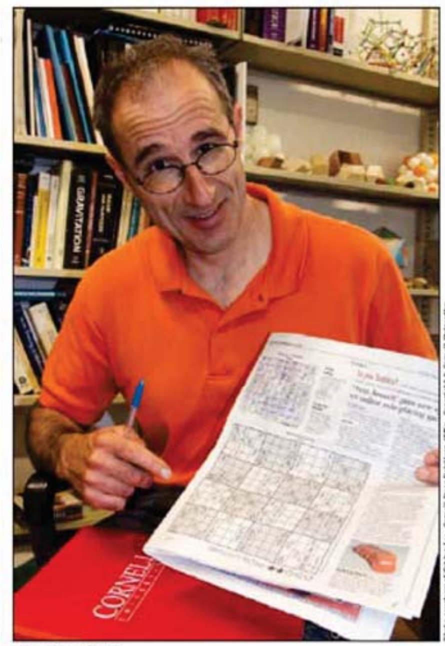
- JMB-Tam (2013) have found a promising cyclic reflection method.

ELSER, QUEENS and SUDOKU

2006 Veit Elser, see [**E1**] and [**E2**], at Cornell has had huge success (and press) using divide-and-concur onprotein folding, sphere-packing, 3SAT, **Sudoku** ($\mathbb{R}^{2916}$), and more.

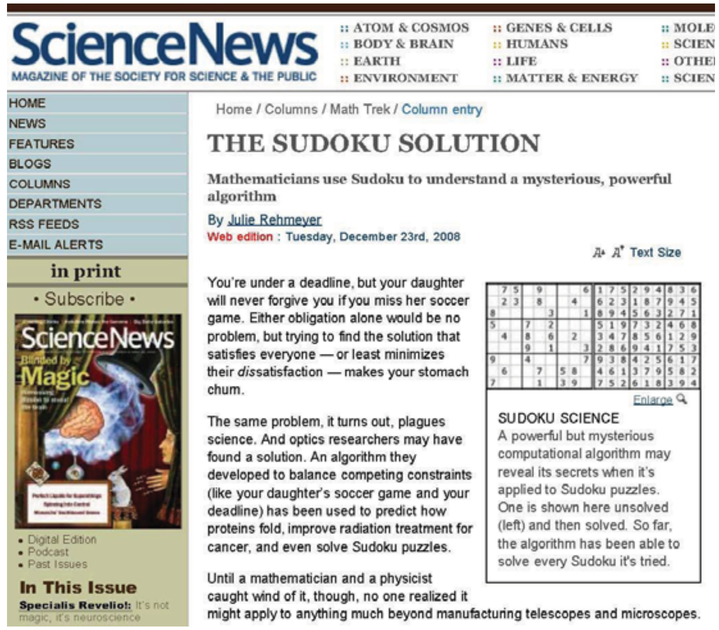
Given a partially completed grid, fill it so that each column, each row, and each of the nine 3×3 regions contains the digits from 1 to 9 only once.

	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	



Veit Elser, Ph.D.

2008 Bauschke and Schaad likewise study **Eight queens problem** ($\mathbb{R}^{256}$) and image-retrieval (*Science News*, 08).


ScienceNews
MAGAZINE OF THE SOCIETY FOR SCIENCE & THE PUBLIC

:: ATOM & COSMOS :: BODY & BRAIN :: EARTH :: ENVIRONMENT :: GENES & CELLS :: HUMANS :: LIFE :: MATTER & ENERGY

HOME
NEWS
FEATURES
BLOGS
COLUMNS
DEPARTMENTS
RSS FEEDS
E-MAIL ALERTS

in print
• Subscribe •

- Digital Edition
- Podcast
- Past Issues

In This Issue
Specialis Revelio!: It's not magic, it's neuroscience

Home / Columns / Math Trek / Column entry

THE SUDOKU SOLUTION

Mathematicians use Sudoku to understand a mysterious, powerful algorithm

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

Text Size

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.

Enlarge

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.

This success (a.e.?) is not seen with alternating projections and cries out for explanation. Brailey Sims and I [**BS**] and then Fran Aragon and I [**AB**] have made some progress, as follows:

FINIS: DOUGLAS-RACHFORD IN THE SPHERE

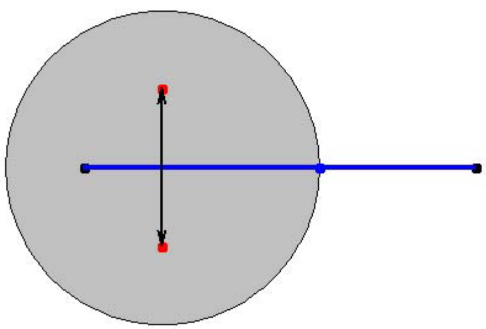
Dynamics for B the unit circle and A the blue line at height $\alpha \geq 0$ are already fascinating. Steps are for

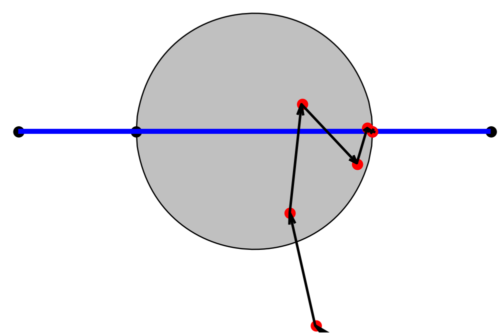
$$T := \frac{I + R_A \circ R_B}{2}.$$

- With θ_n the argument this becomes set

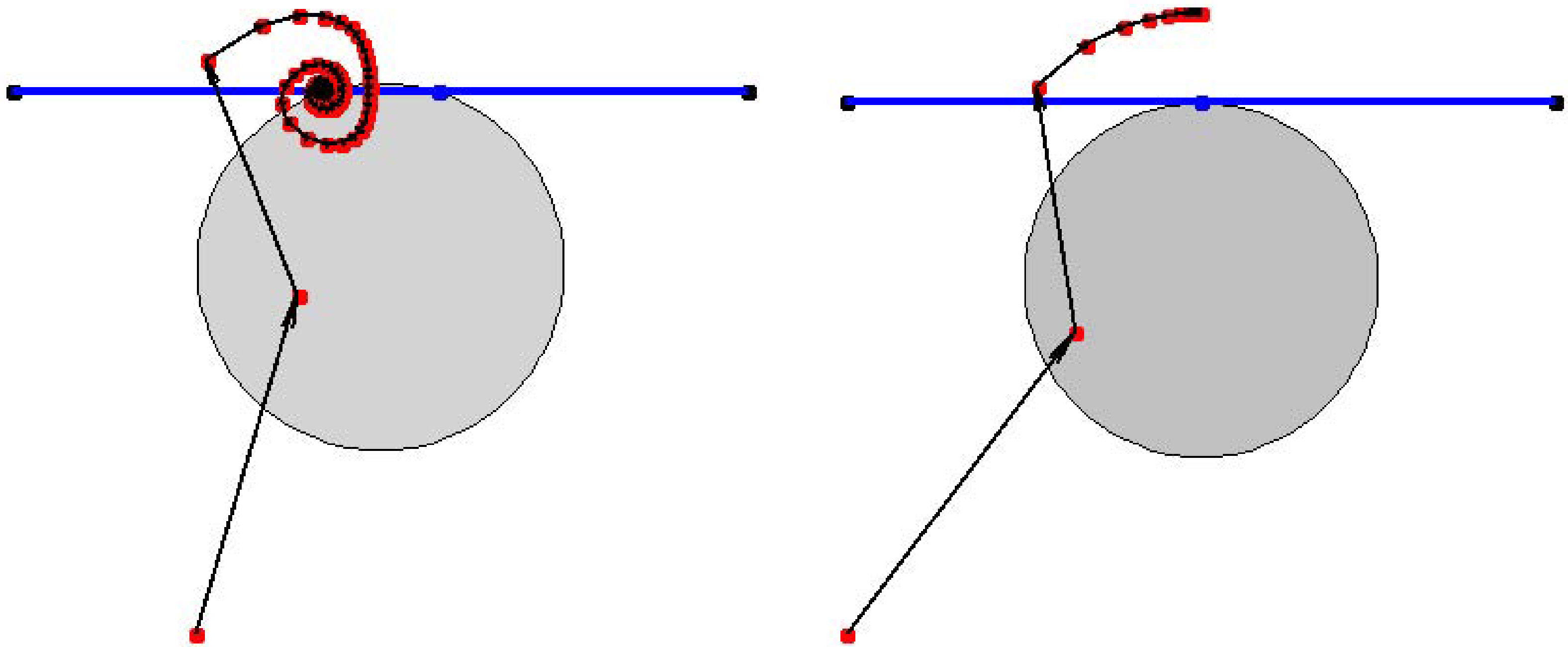
$$\boxed{x_{n+1} := \cos\theta_n, y_{n+1} := y_n + \alpha - \sin\theta_n.}$$

$0 \leq \alpha \leq 1$: converges ('globally' ('13) & locally exponentially asymptotically ('11)) **iff** start off y-axis ('chaos'):





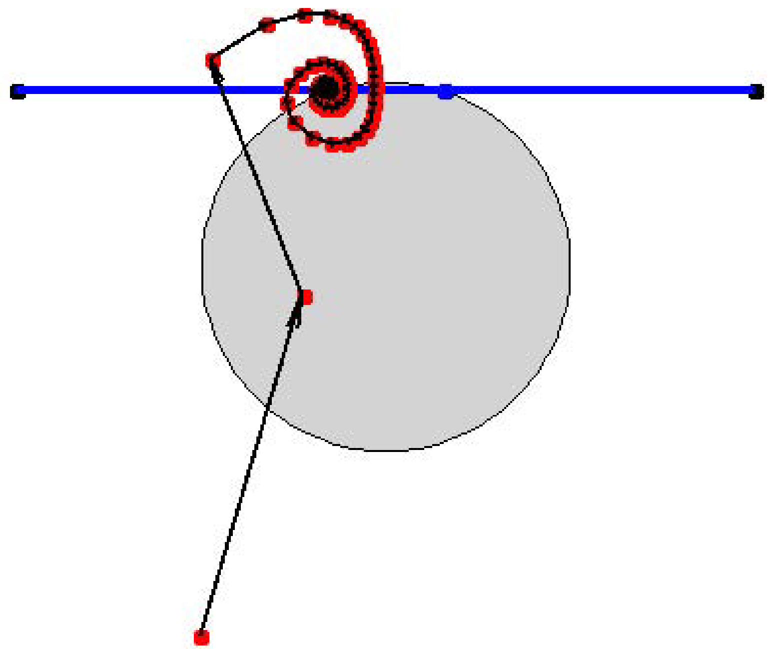
$\alpha > 1 \Rightarrow y \to \infty$, while $\alpha = 0.95$ $(0 < \alpha < 1)$ and $\alpha = 1$ respectively produce:



- The result remains valid for a **sphere** and **any affine manifold** in Euclidean space.

GLOBAL CONVERGENCE

A lot of hard work proved the result in Figure 5 [**AB**]:



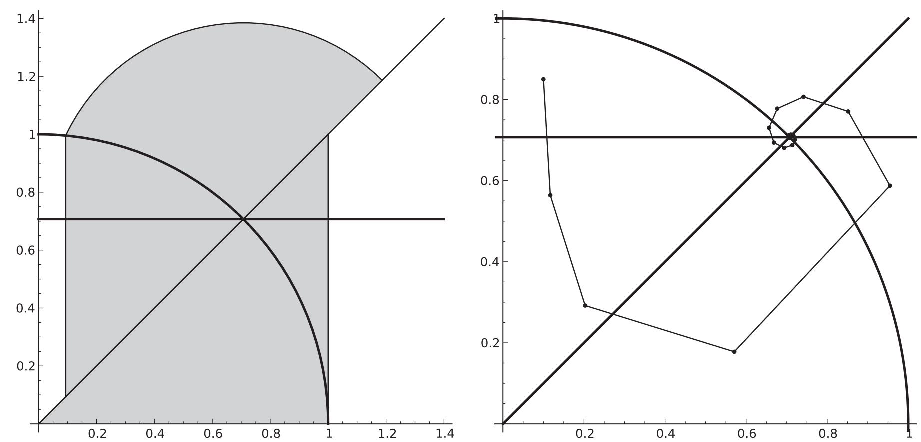
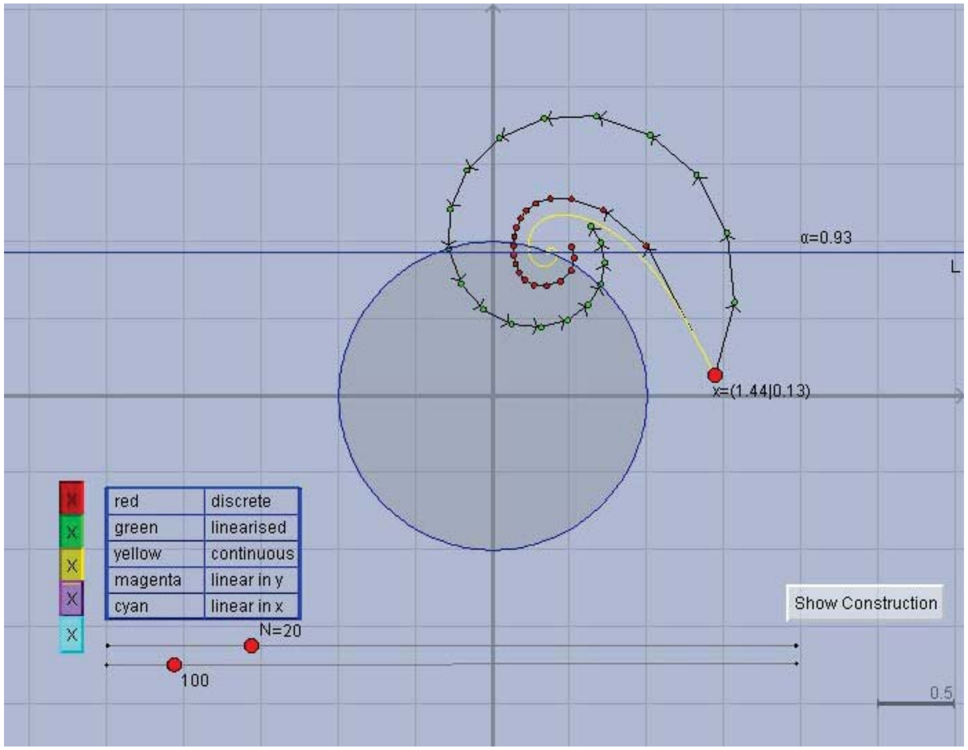


Figure 5: The picture in the left shows the regions of convergence in Theorem 2.1 for the Douglas-Rachford algorithm. The picture in the right illustrates an example of a convergent sequence generated by the algorithm.

DYNAMIC GEOMETRY



- I finish with a *Cinderella* demo based on the recent work with Brailey Sims [**BS**].

The applets are at:

www.carma.newcastle.edu.au/~jb616/composite.html

www.carma.newcastle.edu.au/~jb616/expansion.html

OTHER REFERENCES

AB F. Aragon and J.M. Borwein, "Global convergence of a non-convex Douglas-Rachford iteration." Preprint 2012.

BCM 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.

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

BL2 J. M. Borwein and A. S. Lewis, "Duality relationships for entropy–like minimization problems," *SIAM Control and Optim.*, **29** (1991), 325–338.

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

BN1 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.

BN2 J. Borwein, M. Limber and D. Noll, "Fast heuristic methods for function reconstruction using derivative information," *App. Anal.*, **58** (1995), 241–261.

BS 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.

E2 Gravel, S. and Elser, V., "Divide and concur: A general approach constraint satisfaction," preprint, 2008, `http://arxiv.org/abs/0801.0222v1`.

D 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.

E1 Elser, V., Rankenburg, I., and Thibault, P., "Searching with iterated maps," *Proceedings of the National Academy of Sciences* **104** (2007), 418–423.

PHBH Julia Piantadosi, Phil Howlett, Jonathan Borwein and John Henstridge, "Generation of simulated rainfall data at different time-scales." *Numerical Algebra, Control and Optimization*. **2** (2012), 233–256.

PHB Julia Piantadosi, Phil Howlett and Jonathan Borwein, "Modelling and simulation of seasonal rainfall." MODSIM/ASOR 2013, Adelaide, December 2013.