Entropy in Seismology:

Can we extract velocity from wavefield entropy?

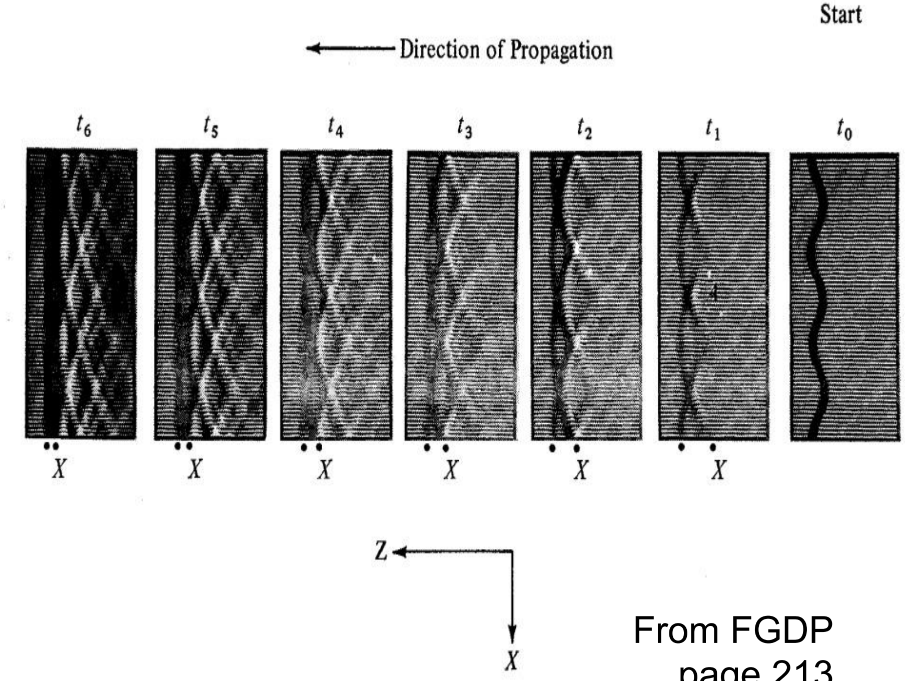
Data space residuals want maximum entropy, minimum sparsity.
Blocky models want minimum entropy, maximum sparsity.
(very different from the L1-norm).

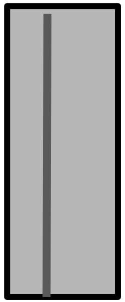
- Remember the “wavefront healing” movie.
- Like tomography but with overwhelming multipathing.
- Complexity grows as entropy grows.
- We will see probability based inverse theory.

- Remember the “wavefront healing” movie.
- Like tomography but with overwhelming multipathing.
- Complexity grows as entropy grows.
- We will see probability based inverse theory.

Multipathing

If going backwards the complexity decreases.

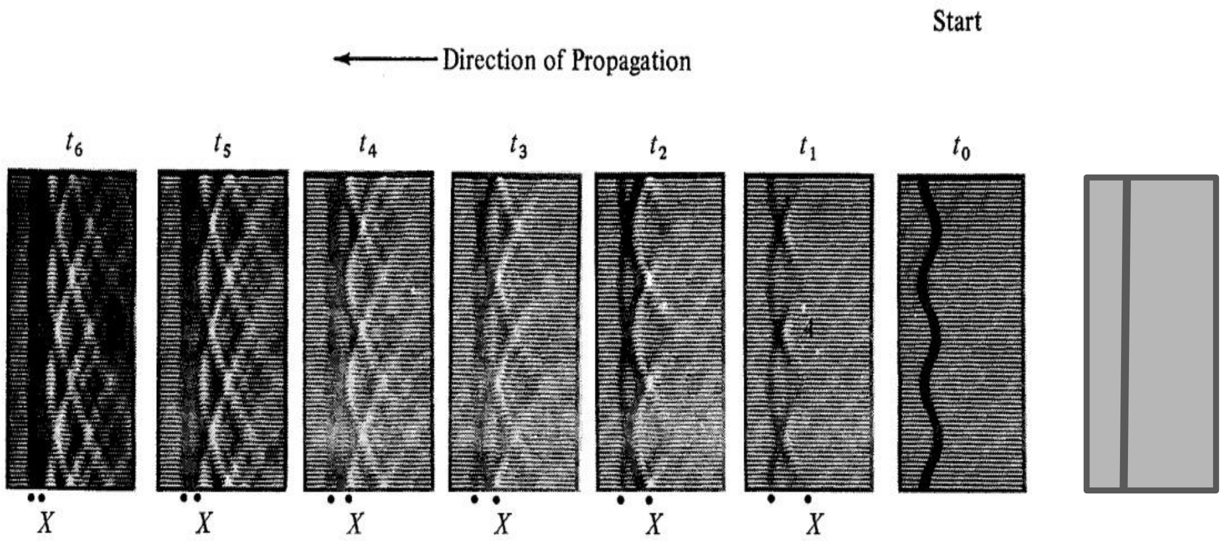


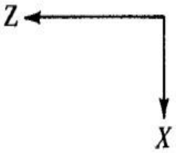


From FGDP page 213

Multipathing

If going backwards the complexity decreases.





From FGDP
page 213

If you have SEPLIB at SEP, view this

```
tube ~prof/bei/fdm/Fig/heal.v3
```

If you don’t have that, slide back and forth over this range (1:00)-(3:00)
http://sep.stanford.edu/sep/jon/WavefrontHealing.mp4

Background

- Seismic Q: Caused by (1) heat flow, (2) inner-bed multiples, or (3) extravagant multipathing?
- Do any of you have computer code to model constant-Q propagation?

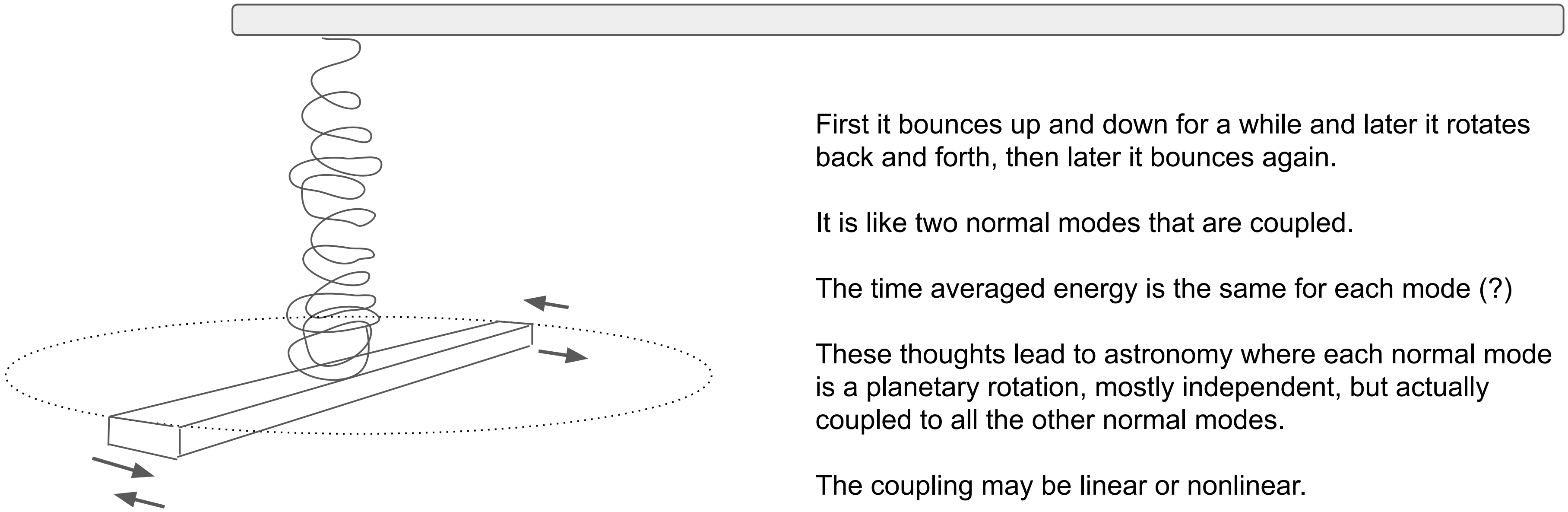
- We seismologists think we know the heart of statistics, but
  statistical Mechanics gives rise to Thermodynamics, and
  you have forgotten how a propane refrigerator works.

- Energy is always conserved, but it is made up of kinetic, potential, and **thermal.**

- Entropy increases, but what is it?
- Entropy and Energy are **extrinsic**.   Temperature is **intrinsic**.
- The more the volume, the more energy and more entropy.

- To work backwards from nature we should *minimize* the ratio Entropy/Energy
- Not the same as minimizing *L1/L2* but it does have a sparseness goal for model space.

A bar suspended from its midpoint by a spring.
It can (1) bounce up and down or (2) rotate.



First it bounces up and down for a while and later it rotates back and forth, then later it bounces again.

It is like two normal modes that are coupled.

The time averaged energy is the same for each mode (?)

These thoughts lead to astronomy where each normal mode is a planetary rotation, mostly independent, but actually coupled to all the other normal modes.

The coupling may be linear or nonlinear.

You have a rock in a vacuum suspended by a thread.
The rock has infinite Q.
You strike the rock.
Internally, P waves convert to S, and S converts to P.
Perform a time average of

Energy in P
Energy in S

What is the ratio of the two energies?

Old theoreticians published a paper on that.
I didn’t understand it, but I do remember the two energies were different.

Look up “equipartition in an elastic solid”

Specific heat of monatomic and diatomic gases

(Specific heat == the ratio of temperature change to energy change)

Randomly colliding monatomic gas molecules have **three** degrees of freedom meaning that each atom has momentum in the three (x,y,z) coordinates.

Diatomic gases like H2, O2, N2 have **three** degrees from **translation** and, **three** degrees of freedom from **vibration** and two degrees from **two** possible **rotation** angles.

Diatomic gases at very low temperature (because of quantum effects) can no longer vibrate so the specific heat loses **three** degrees of freedom.

Butterfly effect

Atmospheric equations are ultrasensitive to precision...
even when the PDEs are simplified to several non-linear ODEs

This discovered by Ed Lorenz at MIT
(while I did my MS degree one flight above him.)
This discovery limits the prospects of weather forecasting.
Some say this discovery led to **Chaos Theory.**

((Iterating towards minimum entropy might be fundamentally limited.))

A Sparsity Example

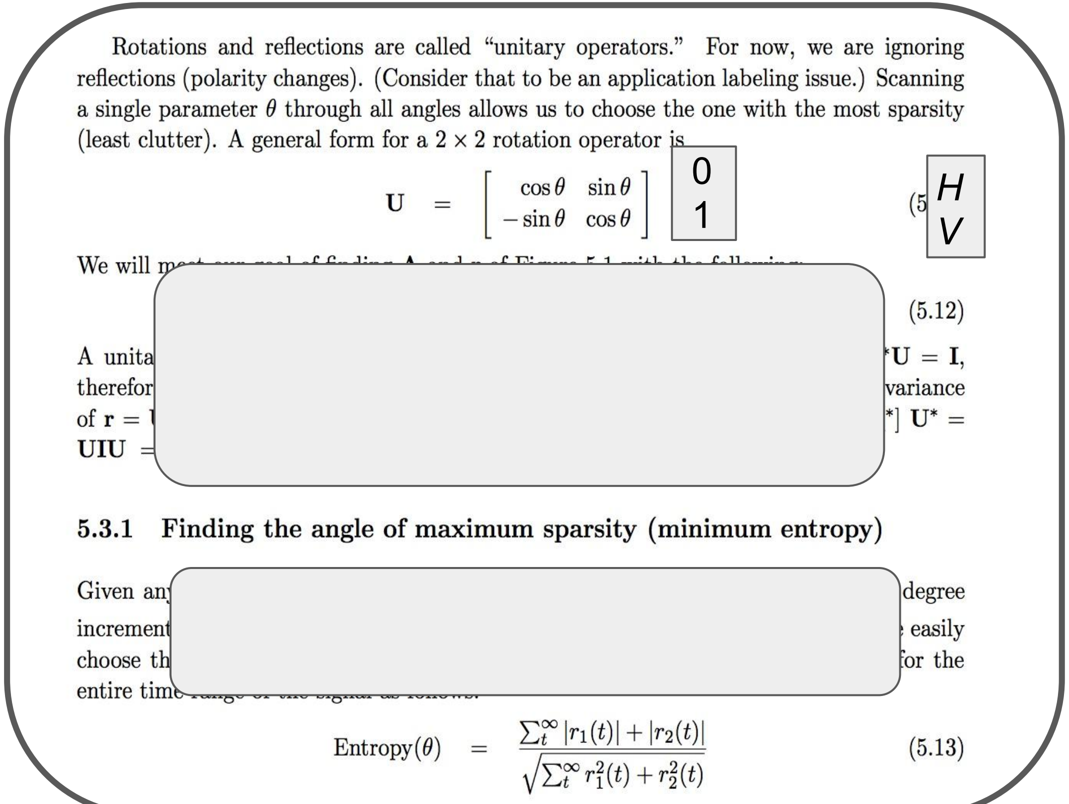
A pragmatic way to achieve sparsity is to minimize L1/L2, a pseudo entropy.

In my booklet DFNS, the chapter on multichannel spectral factorization (Kaiwen) faces the problem of scaling the PEF. Any unitary scale factor does not change the spectrum (conserves energy).

For multichannel signals, the scale factor is a unitary matrix.

The vector (0,1) is more sparse than the vector (sin,cos). At 45 degrees

| sin | + | cos | = sqrt(2)
sqrt(2)/2+sqrt(2)/2 = sqrt(2) > 1



Rotations and reflections are called "unitary operators." For now, we are ignoring reflections (polarity changes). (Consider that to be an application labeling issue.) Scanning a single parameter θ through all angles allows us to choose the one with the most sparsity (least clutter). A general form for a 2 × 2 rotation operator is

$$\mathbf{U} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \quad (5$$

We will m

(5.12)

A unita ... $^*\mathbf{U} = \mathbf{I}$,
therefor ... variance
of $\mathbf{r} =$ ... $^*]\ \mathbf{U}^* =$
$\mathbf{UIU} =$

5.3.1 Finding the angle of maximum sparsity (minimum entropy)

Given any ... degree
increment ... easily
choose th ... for the
entire time range of the signal as follows:

$$\text{Entropy}(\theta) = \frac{\sum_t^\infty |r_1(t)| + |r_2(t)|}{\sqrt{\sum_t^\infty r_1^2(t) + r_2^2(t)}} \quad (5.13)$$

Internet: Entropy is “the lack of order or predictability.”

Wikipedia: In statistical mechanics, entropy is an extensive property of a thermodynamic system. It is closely related to the number Ω of microscopic configurations that are consistent with the macroscopic quantities that characterize the system.

Which convex function shall we play with?

L1/L2 is cheapo entropy.

From where comes p log(p) ?

grad (p log p) = 1 + log p

grad(log(p))=1/p, reminds of AGC and PEF.

Which: p=energy or p=probability ?

Hyperbolic penalty function

More formal documentation:

To utilize all brightness levels equally see SEP 152 in 2014 or “Jensen inequalities” article and video at http://sep.stanford.edu/sep/jon/

Physics 7-dimensional space

position, momentum
m(t, x, y, z, px, py, pz)
t, space, stepout

AGC and PEF boost the entropy.

Nuclear Physics has also spin.
We have vector-valued wave signals, images, volumes.

My wavefront healing images eventually fill space and have all dips,
but each point in space has only one dip.

Equations of Quantum Physics are not physical laws, they are probability estimates.

If you have SEPLIB at SEP, view this

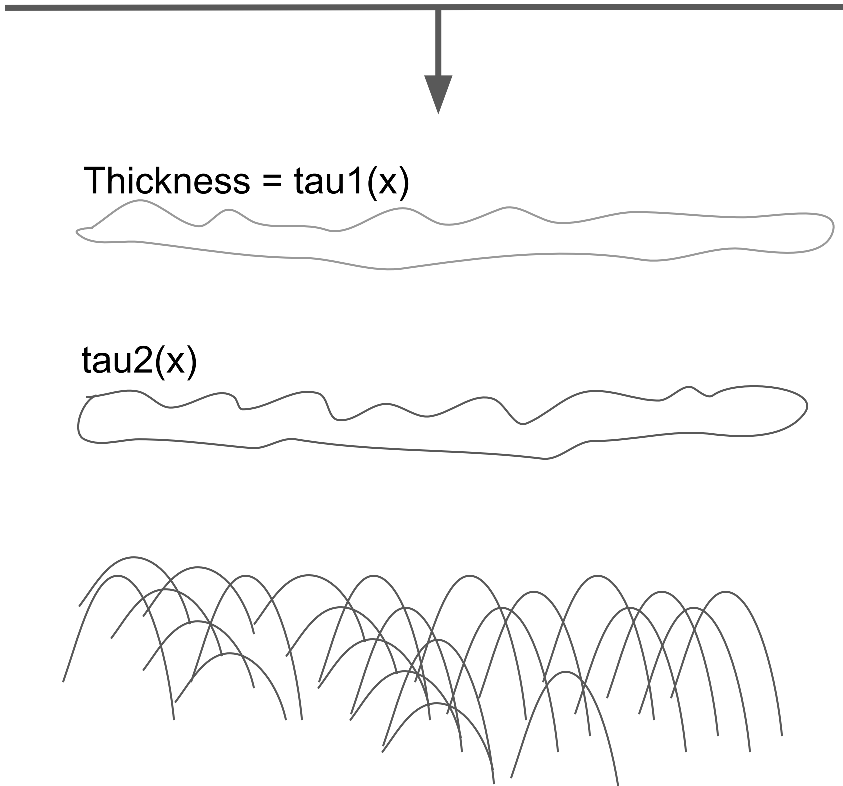
```
tube ~prof/bei/fdm/Fig/heal.v3
```

If you don’t have it, scan parts of this video

http://sep.stanford.edu/sep/jon/WavefrontHealing.mp4

I want a toy to play with a two-lens model.

Thin lens is a simple time shift.



Plane wave
Impulse or white noise.

Steps in exp (i k^2/omega dz)

First thin lens = exp(i omega tau(x) dz)

(want zero tau response between lenses)

Second thin lens

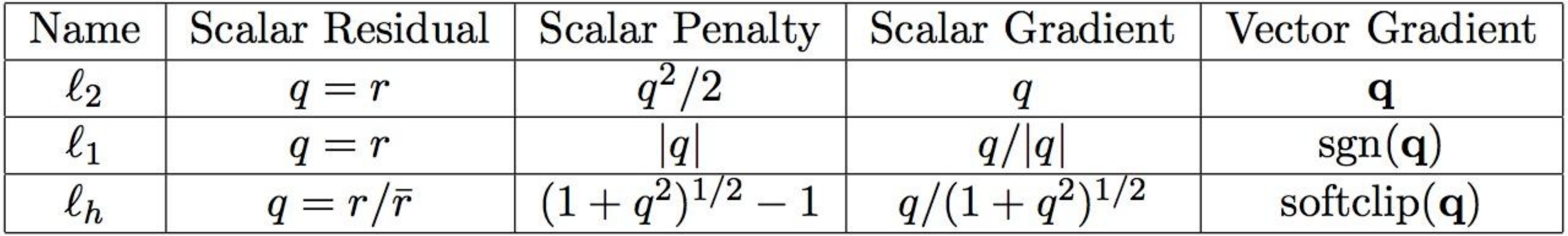
Vision: How to estimate tau(x,z)?
Let us back propagate while minimizing the entropy.

IID = Independent and Identically Distributed
IID = PEF & AGC seem to maximize entropy

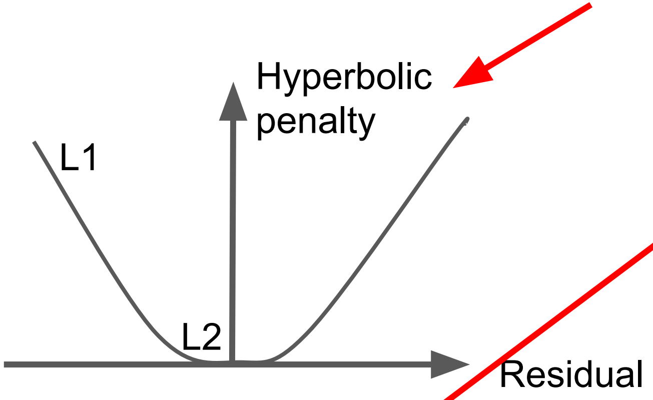
- Entropy = Negentropy = Sparseness = many amplitudes vanish
many dips vanish

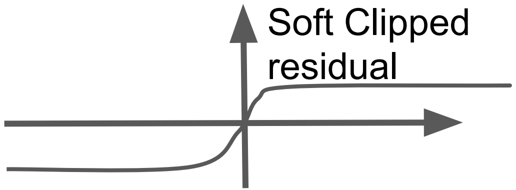
“Jensen inequalities” article and video at   http://sep.stanford.edu/sep/jon/

(From the (1+epsilon) norm I got this idea in 1977 SEP 13
http://sep.stanford.edu/data/media/public/oldreports/sep13/13_01.pdf
Stew guided me to the Jensen inequality literature in my book PVI.
I tested an application to replace our “pclip” in 2014
but it never got installed at SEP for routine work.)



Name	Scalar Residual	Scalar Penalty	Scalar Gradient	Vector Gradient
ℓ_2	$q = r$	$q^2/2$	q	$\mathbf{q}$
ℓ_1	$q = r$	$\|q\|$	$q/\|q\|$	$\mathrm{sgn}(\mathbf{q})$
ℓ_h	$q = r/\bar{r}$	$(1+q^2)^{1/2} - 1$	$q/(1+q^2)^{1/2}$	$\mathrm{softclip}(\mathbf{q})$





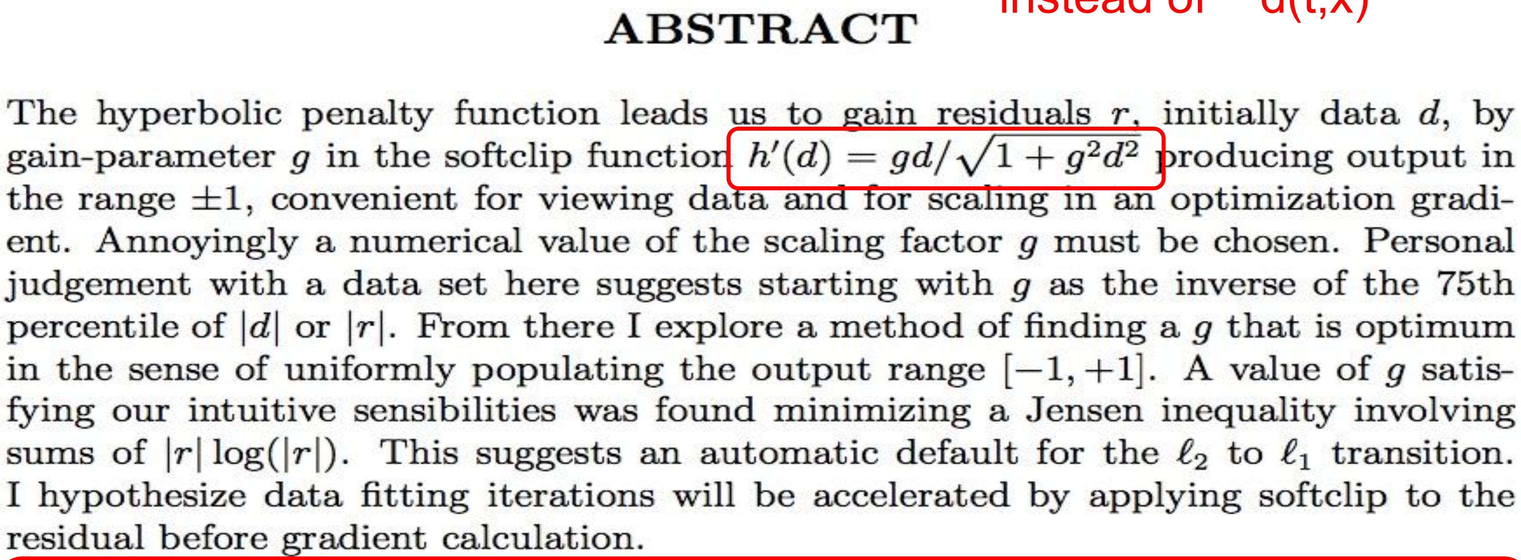
Define g = 1/bar(r) so
g is a gain applied to the residual
q = g r

g defines the L1 to L2 transition.

AUTOMATIC DEFAULT FOR HYPERBOLIC SOFTCLIP

SEP 152

Idea: Plot h'(d(t,x))
instead of d(t,x)

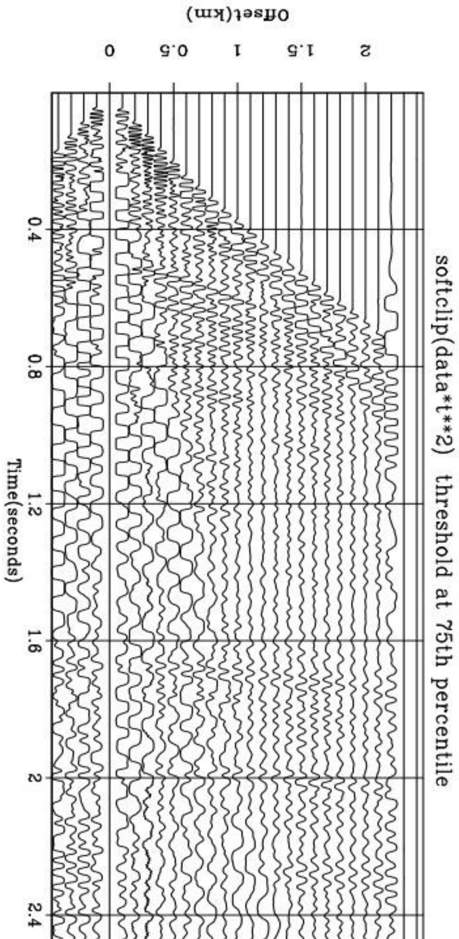


ABSTRACT

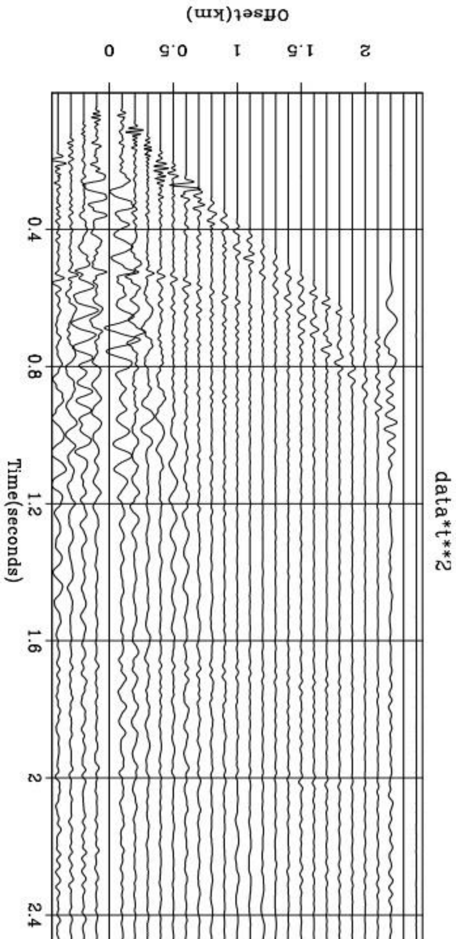
The hyperbolic penalty function leads us to gain residuals r, initially data d, by gain-parameter g in the softclip function $\boxed{h'(d) = gd/\sqrt{1+g^2d^2}}$ producing output in the range ±1, convenient for viewing data and for scaling in an optimization gradient. Annoyingly a numerical value of the scaling factor g must be chosen. Personal judgement with a data set here suggests starting with g as the inverse of the 75th percentile of $|d|$ or $|r|$. From there I explore a method of finding a g that is optimum in the sense of uniformly populating the output range $[-1,+1]$. A value of g satisfying our intuitive sensibilities was found minimizing a Jensen inequality involving sums of $|r|\log(|r|)$. This suggests an automatic default for the ℓ_2 to ℓ_1 transition. I hypothesize data fitting iterations will be accelerated by applying softclip to the residual before gradient calculation.

We will learn that finding g amounts to letting the residuals choose the penalty function!
(Mostly people choose L1 or L2.)

q=SoftClipped(g,**r**)



Input **r**



Softclip applied to field data.
It is a nonlinear data-value stretch
into the range -1 to +1.

Softclip is the derivative of the hyperbolic penalty function.

Softclip has a parameter g
that chooses the L1/L2 threshold.
Here I chose g at 75th percentile.

I made this choice of g subjectively. Then I wondered how to choose g objectively.

Here's how I did it:
"The best g tries to use all brightness levels in equal amounts."

Softest

Best

Hardest

SEP's default pclip

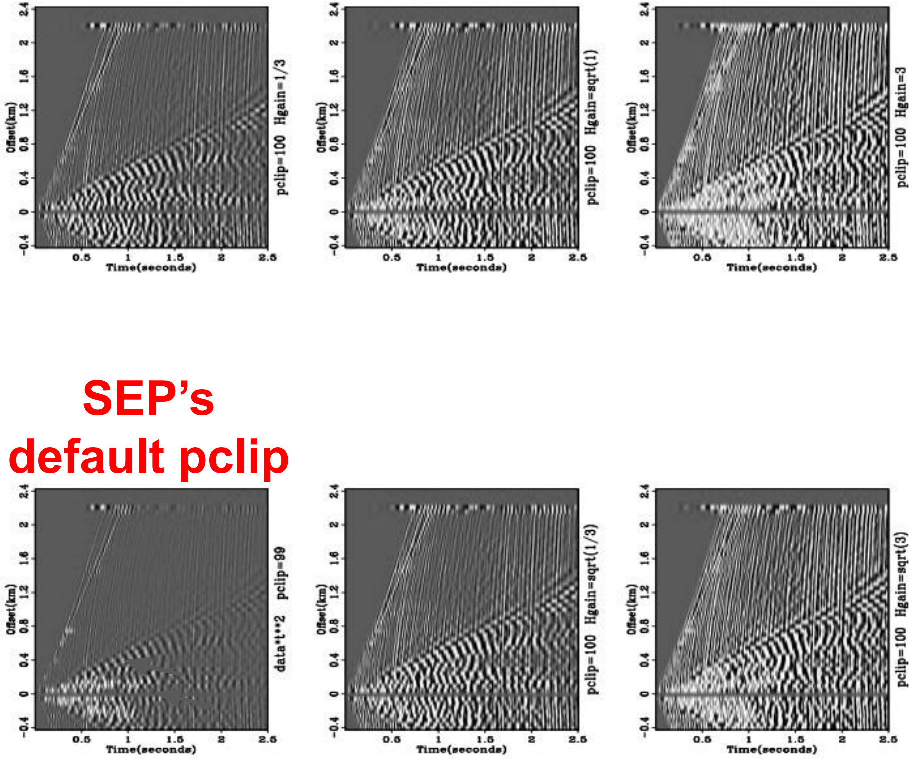
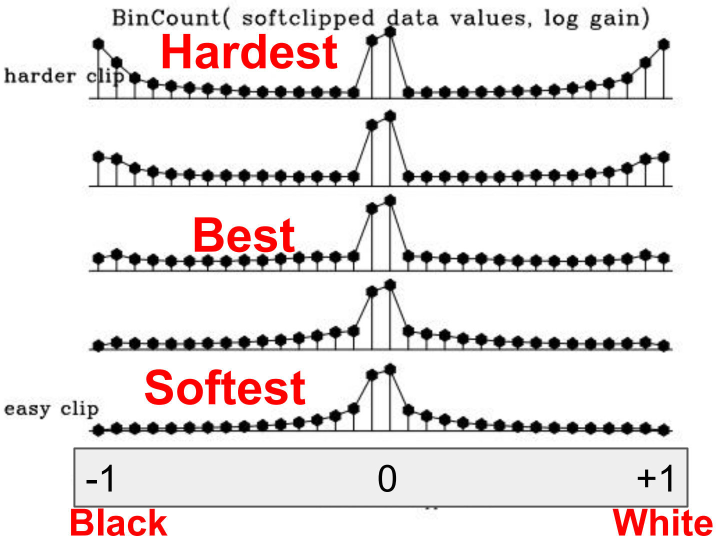


Figure 2: Top left is the default plot in our laboratory. You see the early part of a shot

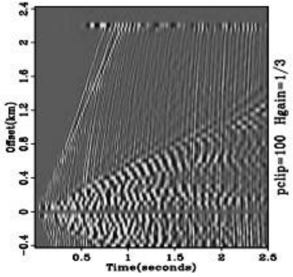
Bin count



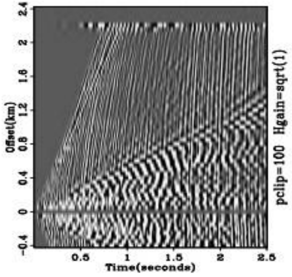
Value of softclipped data

AGC flattens residuals in physical space
PEF flattens residuals in dip space
Softclip flattens residuals in probability --- That’s maximum entropy, I think.

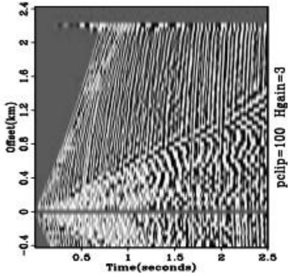
Softest



Best

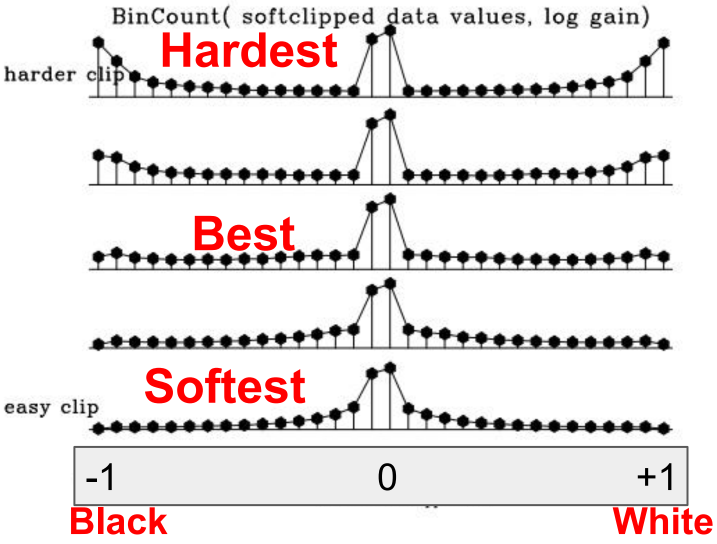


Hardest



plot in our laboratory. You see the early part of a shot

Bin count



Value of softclipped data

Automatic Default for Hyperbolic Softclip (SEP 2014)

(a.k.a. My Jensen paper)

Here is how I found the best g, the location for the L1 to L2 transition.
Let r = residual or data.
Let q = softclip(g r)
Find p=bins containing counts of q values.
Evaluate many g values to best uniformize the counts in the bins. (I will soon define “best uniform”.)

Allowing negative subscripts, take p(-15:15) to be 31 bins.

```
Given:  q(i) = softclip( g, r(i))           #    -1. < q(i) < +1.   by the definition of softclip.
Do i= 1, 80000 {
        p(15.9*q(i)) += 1                   # This accumulates to the 31 probability density bins.
        }
```

WE WILL SEE THAT:

The best g will **minimize** the Jensen inequality of the convex function (**p log p**).
The Jensen inequality will define the intrinsic **Entropy**.
Minimum Entropy attempts to cluster the values of q.

To find this paper and video, open http://sep.stanford.edu/sep/jon/ and search for Jensen.

Amplitude binning is similar to operators we are familiar with.

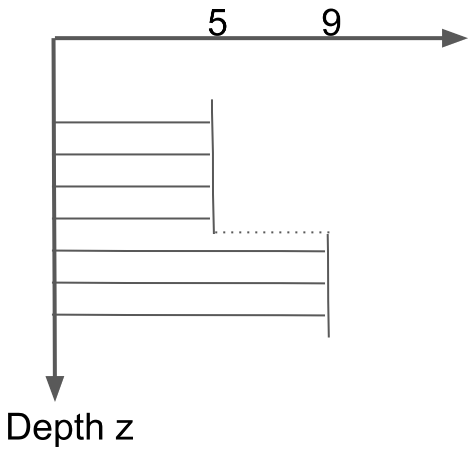
1. Amplitude bin counting
2. Scatter-gather
3. AGC
4. TV-Decon

These operators have adjoints that are easy to code, but
These operators tend to be non-linear.
These operators change if you change the data upon which they were built, but
they become quasi-linear for large statistical windows and small g values.

A conceptual probability function estimate:

Take all the non-sorted data values and replace each by the index of its sorted counterpart.
This is extremely nonlinear but it should resemble the hyperbolic stretch of best g.

(This page might be baloney.)
Now that we understand binning,
we might suspect that a blocky model has sparse bins.



Velocity v(z) = (5,5,5,5,9,9,9)

P = bin(i) = (0,0,0,0,4,0,0,0,3,0,0,0,0)

Large entropy would fill bins about equally.
Small entropy fills bins sparsely, I am guessing.

Want large entropy in residual space. (like AGC & PDF)
Want small entropy in model space?
Entropy is unchanged when layers are intermingled.
Good for a water bucket full of marbles: Only two densities, water and marbles. Sand and shale.

JENSEN INEQUALITY BASICS

This is a clarifying revision of material that appeared earlier in SEP 37 and reprinted in PVI.

Let f be a function with a positive second derivative. Such a function is called "convex" and satisfies the **inequality**

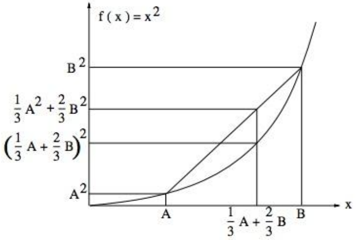
$$\frac{f(a)\ +\ f(b)}{2} \ - \ f\left(\frac{a+b}{2}\right) \quad \geq \quad 0 \qquad (1)$$

The inequality (1) relates the average of the function to a function of the average. The average can be weighted, for example,

$$\frac{1}{3}\,f(a) \ + \ \frac{2}{3}\,f(b) \ - \ f\left(\frac{1}{3}a+\frac{2}{3}b\right) \quad \geq \quad 0 \qquad (2)$$

Figure 4 is a graphical interpretation of equation (2) for the function $f = x^2$. There

Figure 4: Sketch of $y = x^2$ for interpreting equation (2).



is nothing special about $f = x^2$, except that it is convex. Given three numbers a, b, and c, the inequality (2) can first be applied to a and b, then c with the average of a and b. Thus, recursively, an inequality like (2) can be built for a weighted average of three or more numbers. Define weights $w_j \geq 0$ that are normalized ($\sum_j w_j = 1$). The general result for $d^2f/dx^2 > 0$ is

$$F(p_j) \ = \ \sum_{j=1}^{N} w_j f(p_j) \ - \ f\left(\sum_{j=1}^{N} w_j p_j\right) \quad \geq \quad 0 \qquad (3)$$

$$F \ = \ \overline{f(p)} \ - \ f(\bar{p}) \ = \ \mathrm{E(f) - f(E)} \quad \geq \quad 0 \qquad (4)$$

If all the p_j are the same, say $\bar{p}$, then the two terms in (3) both become $f(\bar{p})$ so the inequality becomes an equality. Thus, minimizing F is like urging all the p_j to be identical. Equilibrium is when F is reduced to the smallest possible value which satisfies any constraints that may be applicable. An experimentalist naturally wonders which $f()$ is best for any particular application. Let's look at some.

Examples of Jensen inequalities

The most familiar example of a Jensen inequality occurs when the weights are all equal to $1/N$ and the convex function is $f(x) = x^2$. In this case the Jensen inequality $\overline{f(p)} - f(\bar{p}) \geq 0$ gives the familiar result that the mean of the squares exceeds the square of the mean:

$$Q \ = \ \frac{1}{N}\sum_{i=1}^{N} x_i^2 \ - \ \left(\frac{1}{N}\sum_{i=1}^{N} x_i\right)^2 \quad \geq \quad 0 \qquad (5)$$

In many applications the population consists of positive members only, so the function $f(p)$ need have a positive second derivative only for positive values of p. The function $f(p) = 1/p$ yields a Jensen inequality for the **harmonic mean**:

$$H \ = \ \sum \frac{w_i}{p_i} \ - \ \frac{1}{\sum w_i p_i} \quad \geq \quad 0 \qquad (6)$$

A more important case is the **geometric inequality**. Here $f(p) = -\ln(p)$, and

$$G \ = \ -\sum w_i \ln p_i \ + \ \ln \sum w_i p_i \quad \geq \quad 0 \qquad (7)$$

The more familiar form of the geometric inequality results from exponentiation and a choice of weights equal to $1/N$:

$$\frac{1}{N}\sum_{i=1}^{N} p_i \ - \ \prod_{i=1}^{N} p_i^{1/N} \quad \geq \quad 0 \qquad (8)$$

In other words, the product of square roots of two values is smaller than half the sum of the values.

The function $f(p) = p\ln(p)$ is also convex. That's not obvious, so let us check. First, $f' = 1 + \ln(p)$. Then $f'' = 1/p > 0$, so yes it is convex for $|p| > 0$. The average of the function minus the function of the average $\overline{f(p)} - f(\bar{p}) = E(f) - f(E) \geq 0$ is:

$$S_{\text{extrinsic}} \ = \ \sum w_i p_i \ln p_i \ - \ \left(\sum w_i p_i\right) \ \ln \sum w_i p_i \quad \geq \quad 0 \qquad (9)$$

$$S_{\text{intrinsic}} \ = \ \frac{\sum w_i p_i \ln p_i}{\sum w_i p_i} \ - \ \ln \sum w_i p_i \quad \geq \quad 0 \qquad (10)$$

This inequality is similar to what we may find in Physics and Information Theory. It might be exactly that, but they tend to use integrals instead of sums, so it is not easy to find it expressed in the "programmer ready form" there. No worries at $p = 0$. The logarithm diverges, but p is stronger so the product $p\ln(p)$ is zero.

On this slide p(i) is the count in the i-th bin. So it is the probability of the ordinate of the bin.

Bigger on later slide.

I tried this one for many values of g. The chosen g comes from minimum S.

Compare Jensen inequalities. Why choose p log(p) ?

(Applications might apply them to amplitudes, energies, probabilities, all, or ML.)

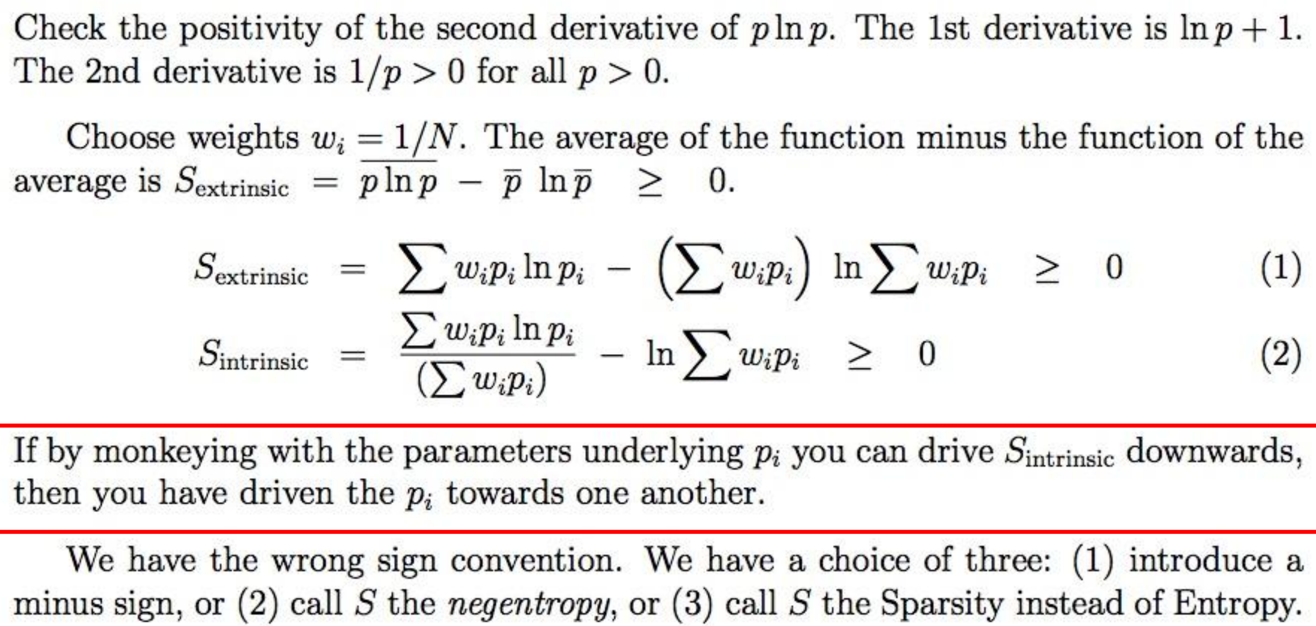
1. p log(p) allows p=0 elements. (Geometric and Harmonic blow up at p=0.)

2. p-squared is *very insensitive* to small p. (Arithmetic inequality)

3. p log(p) respects small p-values although it does attend more to large ones.

4. The Kulback-Leibler Divergence P log(P/Q) found in t-SNE of machine learning is a minor restatement of p log(p) where Q is a prior density.

Specify a thin lens with 200 yet-unknown coefficients, say tau(x).
Now we want the entropy gradient, its derivative by tau(x).
From data r(t,x); q=SoftClip |gr|; p(i) = probability(q) = bincount(q)



Check the positivity of the second derivative of $p\ln p$. The 1st derivative is $\ln p+1$. The 2nd derivative is $1/p>0$ for all $p>0$.

Choose weights $w_i=1/N$. The average of the function minus the function of the average is $S_{\text{extrinsic}} \;=\; \overline{p\ln p} \;-\; \bar{p}\;\ln\bar{p} \quad\geq\quad 0.$

$$S_{\text{extrinsic}} \;=\; \sum w_i p_i \ln p_i \;-\; \left(\sum w_i p_i\right)\;\ln\sum w_i p_i \quad\geq\quad 0 \qquad (1)$$

$$S_{\text{intrinsic}} \;=\; \frac{\sum w_i p_i \ln p_i}{(\sum w_i p_i)} \;-\; \ln\sum w_i p_i \quad\geq\quad 0 \qquad (2)$$

If by monkeying with the parameters underlying p_i you can drive $S_{\text{intrinsic}}$ downwards, then you have driven the p_i towards one another.

We have the wrong sign convention. We have a choice of three: (1) introduce a minus sign, or (2) call S the *negentropy*, or (3) call S the Sparsity instead of Entropy.

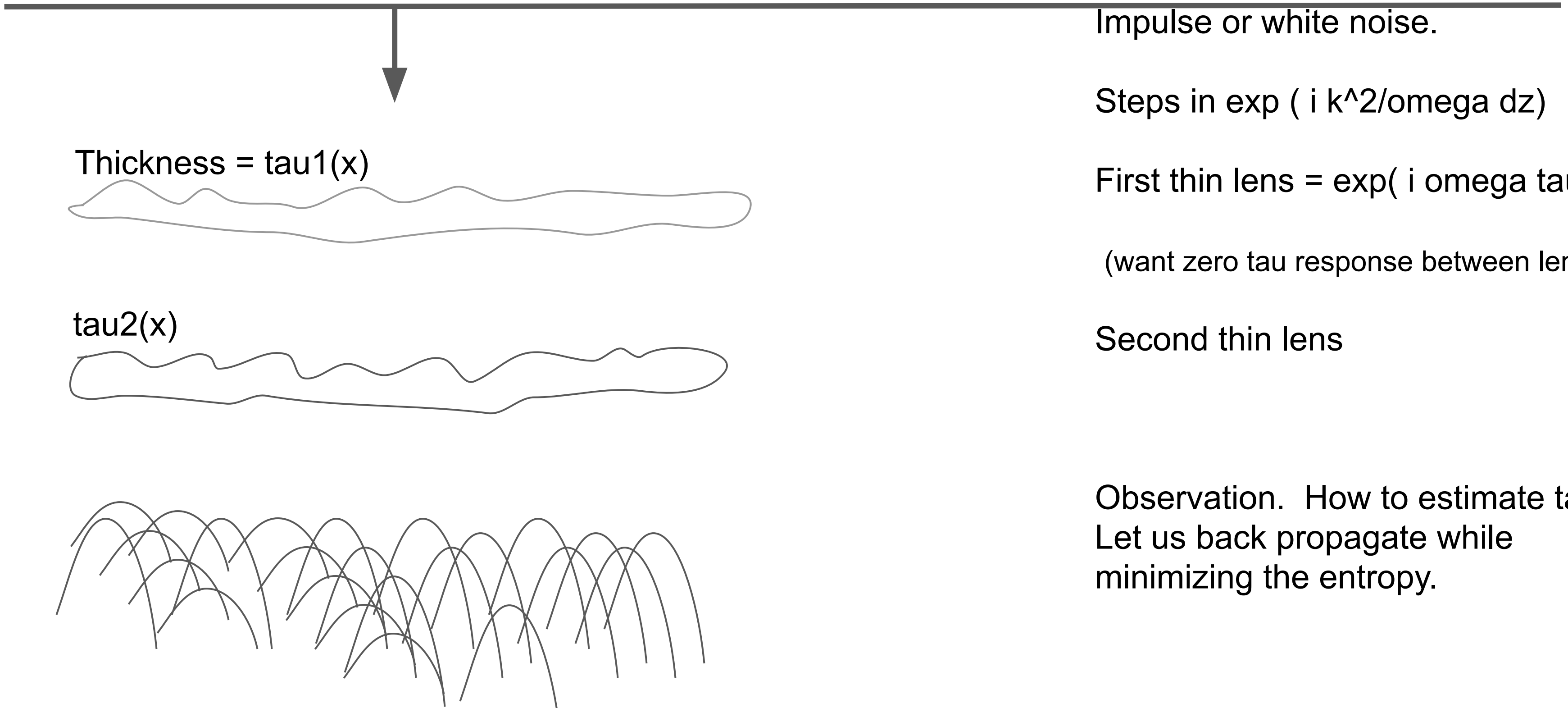
Once we have the gradient of S w.r.t lens thickness tau(x),
we can try various-sized jumps to reduce intrinsic entropy.

Random thoughts about p log(p)

- Physics books often take P as probability.
  - Why? Because they often assume Gaussian probability. Then the most probable is where the derivative vanishes. Most probable implies minimum variance = least squares.
- The Kulback-Leibler Divergence Sum P log(P/Q) found in t-SNE of machine learning is a minor restatement of Sum(w p log(p)) where Q is a prior density and weights w=Q.
- Outside the space of our study, we pad with |r(i)|=0.   Does this affect results?
- I never showed you **the gradient** of the intrinsic Jensen inequality for p log(p).   You can do it. Or ask Mathematica to help you. Haha.   dS/dg = dS/dp dp/dg

I want to play with a two-lens model.

Thin lens is a simple time shift.



Plane wave
Impulse or white noise.

Steps in exp (i k^2/omega dz)

Thickness = tau1(x)

First thin lens = exp(i omega tau dz)

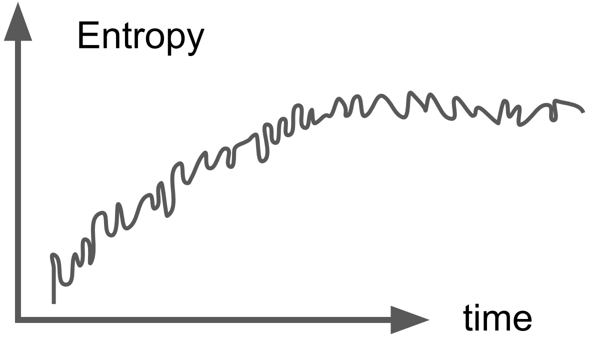
(want zero tau response between lenses)

tau2(x)

Second thin lens

Observation. How to estimate tau?
Let us back propagate while
minimizing the entropy.

In Physics, entropy is said to be a rough function
We want to minimize it?
We better smooth?



While finding the best g, I was maximizing. I never computed a gradient. I simply evaluated the entropy at significantly spaced intervals and chose the g of the largest entropy.

If our unknown is a vector or an image, the roughness might be a problem. Who knows? This is science!

Are there any easier applications to think about?

1. So far, what Jon did was easy: Jon found the L1/L2 transition g on the hyperbolic penalty. You might look for simple applications such as 1-D. What holds back mathematicians is that they seek problems with solutions. For the P log(P) method, all we need is the gradient. We can find step sizes by experimentation --- just the same way mathematicians tell us to find epsilon. (Haha)

2. Antoine and Jon worked out a **sparseness decon** and got very happy results on all 5 datasets tried. Could we repeat that performance with p log(p) binning --- a wholly different approach?

3. In the presence of multiple reflections, the Dix method for interval velocity suddenly becomes very complicated. This binning approach seems ideal for clustering. But, can we build a demonstration?

4. What is the relationship between **machine learning** and **p log(p)** ?

5. Have you any ideas?

Research is a strenuous and devoted attempt to force nature into the conceptual boxes supplied by professional education.

~Thomas Kuhn, *The Structure of Scientific Revolutions*

An environment for **cultivating creativity** emerges through

- Projects
- Peers
- Play
- Passion

--- Mitchell Resnick
Inventor of MIT Scratch

An environment for **cultivating creativity** emerges through

- Projects
- Peers
- Play
- Passion

--- Mitchell Resnick
Inventor of MIT Scratch

Want to explore some toys?
Go to **scratch.mit.edu**. There select ***explore***, then search for ***Machine Learning***.

Recall Kjartansson from IEI

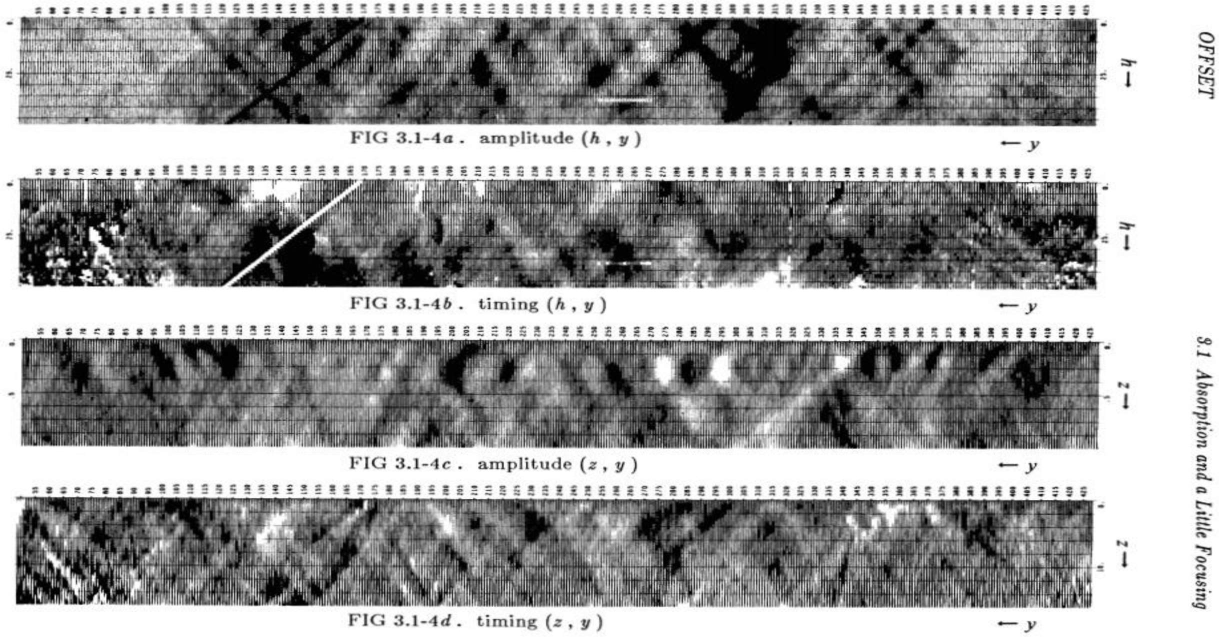


FIG 3.1-4*a*. amplitude (h, y)

FIG 3.1-4*b*. timing (h, y)

FIG 3.1-4*c*. amplitude (z, y)

FIG 3.1-4*d*. timing (z, y)

OFFSET

3.1 Absorption and a Little Focusing