Automatic default for hyperbolic softclip

Jon Claerbout

ABSTRACT

The hyperbolic penalty function leads us to gain residuals r, initially data d, by gain-parameter g in the softclip function $h'(d) = gd/\sqrt{1+g^2d^2}$ 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.

INTRODUCTION

We are looking at a non-linear scaling function for the display of seismic data d, and for scaling residuals r in preparation for model update. Experience has led us to the hyperbolic penalty function $h = \sum_i h_i$ where $h_i^2 = 1/g^2 + r_i^2$. The scalar parameter $1/g$ is the threshold between ℓ_2 and ℓ_1 behavior. Choosing a good value for g is the main topic here. We call the gradient of h the softclip $h'(d_i) = gd_i/\sqrt{1+g^2d_i^2}$. The softclip output is in the range $[-1, +1]$ for any gain value g. Having a bounded range like this is a convenient means to cope with the bounded brightness range available for paper printing and for screen display. Common practice in our lab for plotting is a hard limit at the device limitation called the clip. That parameter is the experimenters' choice. We default it to be the 99th percentile `pclip=99` of data values, but not rarely do we uncover a pitfall arising from failure to examine output of `pclip=100`.

THE TESTS

For a simple test case I selected the early half the time axis and a little over half the offset axis of an old Canadian shot gather, number 25 from the Yilmaz and Cumro shot gathers. I passed this data through the time honored t^2 gain function. After experimentation, I selected a numerical value for the softclip parameter g that led to the subjectively pleasing results you see in Figure 1. My choice of that value turned

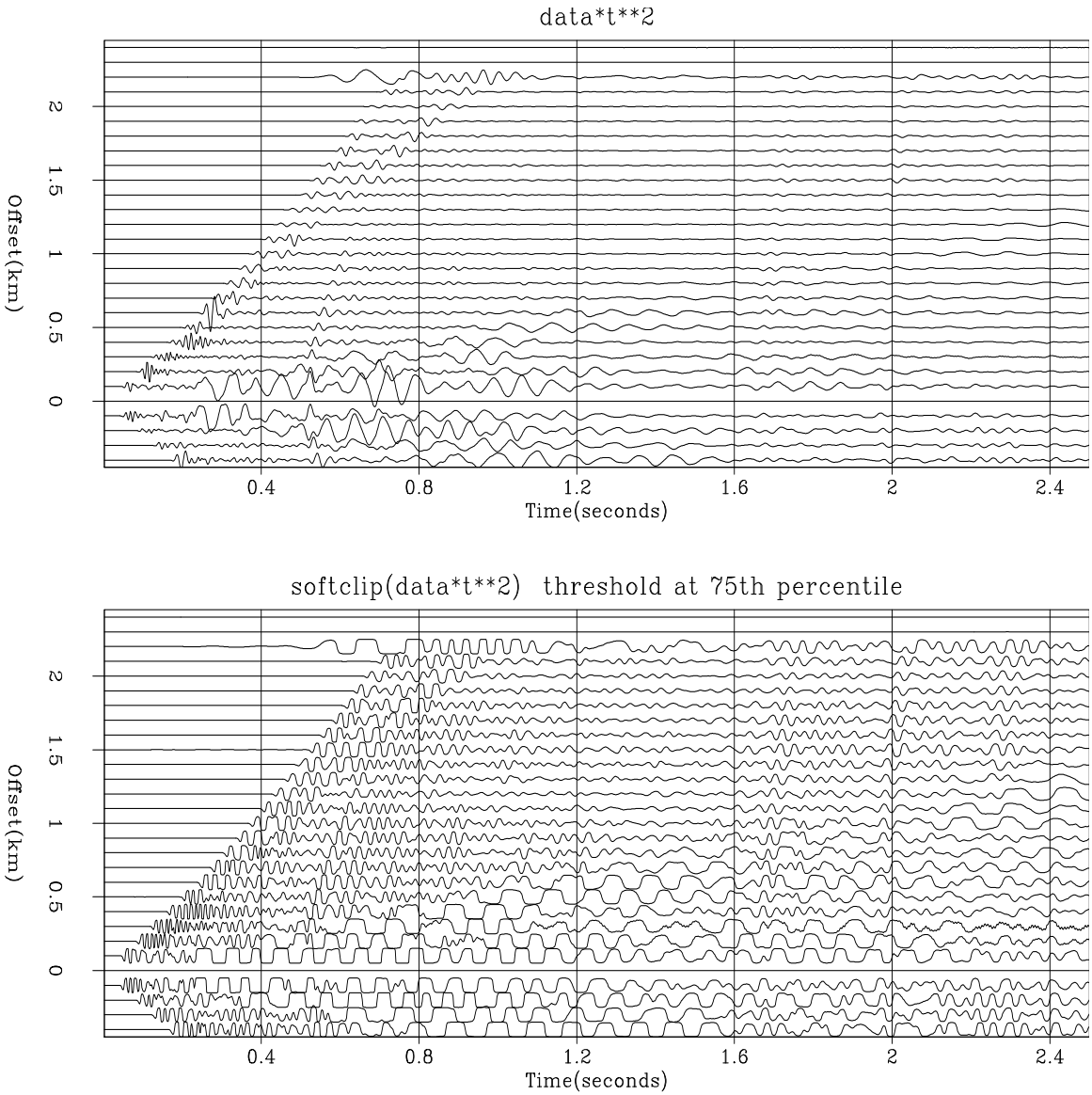

Figure 1: Top is a land shot gather d from Alberta. It has been gained with the traditional t^2 gain. Bottom gains additionally with the nonlinear softclip function $h'(d) = gd/\sqrt{1+g^2d^2}$ where the numerical value of g was chosen subjectively as the inverse of the 75th percentile of data values. Do you like it? [**ER**]

out to be the inverse of the 75th percentile of the data $|d|$. I set the goal of this work to find the g yielding the most pleasing result automatically rather than by experimentation. Others may later test to see if this value speeds convergence in model fitting, both in ℓ_2 and other non-linear (but convex) fitting.

The notion of democracy

In exploring data we have some clear goals and other goals not so clear. Informally we might regard the ground roll in Figure 1 as the dangerous but not necessarily bad oligarchy, while the dead traces and values before first breaks are the street bums, mostly undesirable, but not all bad either. Too many bums with opinions too self-serving, everything could fall apart. On the other hand, the oligarchy have a way of turning up everywhere. We want to listen to them but they have an unfortunate way of coming to dominate the show. We don't want a system allowing that. The ℓ_1 norm gives the bums a vote as big as everyone else, something of doubtful wisdom we'll see here. We start off experimenting with something like equality, but first we have to define it.

Our cultural background includes the old philosophy that residuals and preconditioned models should be IID, independent, identically destributed. Recall the gradient is nothing but the residual as seen in model space (the adjoint into the residual). In practice, the ID part of IID means the variables should have identical variance. The hypothesis injected here, called the "notion of democracy," is that the variables after softclip should be uniformly distributed in amplitude. We want to make equal use of every brightness level our display equipment has. We have predetermined ourselves to using the softclip gain and are seeking the g value that makes them uniformly distributed. It is my hypothesis that this choice of g will lead to subjectively attractive displays, and when applied to residuals, lead to more rapid iterative descents.

Scaling residuals (weighting) and models (preconditioning) is initially based on prior expectations. But experience shows residuals and models can be important where small, and unimportant where large. This suggests the principle of democracy, that to begin with, we'd like all residual values to have an equal chance at moving us towards the solution. That leads towards the notion of ℓ_1 norm with its signum function gradient. But such a population is far from Gaussian and far from uniform. It's time to look again at the experimental result, Figure 2.

It's also time to look at Figure 3 showing amplitude frequencies varying with g. Initially, I chose 256 bins, that being the number of brightness levels in our routine screen plotting. That showed no advantage over what you see here. It also showed a presentation disadvantage for us in that the bums became a narrower and higher peak making the altitude of the bulk population proportionately lower, hence less easily examined. I also tried more than 5 gain levels. Having video presentation facility with denser sampling of gain made picking one's favorite result a bit more fun, but trying to quantify " subjectivity pretty" more precisely pretty much limits the value

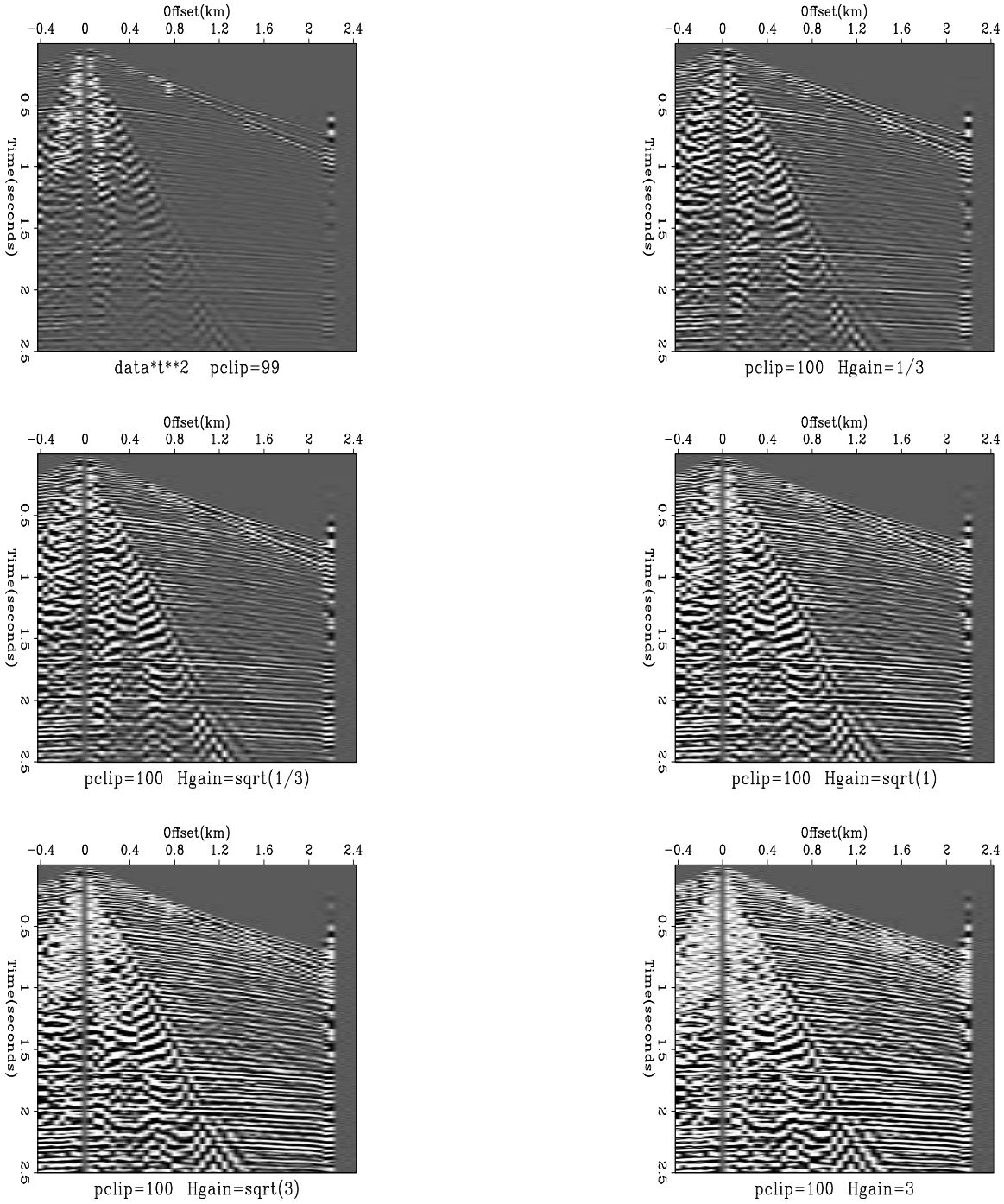

Figure 2: Top left is the default plot in our laboratory. You see the early part of a shot gather after scaling by t^2 and clipping at the 99th percentile. In other words, one percent of the data has gone off scale, so that has been plotted at full scale, the clip value. Since soft clipped data already lies in the range $[+1, -1]$ it is plotted at `pclip=100`. Subsequent plots try various gain constants g in the softclip $t'(d) = gd/\sqrt{1+g^2d^2}$. These are $1/3, \sqrt{1/3}, 1, \sqrt{3}, 3$ all divided by the 75th percentile. Choose your favorite! (Apologies for the paper image display. See the video.) [**ER**]

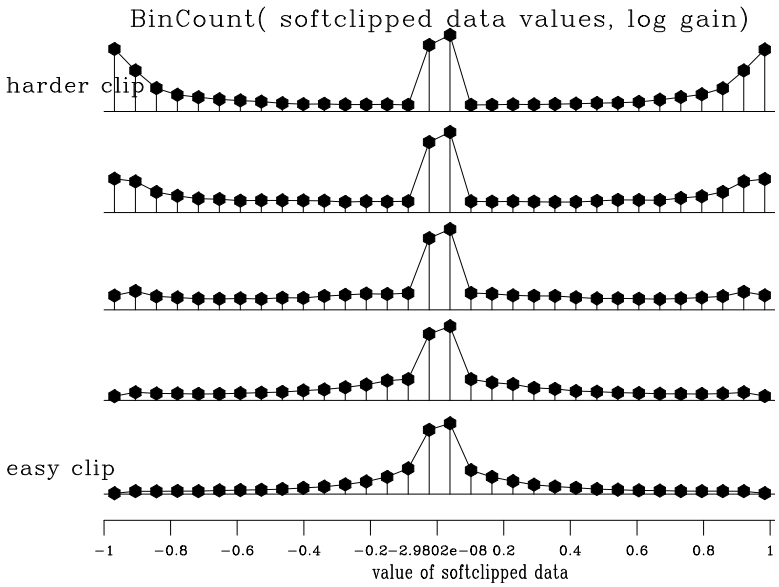

Figure 3: Data values in Figure 2 were sorted into 32 bins. The count in each bin is plotted ranging from the left bin near -1 to the right bin near $+1$. Results are shown for the five gain values displayed in Figure 2. The top plot shows the dominance of nearly clipped data, the ℓ_1 zone, not desirable. The bottom plot is closer to the ℓ_2 zone, not desirable either. The most uniform result, the middle one, is the one that had been chosen aesthetically. It's time to say, "Hooray for the uniform distribution!" [**ER**]

of that kind of fun, moreover Figure 2 was already too crowded already holding 6 plots.

Relation to optimization

Iterative descent has two parts, pointing a direction, and moving a distance. This paper proposes an answer to the first, the value of g for defining a direction $\Delta\mathbf{m} = \mathbf{F}^*\mathbf{h}'(g\mathbf{r})$. A deeper question, approached but not answered here arises next. The g for moving to minimize $h = \sum_i h_i$ where $h_i^2 = 1/g^2 + r_i^2$, what should be the value of this g? Should the moving g have the same numerical value as the pointing g? At the present time the moving g is a numerical parameter the analyst embeds in the objective function. At the outset, it's not easy to choose this g. In practice some idea of the threshold might be suggested, but that would be based on population percentiles. What we will find here is a good g for direction choosing. Perhaps the same g will be suitable for the distance moving? If so, the residual itself would be telling us what to optimize! Is there circular reasoning here? or is this a great discovery? Looks dangerous. Need help. That means examples.

Caution tells us not to change the definition of the problem while we are solving the problem. That means keeping fixed the moving g. On a more familiar topic, focusing our attention on the direction choice only, preconditioning changes the gradient in a predetermined way, scaling it by a positive definite matrix, namely $\Delta\mathbf{m} = \mathbf{F}^*\mathbf{r}$ becomes $\Delta\mathbf{m} = \mathbf{S}\mathbf{S}^*\mathbf{F}^*\mathbf{r}$. Although $\mathbf{S}\mathbf{S}^*$ is commonly fixed at the outset, I believe it need not be. Alternately, the hyperbolic penalty function changes the gradient in a different way, namely $\Delta\mathbf{m} = \mathbf{F}^*g\mathbf{r}$ becomes $\Delta\mathbf{m} = \mathbf{F}^*\mathbf{h}'(g\mathbf{r})$. Should the g of the direction choice with $\mathbf{h}'$ be the same as the g in the moving choice with $h = \sum h_i$. Have we uncovered a "black swan" detector? Experience will tell.

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

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

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

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

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

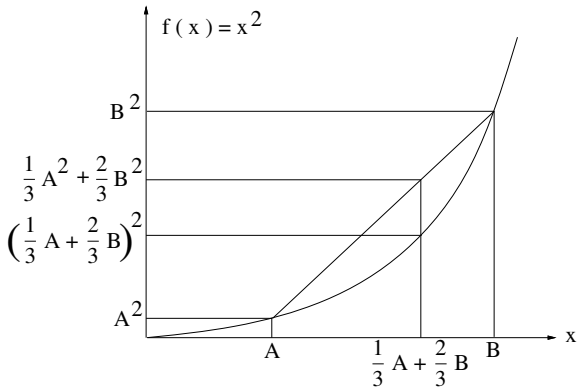

Figure 4: Sketch of $y = x^2$ for interpreting equation (2). [**NR**]

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

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

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

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 $\overline{f(p)} - f(\overline{p}) \geq 0$ gives the familiar result that the mean of the squares exceeds the square of the mean:

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

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:

$$H \;=\; \sum \frac{w_i}{p_i} \;-\; \frac{1}{\sum w_i p_i} \quad\geq\quad 0 \tag{6}$$

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

$$G \;=\; -\sum w_i \ln p_i \;+\; \ln \sum w_i p_i \quad \geq \quad 0 \tag{7}$$

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

$$\frac{1}{N}\sum_{i=1}^{N} p_i \;-\; \prod_{i=1}^{N} p_i^{1/N} \quad \geq \quad 0 \tag{8}$$

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 $\overline{f(p)} - f(\overline{p}) = E(f) - f(E) \geq 0$ is:

$$S_{\text{extrinsic}} \;=\; \sum w_i p_i \ln p_i \;-\; \Big(\sum w_i p_i\Big)\; \ln \sum w_i p_i \quad \geq \quad 0 \tag{9}$$

$$S_{\text{intrinsic}} \;=\; \frac{\sum w_i p_i \ln p_i}{\sum w_i p_i} \;-\; \ln \sum w_i p_i \quad \geq \quad 0 \tag{10}$$

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.

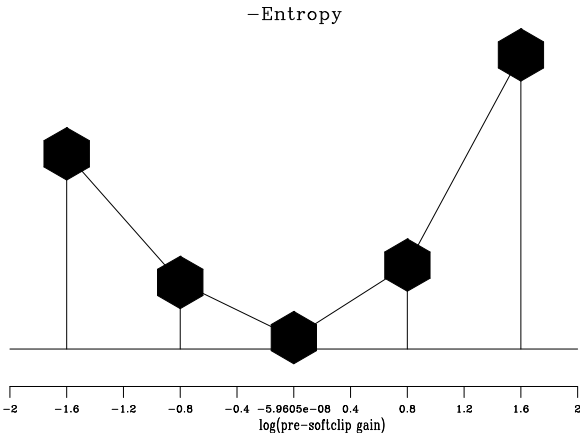

Figure 5: Evaluating the Jensen inequality S of the bins in Figure 3 shows our favorite value of g has the least S. Hooray! Hooray! Hooray! But there was a small problem. Before I used weights to punch out the bums (dead traces, etc), the minimum tilted a bit towards a little harder clip. [**ER**]

Figure 5 now provides me a promising data result from my long-ago theoretical efforts. Hooray again!

CONCLUSIONS

In practice we may wonder which Jensen inequality to use. Seismograms often contain zeros. Notice that a single zero p_i can upset the harmonic H or geometric G inequality,

but a single zero has no horrible effect on S or Q. But Q can be taken over by the oligarchy, so I am happy to conclude with S. Hooray for democracy! Antoine Guitton makes the valuable suggestion that the inverse square-root is another convex function, one dominated by neither large or nor small values.

I hypothesize softclipping the gradient is a general procedure to speed convergence. I also hypothesize the automatically picked g is a good one for the penalty function itself.

Prior and posterior distributions

Random variables have a prior distribution and a posterior distribution. Denote the prior by b_i (for "before") and posterior by a_i (for "after"). Define $p_i = a_i/b_i$, and insert p_i in any of the inequalities above. Now suppose we have an adjustable model parameter like g upon which the a_i all depend. Suppose we adjust that model parameter to try to make some Jensen inequality into an equality. Thus we will be adjusting it to get all the p_i equal to each other, that is, to make all the posteriors equal to their priors. Jensen inequality provides many ways to do this. The one I named entropy should be close to the Shannon or Kullback ideas. Perhaps it is their use of continuous variables instead of sampled ones that make their ideas opaque.

A Gaussian distribution of any variance would soon give us its optimum g and we would observe its final best distribution (like figure 3). That would define the b_i for us to do another iteration.

FUTURE WORK

The idea is well demonstrated for plotting. Whether and what kind of data fitting will be accelerated or improved is speculative. Now we should see if it can take root, to see if some important aspect has been overlooked, or to see what, if any, minor adjustments should be made. The minimization method, examining a handful of gains in the neighborhood of the 75th percentile, is crude, but it's far from certain that a more precise minimum would be worth the extra work.

1. Code like that here should be split out into a SEPLIB utility to receive more use and comment, mostly for plot scaling.
2. If it proves popular, a library subroutine interface should be provided to give access to data fitting projects. Accelerated convergence would be a big plus.
3. We should test Antoine's suggestion basing a Jensen inequality on the negative square root function.
4. We should test whether softclipping the gradient is a general procedure to speed convergence.

5. This idea should play a role in data fitting, but how does it fit with model styling?

6. Maximizing model entropy would drive down model fluctuations without making the model space smoother. That's interesting! Do we have a use for it?

7. Starting with Gaussian random variables of unit variance, what value of g would best flatten it? And how flat would that be? (This might be a clue to a "best" convex penalty function.)

8. Ideas from the community should be explored.

DISCUSSION

I first derived the entropy expression in SEP 37 and also published it in PVI. But I derived it the hard way starting from the convex function $f = |r|^{1+\epsilon}$ as $\epsilon \to 0$ while the method I use above was suggested to me by a book "Inequalities," by Hardy, Littlewood, and Polya, Cambridge University Press, 1934 who also have the result in the "programmer ready form" like mine.

Most likely this result is also implicit in the work of Jensen, of Gibbs, or of Shannon, but I haven't seen it there or elsewhere in the self contained "programmer ready form." It belongs in Wikipedia.

Stew Levin agreed to check my algebra, but better than that, he dug up the Hardy book which provides us the much simpler derivation above. Actually, the Hardy gang expressed it for weights that are not necessarily normalized, $\sum_i w_i \neq 1$, which offers a slight programming advantage. For example, initially I omitted weights choosing them identical and inverse to their number. But when time came to knock out the bums, I had to renormalize. Here is the Hardy result in my notation:

$$S \quad = \quad \frac{\sum w_i p_i \ln p_i}{\sum w_i p_i} \; - \; \ln \frac{\sum w_i p_i}{\sum w_i} \quad \geq \quad 0 \tag{11}$$