12-1-2010

Why Curvature in L-Curve: Combining Soft Constraints

Uram Anibal Sosa Aguirre
University of Texas at El Paso, usosaaguirre@miners.utep.edu

Martine Ceberio
University of Texas at El Paso, mceberio@utep.edu

Vladik Kreinovich
University of Texas at El Paso, vladik@utep.edu

Comments:
Technical Report: UTEP-CS-10-58
Published in *Proceedings of the Fourth International Workshop on Constraint Programming and Decision Making CoProD'11*, El Paso, Texas, March 17, 2011.

Why Curvature in L-Curve: Combining Soft Constraints

Uram Anibal Sosa Aguirre, Martine Ceberio, and Vladik Kreinovich

Computational Sciences Program and Department of Computer Science
University of Texas, El Paso, TX 79968, USA
usosaaguirre@miners.utep.edu, {mceberio,vladik}@utep.edu

Abstract. In solving inverse problems, one of the successful methods of determining the appropriate value of the regularization parameter is the *L-curve method* of combining the corresponding soft constraints, when we plot the curve describing the dependence of the logarithm x of the mean square difference on the logarithm y of the mean square non-smoothness, and select a point on this curve at which the curvature is the largest. This method is empirically successful, but from the theoretical viewpoint, it is not clear why we should use curvature and not some other criterion. In this paper, we show that reasonable scale-invariance requirements lead to curvature and its generalizations.

Keywords: soft constraints, inverse problems, regularization, L-curve, curvature

1 Formulation of the Problem

Inverse problem: a brief reminder. In science and engineering, we are interested in the state of the world, i.e., in the values of different physical quantities that characterize this state. Some of these quantities we can directly measure, but many quantities are difficult or even impossible to measure directly.

For example, in geophysics, we are interested in the density and other properties of the material at different depths and different locations. In principle, it is possible to drill a borehole and directly measure these properties, but this is a very expensive procedure, and for larger depths, the drilling is not possible at all. To find the values of such difficult-to-measure quantities $q = (q_1, \ldots, q_n)$, we measure the values of the auxiliary quantities $a = (a_1, \ldots, a_m)$ that are related to q_i by a known dependence $a_i = f_i(q_1, \ldots, q_n)$, and then reconstruct the values q_j from these measurement results.

In the idealized situation when measurements are absolutely accurate, we can then reconstruct the desired values q_j from the system of m equations $a_1 = f_1(q_1, \ldots, q_n)$, $\ldots$, $a_m = f_m(q_1, \ldots, q_n)$. In real life, measurements are never 100% accurate, so the measured values a_i are only approximately equal to $f_i(q_1, \ldots, q_n)$. Usually, it is assumed that the measurement errors $a_i - f_i(q_1, \ldots, q_n)$ are independent normally distributed random variables with 0 means and the same variance; see, e.g. [3]. In this case, the constraint that the

values q_j are consistent with the observations a_i can be described as a constraint $s \le s_0$ on the sum $s \stackrel{\text{def}}{=} \sum_{i=1}^{m} (a_i - f_i(q_1, \ldots, q_n))^2$. The value s_0 depends on the confidence level: the larger s_0, the more confident we are that this constraint will be satisfied. For each value x_0, the constraint $x \le x_0$ is a *soft constraint*: there is a certain probability that this constraint will be violated.

Often, this constraint is described in a logarithmic scale, as $x \le x_0$, where $x \stackrel{\text{def}}{=} \ln(s)$.

Regularization: how to take into account additional constraints. Often, there are additional constraints on q_j. Usually, the values q_j are more regular than randomly selected values. Methods for taking these additional regularity constraints into account are known as *regularization methods*; see, e.g., [4].

For example, in geophysics, the density values at nearby locations are usually close to each other. In other words, the differences $q_j - q_{j'}$ corresponding to nearby locations should be small.

This constraint can also be described in statistical terms: that there is a prior distribution on the set of all the tuples, in which all the differences $q_j - q_{j'}$ are independent and normally distributed with 0 mean and the same variance. In this case, the constraint that the values q_j are consistent with this prior distribution can be also described as a constraint $t \le t_0$ on the sum $t \stackrel{\text{def}}{=} \sum_{(j,j')} (q_j - q_{j'})^2$. This constraint is also often described in a logarithmic space, as $y \le y_0$, where $y \stackrel{\text{def}}{=} \ln(t)$.

We can combine the two constraints, e.g., by using the Bayesian statistics to combine the prior distribution (describing the regularity of the actual values) and the distribution corresponding to measurement uncertainty. For the resulting posterior distribution, the Maximum Likelihood method of determining the optimal values of the quantities q_j is then equivalent to minimizing the sum $s + \lambda \cdot t$, for some coefficient λ depending on the variance of the prior distribution.

There are also other more complex regularization techniques; see [4].

How to determine a regularization parameter. As we have mentioned, the actual value of the regularization parameter depends on the prior distribution and is, therefore, reasonably subjective. It is therefore desirable to find the value of this parameter based on the data.

For each value of the parameter λ, we can find the corresponding solution $q_j(\lambda)$, and, based on this solution, compute the values $x(\lambda)$ and $y(\lambda)$ of the quantities x and y. These two values represent a point on a plane. Points corresponding to different values λ form a curve. In these terms, the question of which value λ to choose can be reformulated as which point on the curve should we choose?

In practice, often, this curve has a clear turning point, a point that is distinct from others – as a point at which the curve "curves" the most. In such cases,

when we have an L-shaped curve, it is reasonable to select the turning point as the point corresponding to the solution. This idea often leads to a good solution; see, e.g., [1, 2].

In line with the above description, the desired point is selected as a point at which the absolute value $|C|$ of the curvature $C = \dfrac{x'' \cdot y' - y'' \cdot x'}{((x')^2 + (y')^2)^{3/2}}$ takes the largest possible value; here, as usual, x' denotes the derivative $\dfrac{dx}{d\lambda}$, and x'' denotes the second derivative of x with respect to the parameter λ.

Remaining open problem. Empirically, the method of selecting a point with the largest curvature works well. It is therefore desirable to come up with a theoretical justification for the use of curvature function – or at least for a class containing the curvature function.

What we do in this paper. We provide such a justification: specifically, we show that reasonable properties select a class of functions that include curvature.

2 Analysis of the Problem

Let us first analyze the invariance properties of curvature.

Scale-invariance. The numerical values of each quantity depend on the selection of a measuring unit. For example, if instead of meters, we use centimeters, then all numerical values get multiplied by 100. In general, if we select a new measuring unit which is c times smaller than the previous one, then all numerical values get multiplied by c.

If we change a measuring unit for a to a new one which is c_a time smaller, then the numerical values of a_i and $a_i - f_i(q_1, \ldots, q_n)$ get multiplied by c_a. As a result, the sum $s = \sum\limits_{i=1}^{n} (a_i - f_i(q_1, \ldots, q_n))^2$ gets multiplied by c_a^2, and the original value $x = \ln(s)$ changes to $x + \Delta_x$, where we denoted $\Delta_x \stackrel{\text{def}}{=} \ln(c_a^2)$.

Similarly, if we change a measuring unit for q to a new one which is c_q time smaller, then the numerical values of q_j and $q_j - q_{j'}$ get multiplied by c_q. As a result, the sum $t = \sum (q_j - q_{j'})^2$ gets multiplied by c_q^2, and the original value $y = \ln(t)$ changes to $y + \Delta_y$, where we denoted $\Delta_y \stackrel{\text{def}}{=} \ln(c_q^2)$.

Under these changes $x(\lambda) \to x(\lambda) + \Delta_x$ and $y(\lambda) \to y(\lambda) + \Delta_y$, the derivatives do not change – since Δ_x and Δ_y are constants – and thus, the curvature does not change. Thus, the curvature is invariant under these scale transformations.

Invariance under re-scaling of parameters. Instead of the original parameter λ, we can use a new parameter μ for which $\lambda = g(\mu)$. This re-scaling of a parameter does not change the curve itself and thus, does not change its curvature. So, the curvature is invariant under these scale transformations.

Our idea. Our main idea is to describe all the functions which are invariant with respect to both types of re-scalings.

3 Main Result

Definition. *By a* parameter selection criterion *(or simply* criterion*, for short), we mean a function $F(x, y, x', y', x'', y'')$ of six variables. We say that the parameter selection criterion $F(x, y, x', y', x'', y'')$ is:*

- scale-invariant *if for all possible values Δ_x and Δ_y, we have*

$$F(x + \Delta_x, y + \Delta_y, x', y', x'', y'') = F(x, y, x', y', x'', y'');$$

- invariant w.r.t. parameter re-scaling *if for every function $g(z)$ and for the functions $\widetilde{x}(\mu) = x(g(\mu))$ and $\widetilde{y}(\mu) = y(g(\mu))$, we have*

$$F(\widetilde{x}, \widetilde{y}, \widetilde{x}', \widetilde{y}', \widetilde{x}'', \widetilde{y}'') = F(x, y, x', y', x'', y'').$$

Notation. By $C(x, y, x', y', x'', y'')$, we denote the parameter selection criterion corresponding to curvature.

Comment. Once a criterion is selected, for each problem, we use the value λ for which the value $F(x(\lambda), y(\lambda), x'(\lambda), y'(\lambda), x''(\lambda), y''(\lambda))$ is the largest.

Main result. *A parameter selection criterion which is scale-invariant and invariant w.r.t. parameter re-scaling if and only if it has the form*

$$F(x, y, x', y', x'', y'') = f\left(C(x, y, x', y', x'', y''), \frac{x'}{y'}\right)$$

for some function $f(C, z)$.

Proof.

1°. For each tuple (x, y, x', y', x'', y''), by taking $\Delta_x = -x$ and $\Delta_y = -y$, we conclude that $F(x, y, x', y', x'', y'') = F(0, 0, x', y', x'', y'')$. Thus, we conclude that $F(x, y, x', y', x'', y'') = F_0(x', y', x'', y'')$, where we denoted $F_0(x', y', x'', y'') \stackrel{\text{def}}{=} F(0, 0, x', y', x'', y'')$, i.e., we conclude that the value of the parameter selection criterion does not depend on x and y at all.

In terms of the function F_0, invariance w.r.t. parameter re-scaling means that $F_0(\widetilde{x}', \widetilde{y}', \widetilde{x}'', \widetilde{y}'') = F_0(x', y', x'', y'')$.

2°. When we go from the original function $x(\lambda)$ to the new function $\widetilde{x}(\mu) = x(g(\mu))$, the chain rule for differentiation leads to $\widetilde{x}' = x' \cdot g'$ and thus, $\widetilde{x}'' = x'' \cdot (g')^2 + x' \cdot g''$. Similarly, $\widetilde{y}' = y' \cdot g'$ and $\widetilde{y}'' = y'' \cdot (g')^2 + y' \cdot g''$.

In particular, at the point where $g' = 1$, we have $\widetilde{x}' = x$, $\widetilde{x}'' = x'' + x' \cdot g''$, $\widetilde{y}' = y'$, and $\widetilde{y}'' = y'' + y' \cdot g''$, and thus, invariance w.r.t. parameter re-scaling means that $F_0(x', y', x'' + x' \cdot g'', y'' + y' \cdot g'') = F_0(x', y', x'', y'')$. This is true for

every possible values of g''. In particular, for $g'' = -\dfrac{y''}{y'}$, we have $y'' + y' \cdot g'' = 0$ and thus,

$$F_0(x', y', x'', y'') = F_0\left(x', y', x'' - x' \cdot \frac{y''}{y'}, 0\right).$$

Since

$$x'' - x' \cdot \frac{y''}{y'} = C \cdot \frac{((x')^2 + (y')^2)^{3/2}}{y'},$$

we thus conclude that

$$F_0(x', y', x'', y'') = h(C, x', y'),$$

where

$$h(C, x', y') \stackrel{\text{def}}{=} F_0\left(x', y', C \cdot \frac{((x')^2 + (y')^2)^{3/2}}{y'}, 0\right).$$

For the new function $h(C, x', y')$, since the curvature is invariant w.r.t. parameter re-scaling, invariance means that $h(C, \widetilde{x}', \widetilde{y}') = h(C, x', y')$. This means that

$$h(C, x', y') = h(C, x' \cdot g', y' \cdot g').$$

This is true for every possible values of g'. In particular, for $g' = \dfrac{1}{x'}$, we have $x' \cdot g' = 1$ and thus,

$$F(x, y, x', y', x'', y'') = F_0(x', y', x'', y'') = h(C, x', y') = h\left(C, 1, \frac{y'}{x'}\right),$$

i.e., $F(x, y, x', y', x'', y'') = f\left(C, \dfrac{y'}{x'}\right)$ for $f(C, z) \stackrel{\text{def}}{=} h(C, 1, z)$.

The statement is proven.

Acknowledgments. This work was supported in part by the National Science Foundation grants HRD-0734825 and DUE-0926721 and by Grant 1 T36 GM078000-01 from the National Institutes of Health.

References

1. Hansen, P.C. Analysis of discrete ill-posed problems by means of the L-curve, SIAM Review 34(4), 561–580 (1992)
2. Moorkamp, M., Jones, A.G., Fishwick, S.: Joint inversion of receiver functions, surface wave dispersion, and magnetotelluric data. Journal of Geophysical Research 115, B04318 (2010)
3. Rabinovich, S.: Measurement Errors and Uncertainties: Theory and Practice. American Institute of Physics, New York (2005)
4. Tikhonov, A.N., Arsenin, V.Y.: Solutions of Ill-Posed Problems, W.H. Whinston & Sons, Washington, D.C. (1977)