A Bayesian Framework for Regularization

Daniel Keren
Department of Mathematics
and Computer Science
The University of Haifa
Haifa 31905, Israel
dkeren@mathcs1.haifa.ac.il

Michael Werman
Institute of Computer Science
The Hebrew University
Jerusalem 91904, Israel
werman@cs.huji.ac.il

This research was sponsored by ARPA through the U.S. Office of Naval Research under grant N00014-93-1-1202, R&T Project Code 4424341—01.

Abstract

Regularization looks for an interpolating function which is close to the data and also "smooth". This function is obtained by minimizing an error functional which is the weighted sum of a "fidelity term" and a "smoothness term". However, using only one set of weights does not guarantee that this function will be the MAP estimate. One has to consider *all* possible weights in order to find the MAP function. Also, using only one combination of weights makes the algorithm very sensitive to the data.

The solution suggested here is through the Bayesian approach: A probability distribution over all weights is constructed and all weights are considered when reconstructing the function or computing the expectation of a linear functional on the function space.

1 Introduction and Previous Work

In computer vision, regularization [19] is used to reconstruct objects from partial data [17, 18, 5, 1]. The data can be sparse — e.g. the height of a small number of points on a surface, or dense but incomplete — e.g. the case of optical flow and shape from shading [4] where data is available at many points but consists of the function's or its derivative's value in a certain direction only. The first difficulty in solving this problem stems from the multitude of possible solutions, each satisfying the partial data; which one should be chosen? Also, data instances which are not compatible with others can cause singularities in the solution.

The regularization approach overcomes these difficulties by choosing among the possible objects one which approximates the given data and is also "smooth". This embodies an important assumption — that the "smoother" the object, the more probable it is. Formally, a *cost functional* $M(f)$ is defined for every object f by $M(f) = D(f) + \lambda S(f)$, where $D(f)$ measures the distance of f from the given data, $S(f)$ measures the smoothness of f, and $\lambda > 0$ is a parameter. The f chosen is the one minimizing $M()$.

In the one-dimensional case, one minimizes $M(f) = \sum_{i=1}^{n} \frac{[f(x_i) - y_i]^2}{\sigma^2} + \lambda \int_0^1 f_{uu}^2 du$. In the two-dimensional case, the functional to minimize is $M(f) = \sum_{i=1}^{n} \frac{[f(x_i, y_i) - z_i]^2}{\sigma^2} + \lambda \int_0^1 \int_0^1 (f_{uu}^2 + 2f_{uv}^2 + f_{vv}^2) dudv$.

The question is, how does one choose λ and σ? There are various methods for doing this. However, all regularization schemes we are familiar with choose one combination of weights and use them alone to interpolate the function.

But there's a serious problem with this approach: It fails to find the MAP estimate for the interpolant f, as it uses only one set of weights λ and σ to construct f. But, what happens if the chosen function has a relatively small probability for a wide range of other weights? The MAP estimate should maximize the following:

$$\int_w Pr(f/w) Pr(w) dw$$

where w varies over the set of all possible weights. The main contribution of this work is the computation of $Pr(w)$.

The most popular method for determining the smoothing parameter λ is that of the Generalized

1051-4651/94 $04.00 © 1994 IEEE

Cross Validation (GCV) [2]. The problem with GCV is that the choice of the value of λ is sometimes very sensitive to the data. Since this value is crucial to the shape of the fitted curve or surface, it turns out that sometimes a small change in the data drastically changes the shape of the fitted function. Another problem is that although it can be proved that GCV has some nice asymptotic properties, the choice of the "optimal" values of λ and σ is heuristic in nature. One goal of this work is to suggest an improvement for the GCV algorithm. Bayesian approaches for choosing the weights are suggested in the pioneering work of Szeliski [15] and more recently in [10]. Here we suggest a different approach, namely, computing the probability distribution by directly integrating over the (infinite-dimensional) space of all possible interpolants. Another novelty is in using all possible weights, not only the "optimal" ones.

2 Computing the Joint Probability for λ, σ

Suppose we have a data set D and we want to describe or fit it with a member of some model M. The Bayes solution is to find f which satisfies

$$\max_{f \in M} Pr(f/D) = \max_{f \in M} \frac{Pr(D/f)Pr(f/M)}{Pr(D)}$$

$$\propto \max_{f \in M} Pr(D/f)Pr(f/M)$$

Regularization, for instance, can be formalized in this way because

$$Pr(D/f) = \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n} \exp(-\sum_{i=1}^{n} \frac{[f(x_i) - y_i]^2}{2\sigma^2})$$

(assuming uncorrelated Gaussian noise of constant variance) and the prior distribution is

$$Pr(f) \propto \exp(-\lambda \int f_{uu}^2 du)$$

which resembles the Boltzmann distribution [15, 3, 6, 8, 14, 13, 11, 9, 16, 12]. Multiplying, we get that the f chosen from M should maximize $\exp(-M(f))$, or minimize $M(f)$. This simple analysis shows how regularization is consistent with Bayes rule for choosing the MAP estimate, given λ.

Now, what if a few models are possible? As explained before, the first step is to compute the probability of each model. In our case, the models are indexed by two continuous parameters, λ and σ. Thus, a specific choice of the two parameters is equivalent to the claim "this function was sampled from a function space with a specific prior, and the measurement was corrupted by a specific noise".

Call the model that assumes λ as a smoothing parameter and σ as the measurement noise $M_{\lambda,\sigma}$. In this model, $Pr(f) \propto \exp(-\lambda \int f_{uu}^2 du)$. Given a data set D, we compute $Pr(M_{\lambda,\sigma}/D)$. Using Bayes rule:

$$Pr(M_{\lambda,\sigma}/D) = \frac{Pr(D/M_{\lambda,\sigma})Pr(M_{\lambda,\sigma})}{Pr(D)} \quad (1)$$

$$\propto Pr(D/M_{\lambda,\sigma}) = \frac{\int_{M_{\lambda,\sigma}} Pr(D/f)Pr(f/M_{\lambda,\sigma})\mathcal{D}f}{\int_{M_{\lambda,\sigma}} Pr(f/M_{\lambda,\sigma})\mathcal{D}f}$$

where the denominator is introduced to turn the distribution on the functions f into a probability, by normalizing its integral on the whole space to 1.

Since the data is given, it is the same for all models and can be eliminated from consideration.

It turns out that although the space $M_{\lambda,\sigma}$ is infinite dimensional, it is possible to reduce the integral to a quotient of two integrals defined on a finite dimensional space. There is not enough space here to present the computation (see [7]); however, the integral turns out to be equal to

$$\frac{\lambda^{\frac{n}{2}}}{\pi^{\frac{n}{2}}}|A + 2\lambda\sigma^2 I|^{-\frac{1}{2}} \exp(-\lambda Y(A + 2\lambda\sigma^2 I)^{-1}Y^t) \quad (2)$$

where Y is the data vector and A is a matrix which can be easily computed from the sample points (it does not depend on the data).

3 Computing the Expectation of the Value at x

If L is a functional on f, its expectation given D is

$$E[L(f)/D] = \int L(f)Pr(f/D)\mathcal{D}f$$

if we want to compute the value of a function at a point x, then $L(f)$ is simply the evaluation at x, and the expectation can be computed, according to Fubini's theorem, by first evaluating it for each $\{\lambda, \sigma\}$ pair, and then integrating over all such pairs, weighing each one by its probability conditioned by the data D:

$$E[f(x)/D] = \int\int E[f(x)/D, M_{\lambda,\sigma}]Pr(M_{\lambda,\sigma}/D)d\lambda d\sigma$$

which can be shown to equal [7]

$$\frac{\int \frac{1}{\sqrt{v}}|A+vI|^{-\frac{1}{2}}(H_{x_1}(x)..H_{x_n}(x))(A+vI)^{-1}Y^t[Y(A+vI)^{-1}Y^t]^{\frac{t-n}{2}}dv}{\int \frac{1}{\sqrt{v}}|A+vI|^{-\frac{1}{2}}[Y(A+vI)^{-1}Y^t]^{\frac{t-n}{2}}dv} \tag{3}$$

where H_{x_i} are simple polynomial functions in x, $i = 1..n$.

We have therefore presented a closed-form expression for evaluating the expectation of the value of f at a point, given partial data about f. No optimization and no selection of an "optimal" parameter are necessary. One has to note, though, that the function defined pointwise by this expression is *not* necessarily the function which maximizes the probability $Pr(f/D)$ – this is true only in the simple model which uses only one value of the λ and σ parameters. The problem of finding the most probable f will be addressed in a subsequent paper. Nontheless, the function defined by the pointwise expectations can serve as an interpolant; however, as is obvious from the examples, this interpolant is not very smooth. This is because it is computed *pointwise*, without attempting to force global smoothness. It is a question of what one desires; if one wants to give a good estimate for the value of the function at a certain point, Equation 3 is the expression to use. If one wants to estimate the global function, that is a different matter. Oddly enough, the value of the most probable function at the point x is *not* the "best" estimate for the value at x (that is, the expectation).

4 Experimental Results

Next, we demonstrate how the proposed approach results in stable fits. The algorithm was extended to 2D data [7]. In Figure 1, we show the GCV reconstruction for data which consist of a sinusoidal pattern contaminated by Gaussian noise. The data in Figure 2 differs from that of Figure 1 at only one point. The GCV returns radically different results for the two data sets, because this slight change caused GCV to choose a very different value of λ. Figure 3 shows the reconstruction for the data of Figures 1 and 2 using Equation 3. The fits are nearly similar.

Figure 4 shows the GCV reconstruction to data created by adding Gaussian noise to the function $x(1-x)y(1-y)$. In Figure 5, the GCV fit to a data set differing from that of Figure 4 by one point is given. Again, GCV chooses a very different value of λ and returns a very different fit. Applying Equation 3 to the data of Figure 4 results in the surface presented in Figure 6. Applying Equation 3 to the data of Figure 5 results in a fit which is almost similar.

5 Conclusion and Future Work

A novel approach for looking at regularization is suggested, which considers all possible values for the noise and smoothing parameter, by assigning a probability to each combination of these two parameters. This probability is used to compute the expectation of a linear functional over the function space. The main advantages of this interpolation paradigm are that it is free of instability problems which haunt interpolation schemes that choose a single value for these parameters; also, the computation of the parameter's probabilities is performed in a mathematically rigorous manner. We have also demonstrated that the optimal parameters can be found by solving a simple one-dimensional minimization problem [7].

In the future, we plan to demonstrate how the most probable function is computed, and compare it to the pointwise interpolation presented in this work. We also hope to study other priors/models, notably those that belong to the realm of "robust statistics". These pose a fascinating mathematical challenge because, usually, the probability they assign to models can not be expressed in terms of inner products on the function spaces involved. Finally, an attempt is being made to tie this research to the fascinating field known as "information-based complexity", or "continuous complexity".

6 Acknowledgment

We wish to express our sincere gratitude to Alan Williams, who helped us to implement the 2D Generalized Cross Validation package. We also thank Prof. Israel Elitzur for pointing out the connection between probability distributions on function spaces and Quantum theory.

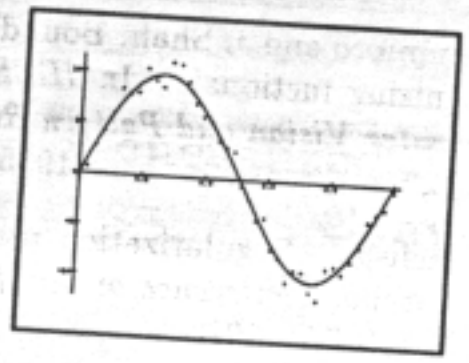

Figure 1: GCV chooses a "standard" value of λ, to interpolate sinusoidal data contaminated by Gaussian noise.

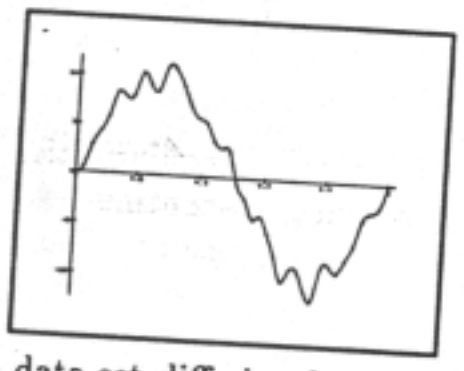

Figure 2: For a data set differing from that of Figure 1 in only one point, GCV chooses a very small value of λ, resulting in a completely different fit.

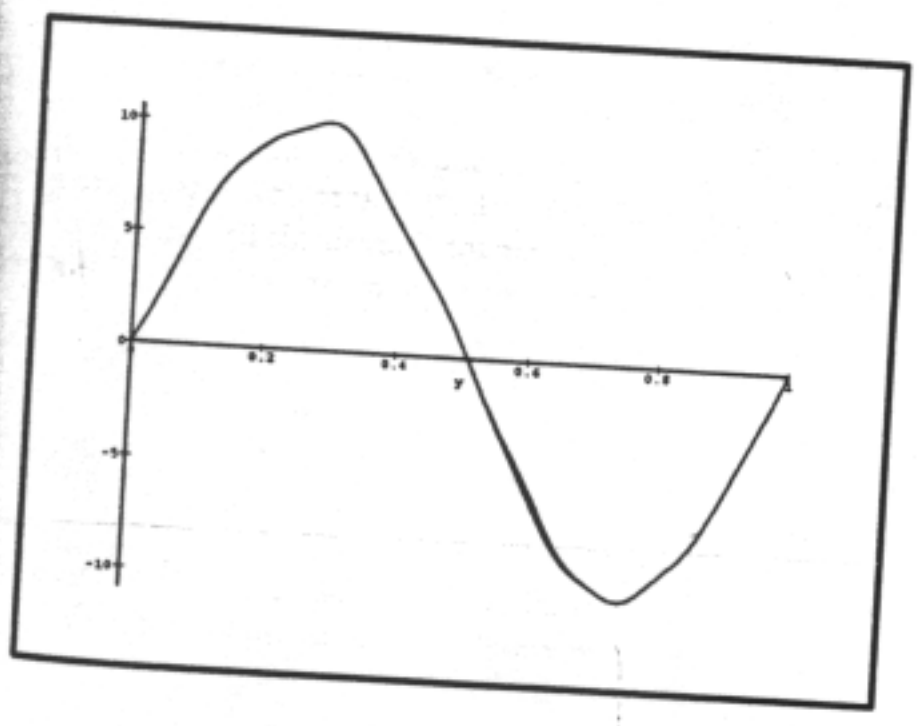

Figure 3: The suggested method for fitting, used on the data sets of Figures 1 and 2. Fits are almost identical.

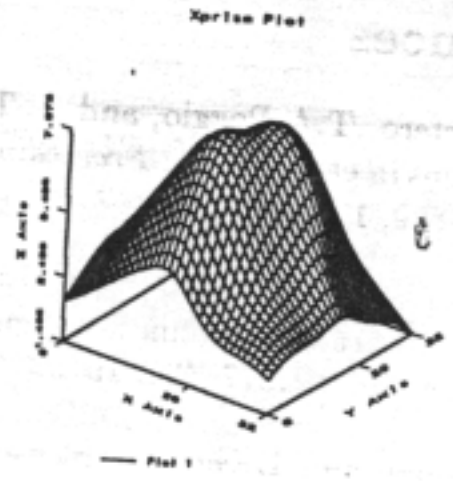

Figure 4: GCV reconstruction for the function $x(1-x)y(1-y)$, contaminated by Gaussian noise. This is a "typical" reconstruction for such data.

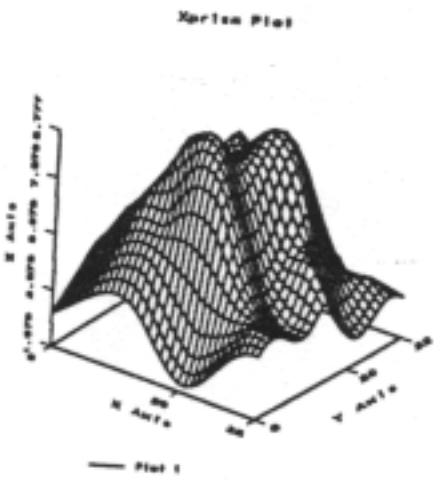

Figure 5: For a data set differing from that of Figure 4 in a single point, GCV finds a radically different interpolation.

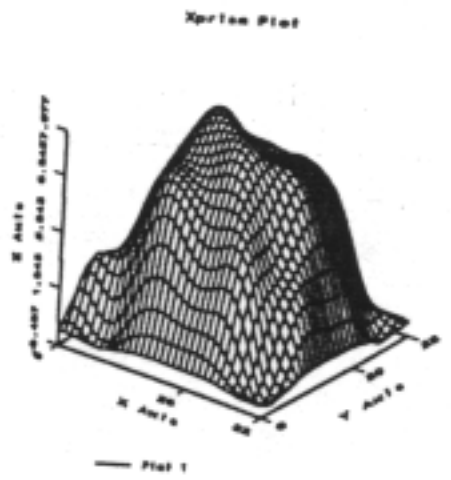

Figure 6: The suggested method for fitting, used on the data set of Figure 4.

References

[1] M. Bertero, T.A Poggio, and V. Torre. Ill-posed problems in early vision. *Proceedings of the IEEE*, 8:869–889, 1988.

[2] P. Craven and G. Whaba. Optimal smoothing of noisy data with spline functions. *Numerische Mathematik*, 31:377–403, 1979.

[3] S. Geman and D.Geman. Stochastic relaxation, gibbs distribution, and the bayesian restorat ion of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 6:721–741, June 1984.

[4] B. Horn. *Robot Vision*. MIT Press, 1986.

[5] B.K.P Horn and B.G. Schunck. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.

[6] D. Keren and M. Werman. Variations on regularization. In *10'th International Conference on Pattern Recognition*, Atlantic City, 1990.

[7] D. Keren and M. Werman. A bayesian framework for regularization. 1994. *submitted to IEEE Transaction on Pattern Analysis and Machine Intelligence.*

[8] D. Keren and M. Werman. Probabilistic analysis of regularization. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 15:982–995, October 1993.

[9] Y.G. Leclerc. Image and boundary segmentation via minimal-length encoding on the co nnection machine. In *Image Understanding Workshop*, pages 1056–1069, 1989.

[10] David J.C. MacKay. *Bayesian Methods for Adaptive Models*. PhD thesis, California Institute of Technology, 1992.

[11] J.L Marroquin. Deterministic bayesian estimation of markovian random fields with app lications to computational vision. In *International Conference on Computer Vision*, pages 597–601, London, May 1987.

[12] L. Matthies, R. Szeliski, and T. Kanade. Incremental estimation of dense depth maps from image sequences. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 366–374, Ann Arbor, June 1988.

[13] D. Mumford and J. Shah. Boundary detection by minimizing fuctionals. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 22–26, San Francisco, June 1985.

[14] R. Szeliski. Regularization uses fractal priors. In *National Conference on Artificial Intelligence*, pages 749–754, 1987.

[15] R. Szeliski. *Bayesian Modeling of Uncertainty in Low-Level Vision*. Kluwer, 1989.

[16] S. Szeliski and D. Terzopoulos. From splines to fractals. In *SIGGRAPH*, pages 51–60, 1989.

[17] D. Terzopoulos. Multi-level surface reconstruction. In A. Rosenfeld, editor, *Multiresolution Image Processing and Analysis*. Springer-Verlag, 1984.

[18] D. Terzopoulos. Regularization of visual problems involving discontinuities. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 8:413–424, August 1986.

[19] A.N Tikhonov and V.Y Arsenin. *Solution of Ill-Posed Problems*. Winston and Sons, 1977.