

Vision: Computational Theory to Neural Systems

Daniel Kersten

Winter, 1999

Lecture 17

1 Bayesian estimation, Regularization Theory and Vision

A central problem in the last half of this course is to understand how perceptual systems can make reliable estimates of scene properties such as shapes, distances, movements and material characteristics of objects from data consisting of luminance values as a function of space and time. So far, we've seen examples from motion flow estimation, shape-from-shading, and reflectance estimation (lightness). Here, we summarize the fundamental computational problems, and then show how the Bayesian framework unifies many of these early to intermediate-level vision problems.

What are the key computational problems? On the one hand, when we consider any particular source of visual information, the local image data often underconstrains the solution of the scene property. That is, there are too many interpretations of the world that could give rise to the same image. On the other hand, the problem of how to satisfy multiple constraints is not straightforward either. We've seen how requiring smoothness in the estimated scene properties can lead to unique solutions. But many constraints can lead to non-linear formulations through non-quadratic cost functions. Bayesian inference unifies a number of image understanding problems in early vision that have been discussed in previous lectures. Early theoretical work along these lines incorporated *Regularization theory*, which can be viewed as a special case of Bayesian inference. The "regularizers" are the smoothness operators, which we will formalize below as Bayesian priors. Certain forms of regularization are well-suited to neural network implementations (Poggio et al, 1985; Knill et al., 1991). The statistical framework provides tools to model multiple constraints. Geman and Geman's results (1984) on Markov Random Fields provided tools for establishing constraints and for solving non-linear scene- from-image problems (Kersten, 1990). Let's look at the formal structure of the "scene from image" problems we've studied so far.

The computer graphics problem of calculating an image function, i from a scene representation s can be represented:

$$i = As$$

where A may be a non-linear operator representing the generative rendering model. We've noted that modeling image formation is non-trivial, especially when one bases the model on the physics of imaging. However, from a mathematical point of view, imaging is well-posed. Given a representation of the scene, and a model of image formation, the image can, in principle, be computed. However, the inverse problem, of computing s from i , given A , is often ill-posed in the sense that there is not a unique s which satisfies the equation. Additional constraints have to be imposed in order to force a unique solution. In regularization theory, one incorporates a suitable constraint operator, P , which, together with a choice of norms, $\|\bullet\|$, serves to constrain the search for s . One approach is then to find s that minimizes:

$$\|As - i\|^2 + \lambda\|Ps\|^2$$

For example, in discrete standard regularization theory, s and i are vectors, A and P matrices and the norm is vector length. There are standard methods for computing solutions using the pseudoinverse matrix (see: Strang, 1988). The form of P is usually arrived at by a combination of heuristics, mathematical convenience and experiment. We have seen an example from motion measurement where we assumed optic flow fields calculated from contours are smooth almost everywhere along the contour. This heuristic can be quantified by using a derivative constraint operator, P on the velocity field (Hildreth, 1983). If the imaging operator, A is linear, then above equation is a quadratic form in s , and thus in general convex with no local minima.

2 Bayesian Inference

How can scene-from-image problems be formulated statistically? We can try to solve the inverse problem as maximum a posteriori estimation (MAP) (Geman and Geman, 1984; Marroquin, 1985). This can be illustrated by showing how to relate it to the cost functions of regularization theory. Let the image and scene parameters be represented discretely by vectors,

$$i = (i_1, i_2, \dots, i_n)$$

$$s = (s_1, s_2, \dots, s_n),$$

respectively. Suppose the probability of s conditional on i is known. To estimate a set of scene characteristics from an image, we must specify some criterion on the posterior distribution, $p(s|i)$, which defines an optimal solution. The two most common are the mode and the mean of the distribution. In the first case, the optimal solution is defined as the most probable scene given an image. This is referred to as *maximum a posteriori* estimation (MAP). MAP estimation minimizes the probability of error (van Trees, 1968).

In the second case, the optimal solution is the one which minimizes the mean squared error of the estimates. It is referred to as minimum mean squared error (MMSE) estimation. We noted early in the course, that Bayesian decision theory provides a general means to specify the task goals through risk and loss functions.

Suppose the action Σ is one of the possible guess of scene parameter s . A *loss function* $L(\Sigma, s)$ specifies the penalty for performing action Σ if the scene is s . The *risk* $R(\Sigma; i)$ of taking action Σ when the input is i is defined to be the expected loss:

$$R(\Sigma; i) = \sum_s L(\Sigma, s)P(s | i),$$

with respect to the a posterior probability, $P(s|i)$ of s . One possible loss function is that you pay penalty 0 for making any wrong interpretations and receive the reward +1 for all of the correct interpretations. In other words, $L(\Sigma, s) = -1$ if $\Sigma = s$ and = 0 otherwise. In this case the risk reduces to

$$R(\Sigma; i) = -P(\Sigma | i),$$

and hence the best strategy is to pick the most likely interpretation. This is then just *MAP estimation*. Other estimation strategies are sometimes better (e.g. Brainard and Freeman, 1997). (Note that decision theory is used in Signal Detection theory and adjusting payoff rewards can be used to affect the observers' decisions).

It is often difficult to write directly an expression for the conditional probability. However, Bayes rule enables us to break the probability into two parts:

$$p(s|i) = \frac{p(i|s)p(s)}{p(i)}$$

where $p(i|s)$ and $p(s)$ derive from the image formation and scene models, respectively. Since i is held fixed, while searching for s , $p(i)$ is constant. MAP estimation is equivalent to maximum likelihood estimation when the prior distribution, $p(s)$ is uniform. Given certain assumptions, MAP estimation is equivalent to regularization theory. If we assume:

$$i = As + noise$$

where the noise term is multivariate Gaussian with a (usually assumed) constant diagonal covariance matrix, then

$$p(i|s) = k \exp\left(-\frac{1}{2\sigma_n^2}(i - As)^T(i - As)\right)$$

where k is a normalization constant. Superscript T indicates transpose. Further, suppose $p(s)$ is multivariate Gaussian:

$$p(s) = \frac{\exp\left(-\frac{1}{2}s^T B s\right)}{\sqrt{2\pi|B^{-1}|}}$$

where B is the inverse of the covariance matrix μ ($\mu = B^{-1}$). And s is adjusted to have zero mean. By substitution and taking the logarithm, maximizing $p(s|i)$ is equivalent to minimizing the cost function:

$$(As - i)^T(As - i) + \lambda s^T B s \tag{1}$$

where λ is a Lagrange multiplier (e.g. equal to the ratio of the noise to scene variance). If the data are noise-free, MAP estimation is equivalent to minimizing

$$s^T B s$$

for those s that exactly satisfy

$$i = As$$

Thus, for quadratic norms, the MAP formulation gives the regularization cost function if the image and scene are represented by discrete vectors, and when $B = P^T P$.

When the image formation process is linear with Gaussian noise, and the prior distribution Gaussian, the cost function is quadratic (standard regularization theory). Thus, the cost function is convex, and gradient descent or

more efficient numerical techniques can be used to find the global minimum. Further, fast matrix operators can be implemented with neural hardware, avoiding the need for iteration. However, linear methods are extremely restrictive. They do not handle discontinuities, or many multiple constraint interactions, essential to understanding vision. Research has shown how the MAP formulation can be extended to solving non-linear problems involving discontinuities and multiple constraints. An example, we show an application of the MAP approach to the non-linear reflectance estimation problem. Learning also provides an alternative to sculpting the cost function landscape (next section).

Expressing regularization theory as a problem in statistical estimation has several advantages. A major problem, not only in early and intermediate-level vision, but in the programming of model neural networks generally, is sculpting the topography of the cost function. This, for example, amounts to finding a suitable B . We will look at two ways of doing it. One is to develop a language for directly programming in suitable constraints, and a second way is to learn the constraints. The statistical approach allows the constraint term to be based on verifiable scene statistics in addition to heuristics. However, it is difficult, in practice, to gather samples and an alternative is to develop statistical models of scene and image synthesis. Finding a good statistical model of natural scene parameters is a difficult problem in itself. Here we may benefit from the rapidly expanding field of computer synthesis of naturalistic scenes (e.g. Mandelbrot, 1977; Fournier et al., 1982), and the recent work in density estimation of textures (Zhu et al., 1997). We will see how the relationship between Markov Random Fields and Gibbs distributions provides potentially powerful tools for devising appropriate prior models of the scene which are more general than the Gaussian model.

Finding the MAP estimate, in general, is non-trivial. Gradient ascent can be used, but is often inefficient and has to deal with local minima. Mean field theory is one approach. But the statistical approach also shows that consideration of the uncertainty in the estimates may also be important. Algorithms for rapid convergence have been demonstrated for vision problems using techniques from Bayes nets and graphical models (Weiss, 1997).

3 Learning scene attributes from images by example

So far, we have considered a formulation of the ideal observer independent of the algorithm or hardware. It is useful to consider suboptimal observers that are optimal except that they are constrained to compute an approximation in a particular way. For example, the energy or cost function topography, could be sculpted by a particular associative learning algorithm. The probabilistic model of scenes provides input/output pairs to use associative algorithms that can learn to estimate scene parameters from images (Kersten et al., 1987; Hurlbert and Poggio, 1988; Lehky and Sejnowski, 1988; Knill and Kersten, 1990). This becomes particularly interesting when it is difficult to compute the posterior distribution, and we have a complete description of the prior distribution and the image formation model. We may not be able to directly address the optimality of the learning algorithm, but we may find out whether a given architecture is capable of computing the desired map to within some desired degree of accuracy. For example, suppose \mathbf{A} and \mathbf{B} in equation (1) are linear, but unknown. Then, the generalized inverse of \mathbf{A} , \mathbf{A}^* , is the matrix (in general, rectangular) mapping \mathbf{i} to \mathbf{s} , which minimizes the squared error over the set of training pairs $\mathbf{s}_k, \mathbf{i}_k$. With training, this approaches the MAP estimate when A is linear and the prior distribution and noise are Gaussian. \mathbf{A}^* can be estimated by associative learning over the training set using Widrow-Hoff error correction. (Duda and Hart, 1973; Kohonen, 1984). Recent developments in more powerful associative algorithms, such as error back-propagation (Rumelhart & McClelland, 1986) may broaden the usefulness of learning constraints. On the other hand, even if \mathbf{A} and \mathbf{B} are non-linear, it is worth finding out how well a linear inverse operator approximates the optimal solution.

In an application of this idea to shape from shading, Knill and Kersten (1990) estimated pseudoinverse mapping (Albert, 1972) from images to surfaces

$$\mathbf{N} = \mathbf{A}^*\mathbf{L},$$

such that the average error between the estimated surface normals, represented by a vector \mathbf{N} (made up of pairs of the x and y components of surface normals at each spatial location), and the real surfaces was minimized. The image formation constraint was based on a lambertian model, with a known

3 LEARNING SCENE ATTRIBUTES FROM IMAGES BY EXAMPLE7

point source illumination, and assumed constant reflectance

$$L(x, y) \propto \mathbf{n} \cdot \mathbf{e}.$$

Rather than image intensities, luminance at each point, normalized by the average value, was used as image data. When the imaging function is linear and the distribution of \mathbf{N} is Gaussian, the optimal mapping is linear. For shape from shading, neither of these conditions holds. A linear mapping may, however, be near optimal. If the sample surfaces are randomly drawn from the prior distribution, $p(\mathbf{N})$; span its space, and the images are calculated using the lambertian image formation function, the derived mapping is the best linear estimate of the mean of $p(\mathbf{N}|\mathbf{L})$. The actual form of the distribution need not be known, as it is implicitly defined by the set of sample surfaces. An optimal mapping for a particular space of surfaces may thus be learned through appropriate selection of samples.

Given pairs of sample surfaces and images, the Widrow-Hoff algorithm was used to derive a linear mapping between them which minimizes the mean squared error between estimated and real surfaces

$$\mathbf{A}^*_{k+1} = \mathbf{A}^*_k + \rho(\mathbf{N}_k - \mathbf{A}^*_k \mathbf{L}_k) \mathbf{L}_k^T.$$

where \mathbf{A}^* is the mapping being learned. The term in parentheses is the error between estimated and actual values of \mathbf{N}_k . Iterative application of this rule to example vectors, \mathbf{N}_k and \mathbf{L}_k , with appropriate relaxation of the learning constant, ρ , (typically $1/k$) will result in a convergence of \mathbf{A}^* to the desired mapping.

One could either use examples of real surfaces as the training set, or draw them from a statistical model of surfaces. The latter approach avoids the data collection problems of using real surfaces, and has the advantage of providing a tool for analyzing the performance of the shape from shading estimator. Because of the potential usefulness of fractal models to describe surfaces (Pentland, 1986), a statistical fractal model was used to generate surfaces for the training set. If the surface statistics are assumed to be spatially invariant, the linear estimators of surface normal in the x and y directions are just convolution filters as applied to the normalized luminance values of the image. The filters learned were spatially oriented bandpass filters. The figure shows results of an original and reconstructed surface. In general, simulations showed that the linear approximation worked well for a wide range of surfaces, including non-fractals.

For recent work on learning scenes from images, see Freeman and Pasztor (1999).

References

- [1] Albert, A. (1972). **Regression and the Moore-Penrose Pseudoinverse**. New York: Academic Press.
- [2] Brainard, D. H., & Freeman, W. T. (1997). Bayesian color constancy. *J Opt Soc Am A*, **14**, (7), 1393-411.
- [3] Bülthoff, H. H., & Yuille, A. (1991). Bayesian models for seeing surfaces and depth. *Comments on Theoretical Biology*, **2**, (4), 283-314.
- [4] Clark, J. J., & Yuille, A. L. (1990). **Data Fusion for Sensory Information Processing**. Boston: Kluwer Academic Publishers.
- [5] Duda, R. O., & Hart, P. E. (1973). **Pattern classification and scene analysis**, New York: John Wiley & Sons.
- [6] Freeman, W. T., & Pasztor, E. C. (1999). Learning to estimate scenes from images. In M. S. Kearns, S. A. S. a. D. A. C. (Ed.), *Adv. Neural Information Processing Systems 11* Cambridge MA: MIT Press.
- [7] Fournier, A., Fussell, D., & Carpenter, L. (1982). Computer Rendering of Stochastic Models. *Graphics and Image Processing. Communications of the ACM*, **25**, (6), 371-384.
- [8] Geiger, D., & Girosi. (1991). Parallel and Deterministic Algorithms from MRF's: Surface Reconstruction. *I.E.E.E PAMI*, **13**, (5).
- [9] Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Transactions Pattern Analysis and Machine Intelligence*, **PAMI-6**, 721-741.
- [10] Hurlbert, A.C. and Poggio, T. A. (1988) Synthesizing a color algorithm from examples. *Science*, 239, 482-485.

- [11] Kersten, D., O'Toole A., Sereno, M., Knill D., Anderson, J.A., Associative learning of scene parameters from images. *Applied Optics*, 1987, 26, 4999- 5006.
- [12] Kersten, D. (1990). Statistical limits to image understanding. In Blake-more, C. (Ed.), *Vision: Coding and Efficiency*, Chapter 3, Cambridge: Cambridge University Press.
- [13] Knill D. C. and Kersten, D. (1990) Learning a near-optimal estimator for surface shape from shading. *Computer Vision, Graphics and Image Processing*.
- [14] Knill, D. C., & Kersten, D. K. (1991). Ideal Perceptual Observers for Computation, Psychophysics, and Neural Networks. In Watt, R. J. (Ed.), *Pattern Recognition by Man and Machine*(pp. 83-97). MacMillan Press.
- [15] Kohonen, T. (1978). **Associative Memory: A System Theoretic Approach.**, Berlin: Springer Verlag.
- [16] Lehky, S. R. and Sejnowski, T. J. (1988) Network model of shape from shading: neural function arises from both receptive and projective fields. *Nature*, 333, 452-454.
- [17] Mandelbrot, B. B. (1977). **Fractals: Form, Chance and Dimension**, . W.H. Freeman, San Francisco.
- [18] Marroquin, J. L. (1985). Probabilistic solution of inverse problems. M.I.T. A.I. Tech Memo 860.
- [19] Papoulis, A. (1965). **Probability, Random Variables, and Stochastic Processes**. New York: McGraw-Hill Book Company.
- [20] Poggio, T., Torre, V., & Koch, C. (1985). Computational vision and regularization theory. *Nature*, **317**,314-319.
- [21] Strang, G. (1988). **Linear Algebra and Its Applications**, (3rd ed.). Saunders College Publishing Harcourt Brace Jovanovich College Publishers.

- [22] Szeliski, R. (1990). Bayesian modeling of uncertainty in low-level vision. *International Journal of Computer Vision*, **5**, (3), 271-301.
- [23] Van Trees, H. L. (1968). **Detection, Estimation and Modulation Theory**, New York: John Wiley and Sons.
- [24] Weiss, Y. (1997). Interpreting images by propagating Bayesian beliefs. In M.C. Mozer, M. I. J. a. T. P. (Ed.), *Advances in Neural Information Processing Systems 9*(pp. 908-915). Cambridge MA: MIT Press.
<http://www-bcs.mit.edu/people/yweiss/nips96.pdf>
- [25] Yuille, A. L., Geiger, D., & Bülthoff, H. H. (1991). Stereo integration, mean field theory and psychophysics. *Network*, **2**, 423-442.
- [26] Zhu, S. C., Wu, Y., & Mumford, D. (1997). Minimax Entropy Principle and Its Applications to Texture Modeling. *Neural Computation*, **9**, (8), 1627-1660.