Course: Vision as Bayesian Inference. Lecture 2.

Alan Yuille
Department of Statistics, UCLA
Los Angeles, CA 90095
yuille@stat.ucla.edu

Abstract

Piecewise smooth models. Markov Random Fields. EM. Mean Field Theory.
NOTE: NOT FOR DISTRIBUTION!!

1 Introduction

In this lecture we describe another way to segment images which complements edge detection. While edge detection concentrates on finding places where the intensity changes rapidly we now attempt to group pixels into regions with similar intensity values. This work is based on three similar, but independent, models developed by Geman and Geman, Mumford and Shah, and Blake and Zisserman. A later, and influential model, is by Rudin-Osher-Fatemi (partially because it is computationally simpler and has an efficient algorithm). These models essential assume that images are piecewise smooth (or weak membrane – explain the relationship to membranes. This is an example of a prior assumption which was rapidly applied to other properties such as depth, motion flow, and disparity. There is now evidence (which we will discuss) which relates this prior assumption to empirical statistical analysis of images, depth, and motion flow.

Only the model by Geman and Geman has an explicit probabilistic formulation – so we will concentrate on it (Blake and Zisserman is also straightforward to reformulate in these terms). The Mumford-Shah and Rudin-Osher-Fatemi models are formulated in terms of energy functionals defined on continuous $\vec{x}$ space rather than a lattice. This can make the probabilistic interpretation more handwaving than rigorous as we will describe.

As we will discuss, it is debatable whether some of these models are realistic enough to be very useful in practice. But, in any case, they are historically important and serve as a good way to introduce many concepts and algorithms. The probabilistic models can be extended to a larger class of Markov Random Field models (ADD CAVEATS) and we will put the work in this bigger context towards the end of these notes.

2 The Piecewise Smooth/Weak Membrane Model

We assume that the observed image I is a corruption of an ideal image J (one application, probably the best one, is to enhancing degraded images). There is a prior on J which includes line processes l – this prior favors images J which are piecewise smooth, see figure (1). In terms of the big picture, $W = (J, l)$, and this is a generative model $P(I|W)P(W)$. The model is formulated in one-dimension (for simplicity) but is easily extended to two, or higher, dimensions. (Simple way to do this – for next lecture!!).

The likelihood function is:

$$P(I|J) = \prod_{i=1}^{N} \frac{1}{Z_i} \exp\{-C(I_i - J_i)^2\} = \frac{1}{Z_i} \exp\{-\sum_{i=1}^{N} C(I_i - J_i)^2\}, \tag{1}$$

This says that the image pixels are generated independently from the ideal image by additive zero-mean Gaussian noise (with variance $\sigma^2 = 1/(2C)$ – $Z_i = \sqrt{2\pi\sigma^2}$). This is the natural model to choose but can be modified to

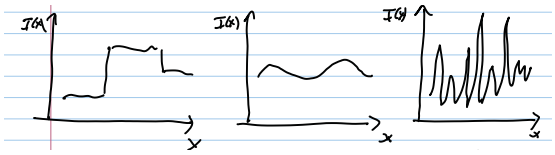

Figure 1: Left Panel: a piecewise smooth image. Center Panel: a smooth image. Right Panel: an un-smooth image

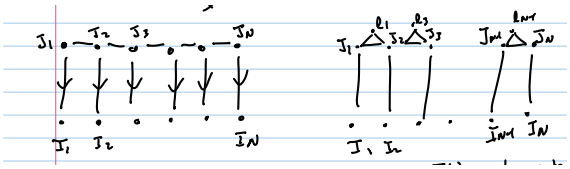

Figure 2: Left Panel: graphical representation of the Gaussian model. Right Panel: graphical representation of the weak membrane model.

other types of noise (e.g. shot noise, and so on – see Appendix for non-robustness of Gaussian). We use the identity $\prod_i \exp\{K_i\} = \exp\{\sum_i K_i\} \; \forall \; \{K_i\}$ to represent this as a *Gibbs distribution* of form $P(\vec{x}) = (1/Z)\exp\{-E(\vec{x})\}$ in the second line. (Properties of Gibbs distributions!!).

The weak membrane prior expresses piecewise smoothness and is of form:

$$P(J,l) = \frac{1}{Z}\prod_{i=1}^{N-1}\exp\{-A(J_{i+1}-J_i)^2(1-l_i) - Bl_i\} = \frac{1}{Z}\exp\{-\sum_{i=1}^{N-1}\big(A(J_{i+1}-J_i)^2(1-l_i) + Bl_i\big)\}. \quad (2)$$

The J_i are continuous valued variables and the line process variables are binary-valued $l_i \in \{0,1\}$. This is an improper prior in the sense that we cannot normalize the distribution (i.e., Z is infinite). See figure (2) for graphical representations of the model without line processors (Gaussian model) and the full weak membrane model.

Firstly, we consider this model when the line process variables $\{l_i\}$ are all set to zero. In this case the prior reduces to:

$$P(J) = \frac{1}{Z}\exp\{-\sum_{i=1}^{N-1} A(J_{i+1}-J_i)^2\}, \quad (3)$$

which is a multivariate Gaussian.

Together the likelihood in equation (1) and the Gaussian Prior in equation (3) define a generative model. In our *big picture* terminology, $W = J$ is the ideal image – and these models define $P(I|W)$ and $P(W)$.

The posterior distribution $P(J|I)$ can be expressed in Gibbs form:

$$P(J|1) = \frac{1}{Z}\exp\{-E[J;I]\},$$
$$\text{where } E[J;I] = \sum_{i=1}^{N-1} A(J_{i+1}-J_i)^2 + \sum_{i=1}^{N} C(J_i - I_i)^2\}. \quad (4)$$

This model can be easily understood. The second term tries to keep the value of J_i close to the observed intensity I_i while the first (prior) terms tries to keep J_i similar to the values of its neighbors J_{i+1}, J_{i-1}. The result is to make J a spatially smoothed version of I.

The MAP estimate $J^* = \arg\max P(J|I)$ is equivalent to $J^* = \arg\min E[J;I]$. Because $E[J;I]$ is quadratic in J, the solution is obtained by taking the derivatives of $E[J;I]$ with respect to J and solving the following linear equations (MAP estimation on Gaussian distributions reduces to solving linear equations):

$$C(J_i - I_i) + A(2J_i - J_{i+1} - J_{i-1}) = 0, \; i = 2, ..., N-1$$
$$C(J_1 - I_1) + A(J_1 - J_2) = 0, \quad C(J_N - I_N) + A(J_N - J_{N-1}) = 0. \quad (5)$$

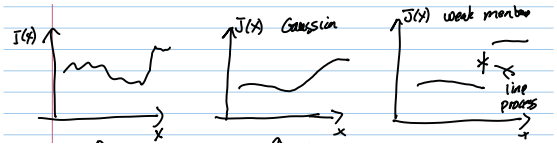

Figure 3: Left Panel: input image. Center Panel: output of Gaussian model. Right Panel: output of weak membrane model

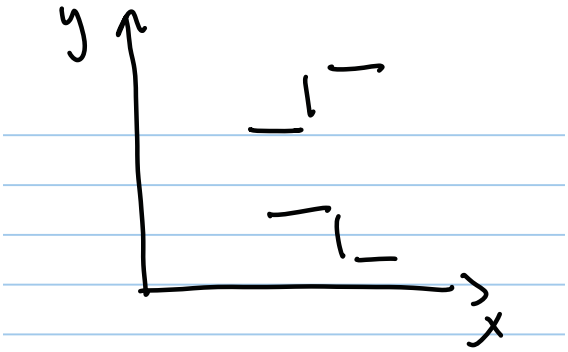

Figure 4: In two-dimensions the line processes are coupled to the neighbors to encourage long continuous lines.

This Gaussian model is easy to analyze, has a simple inference algorithm (i.e., solve these linear equations), and can also be learnt (we discuss learning later!!). But overall it is not a very useful model – the main problem is that it smooths images and, in particular, smoothes out edges, see figure (3). It is a smooth model and not piecewise smooth.

Secondly, we consider the weak membrane model with the linear processes. This gives a posterior distribution (technically an improper posterior since the normalization factor cannot be computed):

$$P(J, l|I) = \frac{1}{Z} \exp\{-E[J, l; I]\},$$

$$\text{where } E[J, l; I] = \sum_{i=1}^{N-1} \{A(J_{i+1} - J_i)^2(1 - l_i) + Bl_i\} + \sum_{i=1}^{N} C(J_i - I_i)^2\}. \tag{6}$$

MAP estimation of $(J^*, l^*) = \arg\max_{(J,l)} P(J, l|I)$ corresponds to minimizing the energy function $E(J, l; I)$ with respect to (J, l). Studying $E(J, l; I)$ gives the intuition for the weak membrane model. Each J_i wants to keep close to the data I_i and its neighbors J_{i+1}, J_{i-1} but if one of its neighbors takes too different a value – e.g., $|J_{i+1} - J_i|$ is very large – then the energy favors switching on the line process by setting $l_i = 1$ and breaking the connection between J_i and J_{i+1}. Hence the prior will smooth values of $\{I_i\}$ if they are fairly similar to their neighbors but will break the smoothness and activate an edge $\{l_i\}$ if they differ a lot. This results in J being a piecewise smoothed version of I, see figure (3).

Extending this model to two or more dimensions is straightforward allowing each pixel to be connected to its neighbors in all spatial directions – see figure (4). But there is an important difference. In one-dimensions activating a line process put a break which decomposes the image into disjoint parts but in two dimensions there is no guarantee that the line processes will link up to form a closed contour that divides the image into disjoint pieces. This has lead to the introduction of additional terms in the prior that couple adjacent line processes – so that if one line process is activated then it becomes energetically more favorable to activate its neighboring line processes (will be discussed later!!).

3 Algorithms for Weak Membrane

Performing inference on weak membrane models is difficult because it involves minimizing an energy function $E(J, l|I)$ which is non-convex – see figure (7) – and contains both continuous variables J and discrete variables l.

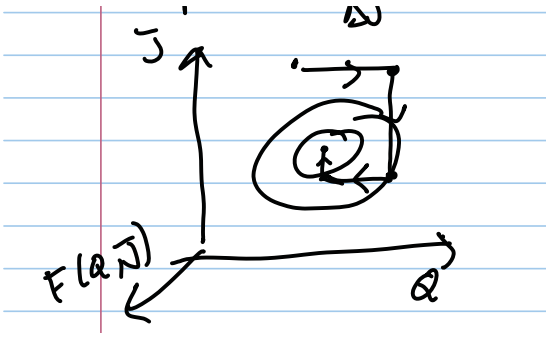

Figure 5: EM proceeds by coordinate descent in $F(J, q)$. The algorithm decreases $F(J, q)$ with respect to J keeping q fixed, then decreases $F(J, q)$ with respect to q keeping J fixed, and repeats until convergence.

It is not surprising that Geman and Geman invented simulated annealing (independently from Kirkpatrick). Blake and Zisserman developed a continuation method called graduated non-convexity which involve defining a family of different energy functions and which, surprisingly, relates to annealing.

We will start my methods that transform this problem into minimizing a function of continuous valued variables (EM and elimination). Then we will discuss continuation methods.

3.1 EM, Free Energy

We now present the Expectation Maximization (EM) algorithm for estimating J from $P(J|I)$. EM has many advantages to the previous section: (i) it can be used even when we cannot analytically perform the summation $P(J|I) = \sum_l P(J, l|I)$, (ii) it gives soft (i.e., probabilistic) estimates of l, (iii) it has many more extensions and applications. We follow the free energy formulation of EM (Neal and Hinton) rather than the original derivation (Dempster et al.).

We define a *free energy* $F(J, q)$:

$$\begin{aligned} F(J, q) &= -\log P(J|I) + \sum_l q(l) \log \frac{q(l)}{P(l|I, J)} \\ &= +\sum_l q(l) \log q(l) - \sum_l q(l) \log P(J, l|I), \end{aligned} \tag{7}$$

where $q(l)$ is a probability distribution on l (satisfying $\sum_l q(l) = 1$). The term $\sum_l q(l) \log \frac{q(l)}{P(l|I,J)} = D(q||P)$ is the *Kullback-Leibler divergence*. It attains the value 0, if and only if $q(l) = P(l|I, J)$.

Minimizing $F(J, q)$ with respect to J and q is the same as minimizing $-\log P(J|I)$ with respect to J. (Because of properties of the Kullback-Leibler divergence).

The EM algorithm is obtained by coordinate descent of $F(J, q)$ – i.e., fix q and minimize wrt J, then fix J and minimize wrt q, and repeat, see figure (5). This gives an update rule:

$$q^{t+1}(l) = P(l|I, J^t), \quad J^{t+1} = \arg\min_J \sum_l q^{t+1}(l) \log P(J, l|I). \tag{8}$$

This converges to a solution $J^*, q^*(l)$. But the free energy may have many local minima and so the algorithm can converge to any one of them, see figure (6). Convex functions, see figure (7), have only a single minima (caveats). It is often easy to check whether functions are convex or not. But convexity may be too strong a requirement.

What does the EM algorithm do for the weak membrane model? Substituting from the weak membrane model (equations!!), we find that $q(L)$ is a factorized distribution $\prod_i q_i(L_i)$, where

$$q_i^{t+1}(l_i) = \frac{1}{Z_i} \exp\{-A(J_i^t - J_{i+1}^t)^2(1 - l_i) - Bl_i\}, \tag{9}$$

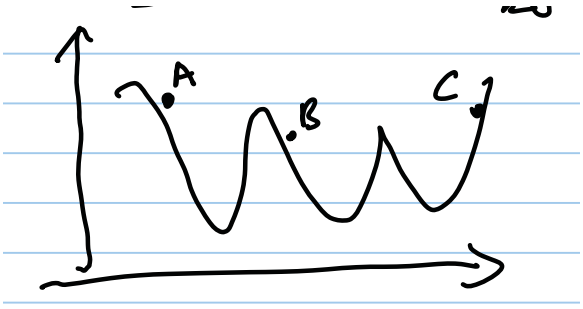

Figure 6: An energy function with local minima is hard to minimize. Steepest descent algorithms will converge to different minima depending on their starting points – e.g., A, B or C.

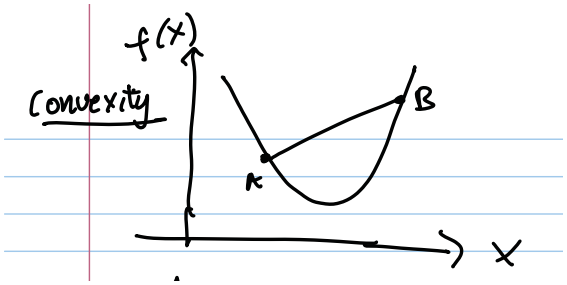

Figure 7: A function is convex if for all points A, B, the line connection A to B lies entirely within the region bounded by the function.

$$J^{t+1} = \arg\min \sum_i \{A(J_{i+1} - J_i)^2(1 - q_i^{t+1}) + C(J_i - I_i)^2, \tag{10}$$

where $q_i^{t+1} = \sum_{l_i} q_i^{t+1}(l_i) l_i = q_i^{t+1}(l_i = 1)$.

The E-step is given analytically by equation (9) and the M-step is a set of linear equations which can be solved by standard techniques. The EM algorithm is very intuitive for the weak membrane model. It consists of solving a series of Gaussian models (by the M-step) where the variances of the prior neighborhood terms are updated and vary with the pixels (i.e, $A \mapsto A(1 - \sum_{L_i} q_i^{t+1}(L_i)L_i)$). The variances become bigger – and hence the connections smaller – the more likely there is an edge.

Notice that the EM algorithm gives a way to minimize $E_{eff}(J)$ (from previous section) in an iterative two step manner. We will show (later section) that this is an example of a very general technique for finding discrete iterative algorithms for minimizing energy functions.

3.2 Eliminating the Line-Process Analytically

In very special cases – like this model – it is possible to remove the line processes l by summing them out to obtain $P(J|I) = \sum_l P(J, l|I)$. This is justified by Bayesian decision theory if the loss function depends only on J (i.e., $W = (J, l)$, $\vec{\alpha}(I) = (\alpha_1(I), \alpha_2(I))$ where $\alpha_1(I), \alpha_2(I)$ estimate J, l respectively, but the loss function $L(\vec{\alpha}(I), W) = -\delta(\alpha_1(I) - J)$ depends only on J – in other words, we pay no penalty if our estimate of l is incorrect, but we pay infinite penalty if we estimate J incorrectly). Surprisingly, this summation wrt l can be performed analytically for this model (Geiger and Girosi).

To perform this summation, observe that we can express $P(J, l|I)$ as a factorized form in l:

$$P(J, l|I) = \frac{1}{Z} \exp\{-C \sum_{i=1}^{N} (I_i - J_i)^2\} \times \prod_{i=1}^{N-1} \exp\{-A(J_i - J_{i+1})^2(1 - l_i) - Bl_i\}. \tag{11}$$

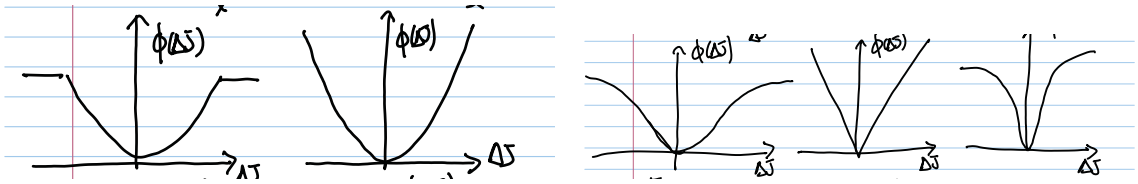

Figure 8: Potential figures. Left Panel:

This enables us to perform the summations over each l_i independently using the results:

$$\sum_{l}\prod_{i=1}^{N-1}\exp\{-A(J_i-J_{i+1})^2(1-l_i)-Bl_i\}=\prod_{i=1}^{N-1}\sum_{l_i}\exp\{-A(J_i-J_{i+1})^2(1-l_i)-Bl_i\}$$

$$=\prod_{i=1}^{N-1}\big(\exp\{-A(J_i-J_{i+1})^2\}+\exp\{-B\}\big)=\prod_{i=1}^{N-1}\exp\{\log\big(\exp\{-A(J_i-J_{i+1})^2\}+\exp\{-B\}\}. \quad (12)$$

This gives:

$$P(J|I)=\frac{1}{Z}\exp\{-C\sum_{i=1}^{N}(I_i-J_i)^2+\sum_{i=1}^{N-1}\phi(J_i,J_{i+1})\},$$
$$\text{where }\ \phi(J_i,J_{i+1})=\log\big(\exp\{-A(J_i-J_{i+1})^2\}+\exp\{-B\}\big). \quad (13)$$

Observe the different forms of the potential functions linking the neighbors: (i) $A(J_i-J_{i+1})^2$, (ii) $A(J_i-J_{i+1})^2(1-l_i)+Bl_i$, (iii) $\big(\exp\{-A(J_i-J_{i+1})^2\}+\exp\{-B\}\big)$. The main point is that the first potential term (Gaussian) pays an extremely large energy penalty as the difference $|J_i-J_{i+1}|$ gets large. This relates to non-robustness of Gaussians (see ROBUST SECTION). By contrast, the other two potentials have energy potentials which increase up to a maximum value of B – the third potential can be thought of as soft version of the second. Are these the best potentials? Not necessarily! All these potentials are quadratic for small $|J_i-J_{i+1}|$ which means that they only pay tiny penalties for small fluctuations and so samples from this prior would not be very smooth (and relate to Brownian processes – Szeliski). Better potentials would be sharper at the origin. An alternative potential (discussed later) is simply $|J_i-J_{i+1}|$ which is less non-robust than the Gaussian, sharper for small $|J_i-J_{i+1}|$, and has computational advantages (convexity – discussed later). Finally, of course, the bottom-line is what prior best corresponds to the empirical data (learning – discussed later).

One advantage of eliminating the line process l is that MAP estimation of J reduces to minimizing an energy function $E(J;I)$ which is a function of a continuous variable J only. This means that standard optimization techniques (steepest descent, Newton-Raphson) can be used. Unfortunately they are not guaranteed to find the global minimum of the energy function and the solution they find will depend on the initial conditions. Fortunately there is a natural set of initial conditions by setting the initial value of J to be the intensity image I – i.e., $J^{t=0}=I$ – which helps mitigate this.

But this approach only outputs an estimate J^* of the ideal image and does not give the state of the line process variables.

3.3 The Weak Membrane and the Mean Field Approximation

The simple analytic form of the EM algorithm in the previous section is due to the factorizable form of the probability distribution $P(J,l|I)$. We new discuss how this can be extended to cases where the distributions are not factorized. This involves an approximation called mean field theory (which relates to variational methods – Jordan).

Extend the weak membrane model to two-dimensions. The pixels are labeled by i and there is a neighborhood structure $j\in N(i)$. We denote $l_{i,j}$ to be the line process between neighboring pixels i,j. We introduce coupling $\phi_{(j-i,k-i)}(l_{i,j},l_{i,k})$ between the neighboring line processes which encourages the formation of continuous lines. We

specify the distribution in Gibbs form:

$$P(J,l|I) = \frac{1}{Z}\exp\{-E(J,l;I)\}, \tag{14}$$

$$E(J,l;I) = C\sum_i (J_i - I_i)^2 + A\sum_i \sum_{j\in N(i)} (J_i - J_j)^2(1-l_{i,j}) + B\sum_i \sum_{j\in N(i)} l_{i,j} + \sum_i \sum_{j\in N(i),k\in N(i),j\neq k} \phi_{(j-i,k-i)}(l_{i,j},l_{i,k}). \tag{15}$$

The coupling terms $\phi_{(j-i,k-i)}(l_{i,j},l_{i,k})$ make it impossible either to sum out the $\{l_{(i,j)}\}$ analytically or compute the distribution $P(l|J,I)$ analytically (in order to update $q(l)$ during the EM algorithm). But there is an approximation (mean field theory/variational) which can be used here (and which can be applied to a large class of problems – see later section – also has some relations to neural network models).

The Free Energy is expressed as:

$$F(J,q(.)) = \sum_l q(l)\log q(l) - \sum_l q(l)\log P(J,l|I). \tag{16}$$

To obtain the EM algorithm we should minimize $F(J,q(.))$ wrt J and $q(.)$. But the minimization wrt $q(.)$ is impractical. Instead we approximate the minimization by restricting t $q(.)$ to be a factorizable distribution $q(l) = \prod_i q_i(l_i)$ (for the models earlier this lecture $q(.)$ is automatically factorized). We then minimize $F(J,q(.))$ approximately by restricting the form of $q(.)$. This leads to minimizing a function:

$$F_{MFT}(J,q(.)) = \sum_{i,j}\sum_{l_{i,j}} q_{i,j}(l_{i,j})\log q_{i,j}(l_{i,j}) + C\sum_i (J_i - I_i)^2 + A\sum_i \sum_{j\in N(i)} (J_i - J_j)^2(1- < l_{i,j} >) + B\sum_i \sum_{j\in N(i)} < l_{i,j} > + \sum_i \sum_{j\in N(i),k\in N(i),j\neq k} \phi_{(j-i,k-i)}(< l_{i,j} >, < l_{i,k} >), \tag{17}$$

where $< l_{i,j} >$ denotes the expectation of $l_{i,j}$ with respect to $q_{i,j}(l_{i,j})$.

Algorithms for minimizing $F_{MFT}(J,q(.))$ will be given in the next lecture.

Appendix A

Appendix A1: Robustness

This section is an aside which is motivated by the line process discussion. Any distribution $P(W|I)$ used to model vision will inevitably be an approximation and so the MAP (or other) estimates will inevitably have some errors. Are there classes of models $P(W|I)$ for which the errors are likely to be serious?

Statisticians have developed the field of *Robust Statistics* to address these issues (Huber) by determining which distributions are particularly sensitive to *outliers* which are generated by some other process. More precisely, suppose the data is generated by a distribution $(1-\epsilon)P(\vec{x}) + \epsilon Q(\vec{x})$ but we use distribution $P(\vec{x})$ to analyze it. All data coming from $Q(\vec{x})$ are contaminants which will cause errors. It is known that the Gaussian distribution is particularly sensitive to outliers because it pays a large penalty (due to the quadratic exponent) for any data that is significantly different from the rest.

To be more precise, consider a one-dimensional Gaussian $P(x|\mu,\sigma) = \frac{1}{sqrt2\pi\sigma}\exp\{-(x-\mu)^2/(2\sigma^2)\}$. Learning the model parameters μ,σ by Maximum Likelihood (ML) from data $\{x_i : i = 1,...,N\}$ reduces to solving $(\mu^*,\sigma^*) = \arg\max_{(\mu,\sigma)} \prod_{i=1}^N P(x_i|\mu,\sigma)$ which can be solved analytically to yield $\mu^* = (1/N)\sum_i x_i$ and $\sigma^{2,*} = (1/N)\sum_{i=1}^N$. The problem arises if some of the data $\{x_i\}$ is contaminated (i.e., not generated by $P(x)$) in which case it can have a very large effect on the estimates of μ and σ (one outlier can ruin your whole dataset). The problem is traced to the

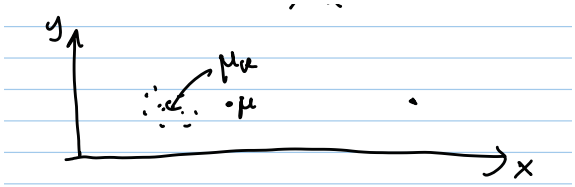

Figure 9: A single outlier (to the extreme right) can contaminate the mean and give estimate μ while the true value is μ_R.

Figure 10: Left Panel: the Gaussian has a quadratic potential and so is non-robust. Right Panel: a better potential assymptotes with distance – but may require more computation.

quadratic energy $(1/2\sigma^2)(x-\mu)^2$ so that outliers pay quadratic penalties. Alternatively, the Gaussian has *short tails* because the distribution $P(x|\mu,\sigma)$ decreases very very rapidly as $|x-\mu|$ gets large. See figure (9) and figure (10).

Most research has concentrated on the difficulties of dealing with outliers x_O such that $P(x_O)$ is small (i.e., data which has very low probability given the model). There are two strategies: (a) propose models that have long tails (e.g., DISTRIBUTIONS INSTEAD OF GAUSSIANS), (b) have an explicit process that detects outlier and rejects them.

Our study of the weak membrane model shows that these two strategies are highly related. The weak membrane potential (with line process) can be seen as mixture model where the interaction between neighbors is either a Gaussian (if $l_i = 0$) or a uniform distribution (if $l_i = 1$).

One way to fix this is to assume that the data is generated by a mixture of distributions $(1-\epsilon)P(x) + \epsilon Q(x)$ where $P(x)$ is the basic process that generates the data and $Q(x)$ is the outlier process. The difficulties are: (i) what value to select for ϵ, (ii) what distribution to select form $Q(.)$, and (iii) how to perform inference/learning on the mixture model.

A2: Relations to Neuroscience

A mean field version of the weak membrane model was proposed (Koch, Marroquin, Yuille 1987) as a model for how the part of V1 processes images. A more realistic variant was proposed by Lee (1996). This is discussed in the review chapter (Lee and Yuille 2007) which can be downloaded from the website. Some of Lee's multielectrode recording experiments seem consistent with a model of this type. But who knows what V1 really does...