models. Notice as pointed out earlier that this modern Bayesian framework includes no emphasis on defining the 'right' prior $\mathcal{R}$ with which we ought to interpolate. Rather, we invent as many priors (regularisers) as we want, and allow the data to tell us which prior is most probable.

### Evaluating the evidence for $\mathcal{A}, \mathcal{R}$

As $\alpha$ and $\beta$ vary, a single evidence maximum is obtained, at $\hat{\alpha}, \hat{\beta}$ (at least for quadratic $E_D$ and $E_W$). The evidence maximum is usually well approximated[7] by a separable gaussian, and differentiating (21) twice we obtain gaussian error bars for $\log\alpha$ and $\log\beta$:

$$(\Delta \log \alpha)^2 \simeq 2/\gamma$$
$$(\Delta \log \beta)^2 \simeq 2/(N-\gamma)$$

Putting these error bars into (26), we obtain the evidence.[8]

$$P(D|\mathcal{A},\mathcal{R}) \simeq P(D|\hat{\alpha},\hat{\beta},\mathcal{A},\mathcal{R}) P(\hat{\alpha},\hat{\beta})\, 2\pi\, \Delta\log\alpha\, \Delta\log\beta \tag{27}$$

How is the prior $P(\hat{\alpha},\hat{\beta})$ assigned? This is the first time in this paper that we have met one of the infamous 'subjective priors', which are supposed to plague Bayesian methods. Here are some answers to this question. (a) Any other method of assigning a preference to alternatives must implicitly assign such priors. Bayesians adopt the healthy attitude of not sweeping them under the carpet. (b) With some thought, reasonable values can usually be assigned to subjective priors, and the degree of reasonable subjectivity in these assignments can be quantified. For example, a reasonable prior on an unknown standard deviation states that $\sigma$ is unknown over a range of $(3 \pm 2)$ orders of magnitude. This prior contributes a subjectivity of about $\pm$ a factor of 2 to the value of the evidence. This degree of subjectivity is often negligible compared to the evidence differences. (c) In the noisy interpolation example, all models considered include the free parameters $\alpha$ and $\beta$. So in this paper I do not need to assign a value to $P(\hat{\alpha},\hat{\beta})$; I assume that it is a flat prior which cancels out when we compare alternative interpolation models.

## 6 A Demonstration

These demonstrations will use two one–dimensional data sets, in imitation of [17]. The first data set, 'X,' has discontinuities in derivative (figure 4), and the second is a smoother data set, 'Y' (figure 8). In all the demonstrations, $\beta$ was not left as a free parameter, but was fixed to its known true value.

The Bayesian method of setting $\alpha$, assuming a single model is correct, has already been demonstrated, and quantified error bars have been placed on the most probable interpolant (figure 4). The method of evaluating the error bars is to use the posterior covariance matrix of the parameters $w_h$, $\mathbf{A}^{-1}$, to get the variance on $y(x)$, which for any $x$ is a linear function of the parameters, $y(x) = \sum_h \phi_h(x) w_h$. The error bars at a single point $x$ are given by $\mathrm{var}\, y(x) = \phi^{\mathrm{T}} \mathbf{A}^{-1} \phi$. However we have access to the full covariance information for the entire interpolant, not just the pointwise error bars. It is possible to visualise the joint error bars on the interpolant by making typical samples from the posterior distribution, performing a random walk around the posterior 'bubble' in parameter space [18]. It is simple to program such a random walk for interpolation problems such as are examined in this paper; however for lack of dynamic paper I will have to leave a demonstration of this to your imagination.

In this section objective comparison of alternative models will be demonstrated; this will be illustrated first with models differing only in the number of free parameters (for example polynomials of different degrees), then with comparisons between models as disparate as splines, radial basis functions and feedforward neural networks. For each individual model, the value of $\alpha$ is optimised, and the evidence is evaluated by integrating over $\alpha$ using the gaussian approximation. All logarithms are to base $e$.

#### Hermite polynomials and Hermite functions: Occam's razor for the number of basis functions

Figure 7a shows the evidence for Hermite polynomials of different degrees for data set X. This figure also shows the evidence for Hermite functions, by which I mean Hermite polynomials multiplied by $e^{-x^2/2}$. A regulariser of the form $E_W = \sum \frac{1}{2} 2^{2h} w_h^2$ was used in both cases because the leading polynomial coefficient of Hermite polynomials is $2^h$, and it was found that the evidence for the flat regulariser $\sum \frac{1}{2} w_h^2$ was much smaller.

Notice that an evidence maximum is obtained: beyond a certain number of terms, the evidence starts to decrease. This is the Bayesian Occam's razor at work. The additional terms make the model more powerful, able to make more predictions. This power is automatically penalised. Notice the characteristic shape of the 'Occam hill.' On the left, the hill is very steep as the over–simple models fail to fit the data; the penalty for misfitting the data scales as $N$, the number of data measurements. The other side of the hill is much less steep; the 'Occam factors' here only scale as $k \log N$, where $k$ is the number of parameters. We note in table 1 the values of the maximum evidence achieved by these two models, and move on to alternative models.

[7] Condition for this approximation: in the spectrum of eigenvalues of $\beta\mathbf{B}$, the number of eigenvalues within $e$–fold of $\alpha$ must be $O(1)$.

[8] There are analytic methods for performing such integrals over $\beta$ [2].

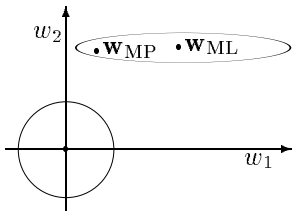


Figure 6: **Good and bad parameter measurements** $w_1$ and $w_2$ are the components in parameter space in two directions parallel to eigenvectors of the data matrix $\mathbf{B}$. The circle represents the characteristic prior distribution for $\mathbf{w}$. The ellipse represents a characteristic contour of the likelihood, centred on the maximum likelihood solution $\mathbf{w}_{\rm ML}$. $\mathbf{w}_{\rm MP}$ represents the most probable parameter vector. $w_1$ is a direction in which $\lambda_1$ is small compared to $\alpha$, *i.e.* the data have no strong preference about the value of $w_1$; $w_1$ is a poorly measured parameter, and the term $\frac{\lambda_1}{\lambda_1+\alpha}$ is close to zero. $w_2$ is a direction in which $\lambda_1$ is large; $w_2$ is well determined by the data, and the term $\frac{\lambda_2}{\lambda_2+\alpha}$ is close to one.

Each eigenvalue $\lambda_a$ measures how strongly one parameter is determined by the data. $\alpha$ measures how strongly the parameters are determined by the prior. The term $\lambda_a/(\lambda_a+\alpha)$ is a number between 0 and 1 which measures the strength of the data relative to the prior (figure 6). A direction in parameter space for which $\lambda_a$ is small compared to $\alpha$ does not contribute to the number of good parameter measurements. $\gamma$ is thus a measure of the effective number of parameters which are well determined by the data. As $\alpha/\beta \to 0$, $\gamma$ increases from 0 to $k$.

This concept is not only important for locating the optimum value of $\alpha$: it is only the $\gamma$ good parameter measurements which are expected to contribute to the reduction of the data misfit that occurs when a model is fitted to noisy data. In the process of fitting $\mathbf{w}$ to the data, it is unavoidable that some fitting of the model to noise will occur, because some components of the noise are indistinguishable from real data. Typically, one unit ($\chi^2$) of noise will be fitted for every well–determined parameter. Poorly determined parameters are determined by the regulariser only, so they do not reduce $\chi^2_D$ in this way. We will now examine how this concept enters into the Bayesian choice of $\beta$.

Recall that the expectation of the $\chi^2$ misfit between the true interpolant and the data is $N$. We do not know the true interpolant, and the only misfit measure we have access to is the $\chi^2$ of the data, $\chi^2_D = 2\beta E_D$. The 'discrepancy principle' of orthodox statistics states that the model parameters should be adjusted so as to make $\chi^2_D = N$. Other orthodox approaches would suggest that we should estimate the noise level so as to set $\chi^2_D = N - k$, where $k$ is the number of free parameters. Let us find out the opinion of Bayes' rule on this matter.

We differentiate the log evidence (21) with respect to $\beta$ and obtain:

$$2\beta E_D = N - \gamma \tag{24}$$

Thus the most probable noise estimate, $\hat{\beta}$, does not satisfy $\chi^2_D = N$ or $\chi^2_D = N - k$; rather, $\chi^2_D = N - \gamma$. Thus the Bayesian estimate of noise level naturally takes into account the fact that the parameters which have been determined by the data inevitably suppress some of the noise in the data, while the poorly determined parameters do not. Note that the value of $\chi^2_D$ only enters into the determination of $\beta$: misfit criteria have no role in the Bayesian choice of $\alpha$ [6].

In summary, at the optimum value of $\alpha$ and $\beta$, $\chi^2_W = \gamma$, $\chi^2_D = N - \gamma$. Notice that this implies that the total misfit $M = \alpha E_W + \beta E_D$ satisfies the simple equation $2M = N$.

The Bayesian choice of $\alpha$ is illustrated by figure 4b. Figure 5b illustrates the functions involved with the Bayesian choice of $\alpha$, and compares them with the 'test error' approach. Demonstration of the Bayesian choice of $\beta$ is omitted, since it is straightforward; $\beta$ is fixed to its true value for the demonstrations in this paper.

These results generalise to the case where there are two or more separate regularisers with independent regularising constants $\alpha_1, \alpha_2 \ldots$ [6]. In this case, each regulariser has a number of well–measured parameters $\gamma_i$ associated with it. Multiple regularisers will be used in the companion paper on neural networks.

Finding the evidence maximum with a head on approach would involve evaluating $\det \mathbf{A}$ while searching over $\alpha, \beta$; the above results (22,24) enable us to speed up this search (for example by the use of re–estimation formulae like $\alpha := \gamma/2E_W$) and replace the evaluation of $\det \mathbf{A}$ by the evaluation of $\mathrm{Trace}\mathbf{A}^{-1}$. For large dimensional problems where this task is demanding, Skilling has developed methods for estimating $\mathrm{Trace}\mathbf{A}^{-1}$ statistically [21].

# 5 Bayesian model comparison

To rank alternative basis sets $\mathcal{A}$ and regularisers (priors) $\mathcal{R}$ in the light of the data, we examine the posterior probabilities:

$$P(\mathcal{A},\mathcal{R}|D) \propto P(D|\mathcal{A},\mathcal{R})P(\mathcal{A},\mathcal{R}) \tag{25}$$

The data–dependent term, the evidence for $\mathcal{A},\mathcal{R}$, appeared earlier as the normalising constant in (15), and it is evaluated by integrating the evidence for $(\alpha,\beta)$:

$$P(D|\mathcal{A},\mathcal{R}) = \int P(D|\mathcal{A},\mathcal{R},\alpha,\beta)P(\alpha,\beta)\,d\alpha\,d\beta \tag{26}$$

Assuming that we have no reason to assign strongly differing priors $P(\mathcal{A},\mathcal{R})$, alternative $\mathcal{A},\mathcal{R}$ are ranked just by examining the evidence. The evidence can also be compared with the evidence found by an equivalent Bayesian analysis of other learning and interpolation models so as to allow the data to assign a preference to the alternative

space to the prior accessible volume. Figure 5a illustrates the behaviour of these various terms as a function of $\alpha$ for the same radial basis function model as illustrated in figure 4.

Now we could just proceed to evaluate the evidence numerically as a function of $\alpha$ and $\beta$, but a more deep and fruitful understanding of this problem is possible.

## Properties of the evidence maximum

The maximum over $\alpha, \beta$ of $P(D|\alpha,\beta,\mathcal{A},\mathcal{R}) = \frac{Z_M(\alpha,\beta)}{Z_W(\alpha)Z_D(\beta)}$ has some remarkable properties which give deeper insight into this Bayesian approach. The results of this section are useful both numerically and intuitively.

Following Gull [6], we transform to the basis in which the Hessian of $E_W$ is the identity, $\nabla\nabla E_W = \mathbf{I}$. This transformation is simple in the case of quadratic $E_W$: rotate into the eigenvector basis of $\mathbf{C}$ and stretch the axes so that the quadratic form $E_W$ becomes homogeneous. This is the natural basis for the prior. I will continue to refer to the parameter vector in this basis as $\mathbf{w}$. Using $\nabla\nabla M = \mathbf{A}$ and $\nabla\nabla E_D = \mathbf{B}$ as above, we differentiate the log evidence with respect to $\alpha$ and $\beta$ so as to find the condition that is satisfied at the maximum. The log evidence, from (20), is:

$$\log P(D|\alpha,\beta,\mathcal{A},\mathcal{R}) = -\alpha E_W^{MP} - \beta E_D^{MP} - \frac{1}{2}\log\det\mathbf{A} + \frac{N}{2}\log\beta + \frac{k}{2}\log\alpha + \frac{k}{2}\log 2\pi \quad (21)$$

First, differentiating with respect to $\alpha$, we need to evaluate $\frac{d}{d\alpha}\log\det\mathbf{A}$. Using $\mathbf{A} = \alpha\mathbf{I} + \beta\mathbf{B}$,

$$\begin{aligned}\frac{d}{d\alpha}\log\det\mathbf{A} &= \mathrm{Trace}\left(\mathbf{A}^{-1}\frac{d\mathbf{A}}{d\alpha}\right)\\ &= \mathrm{Trace}\mathbf{A}^{-1}\mathbf{I} = \mathrm{Trace}\mathbf{A}^{-1}\end{aligned}$$

This result is exact if $E_W$ and $E_D$ are quadratic. Otherwise this result is an approximation, omitting terms in $\partial\mathbf{B}/\partial\alpha$. Now, differentiating (21) and setting the derivative to zero, we obtain the following condition for the most probable value of $\alpha$:

$$2\alpha E_W^{MP} = k - \alpha\mathrm{Trace}\mathbf{A}^{-1} \quad (22)$$

The quantity on the left is the dimensionless measure of the amount of structure introduced into the parameters by the data, *i.e.* how much the fitted parameters differ from their null value. It can be interpreted as the $\chi^2$ of the parameters, since it is equal to $\chi^2_W = \sum w_i^2/\sigma_W^2$, with $\alpha = 1/\sigma_W^2$.

The quantity on the right of (22) is called the number of good parameter measurements, $\gamma$, and has value between 0 and $k$. It can be written in terms of the eigenvalues of $\beta\mathbf{B}$, $\lambda_a$. The eigenvalues of $\mathbf{A}$ are $\lambda_a + \alpha$, so we have:

$$\gamma = k - \alpha\,\mathrm{Trace}\mathbf{A}^{-1} = k - \sum_{a=1}^{k}\frac{\alpha}{\lambda_a+\alpha} = \sum_{a=1}^{k}\frac{\lambda_a}{\lambda_a+\alpha} \quad (23)$$

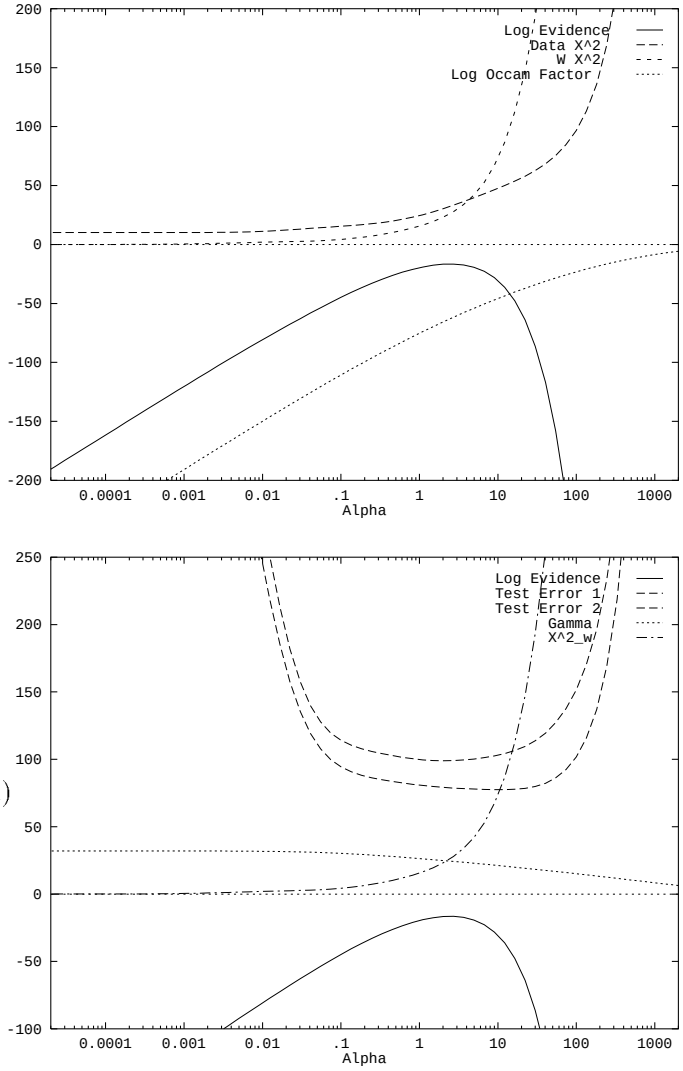


Figure 5: **Choosing $\alpha$**

a) **The evidence as a function of $\alpha$:** Using the same radial basis function model as in figure 4, this graph shows the log evidence as a function of $\alpha$, and shows the functions which make up the log evidence, namely the data misfit $\chi^2_D = 2\beta E_D$, the weight penalty term $\chi^2_W = 2\alpha E_W$, and the log of the Occam factor $(2\pi)^{k/2}\det{}^{-\frac{1}{2}}\mathbf{A}/Z_W(\alpha)$.

b) **Criteria for optimising $\alpha$:** This graph shows the log evidence as a function of $\alpha$, and the functions whose intersection locates the evidence maximum: the number of good parameter measurements $\gamma$, and $\chi^2_W$. Also shown is the test error (rescaled) on two test sets; finding the test error minimum is an alternative criterion for setting $\alpha$. Both test sets were more than twice as large in size as the interpolated data set. Note how the point at which $\chi^2_W = \gamma$ is clear and unambiguous, which cannot be said for the minima of the test energies. The evidence gives $\alpha$ a confidence interval of $[1.3, 5.0]$. The test error minima are more widely distributed because of finite sample noise.

arises for the optimal value of $\beta$; on the other hand, the Bayesian choice of $\alpha$ is unrelated to the misfit.

## Bayesian choice of $\alpha$ and $\beta$

To infer from the data what value $\alpha$ and $\beta$ should have,[6] Bayesians evaluate the posterior probability distribution:

$$P(\alpha,\beta|D,\mathcal{A},\mathcal{R}) = \frac{P(D|\alpha,\beta,\mathcal{A},\mathcal{R})P(\alpha,\beta)}{P(D|\mathcal{A},\mathcal{R})} \qquad (15)$$

The data dependent term $P(D|\alpha,\beta,\mathcal{A},\mathcal{R})$ has already appeared earlier as the normalising constant in equation (12), and it is called the evidence for $\alpha$ and $\beta$. Similarly the normalising constant of (15) is called the evidence for $\mathcal{A},\mathcal{R}$, and it will turn up later when we compare alternative models $\mathcal{A},\mathcal{R}$ in the light of the data.

If $P(\alpha,\beta)$ is a flat prior, the evidence is the function that we use to assign a preference to alternative values of $\alpha$ and $\beta$. It is given by

$$P(D|\alpha,\beta,\mathcal{A},\mathcal{R}) = \frac{Z_M(\alpha,\beta)}{Z_W(\alpha)Z_D(\beta)} \qquad (16)$$

where the $Z$'s have been defined earlier. Occam's razor is implicit in this formula: if $\alpha$ is small, the large freedom in the range of possible values of $\mathbf{w}$ is automatically penalised by the consequent large value of $Z_W$; models that fit the data well achieve a large value of $Z_M$; a model that has to be very finely tuned for it to fit the data is penalised by a smaller value of $Z_M$. The optimum value of $\alpha$ achieves a compromise between fitting the data well and being too powerful a model.

Now to assign a preference to $(\alpha,\beta)$, our computational task is to evaluate the three integrals $Z_M$, $Z_W$ and $Z_D$. We will come back to this task in a moment.

### But that sounds like determining your prior after the data have arrived!

This is an aside which can be omitted on a first reading. When I first heard the preceding explanation of Bayesian regularisation I was discontent because it seemed that the prior is being chosen from an ensemble of possible priors *after* the data have arrived. To be precise, as described above, the most probable prior (most probable value of $\alpha$) is selected; then that prior alone is used to infer what the interpolant might be. This is not how Bayes would have us infer what the interpolant is. It is the combined ensemble of priors that define our prior, and we should integrate over this ensemble when we do inference. Let us work out what happens if we follow this proper approach. The preceding method of using only the most probable prior will emerge as a good approximation.

Let us examine the true posterior $P(\mathbf{w}|D,\mathcal{A},\mathcal{R})$, obtained by integrating over $\alpha$ and $\beta$:

$$P(\mathbf{w}|D,\mathcal{A},\mathcal{R}) = \int P(\mathbf{w}|D,\alpha,\beta,\mathcal{A},\mathcal{R})P(\alpha,\beta|D,\mathcal{A},\mathcal{R})\,d\alpha\,d\beta \qquad (17)$$

In words, the posterior probability over $\mathbf{w}$ can be written as a linear combination of the posteriors for all values of $\alpha,\beta$. Each posterior density is weighted by the probability of $\alpha,\beta$ given the data, which appeared in (15). This means that if $P(\alpha,\beta|D,\mathcal{A},\mathcal{R})$ has a single peak at $\hat{\alpha},\hat{\beta}$, then the true posterior $P(\mathbf{w}|D,\mathcal{A},\mathcal{R})$ will be dominated by the density $P(\mathbf{w}|D,\hat{\alpha},\hat{\beta},\mathcal{A},\mathcal{R})$. As long as the properties of the posterior $P(\mathbf{w}|D,\alpha,\beta,\mathcal{A},\mathcal{R})$ do not change rapidly with $\alpha,\beta$ near $\hat{\alpha},\hat{\beta}$ and the peak in $P(\alpha,\beta|D,\mathcal{A},\mathcal{R})$ is strong, we are justified in using the approximation:

$$P(\mathbf{w}|D,\mathcal{A},\mathcal{R}) \simeq P(\mathbf{w}|D,\hat{\alpha},\hat{\beta},\mathcal{A},\mathcal{R}) \qquad (18)$$

## Evaluating the evidence

Let us return to our train of thought at equation (16). To evaluate the evidence for $\alpha,\beta$, we want to find the integrals $Z_M$, $Z_W$ and $Z_D$. Typically the most difficult integral to evaluate is $Z_M$.

$$Z_M(\alpha,\beta) = \int d^k\mathbf{w}\ \exp -M(\mathbf{w},\alpha,\beta)$$

If the regulariser $\mathcal{R}$ is a quadratic functional (and the favourites are), then $E_D$ and $E_W$ are quadratic functions of $\mathbf{w}$, and we can evaluate $Z_M$ exactly. Letting $\nabla\nabla E_W = \mathbf{C}$ and $\nabla\nabla E_D = \mathbf{B}$ then using $\mathbf{A} = \alpha\mathbf{C} + \beta\mathbf{B}$, we have:

$$M = M(\mathbf{w}_{\rm MP}) + \frac{1}{2}(\mathbf{w}-\mathbf{w}_{\rm MP})^{\rm T}\mathbf{A}(\mathbf{w}-\mathbf{w}_{\rm MP})$$

where $\mathbf{w}_{\rm MP} = \mathbf{A}^{-1}\mathbf{B}\mathbf{w}_{\rm ML}$. This means that $Z_M$ is the gaussian integral:

$$Z_M = e^{-M_{\rm MP}}(2\pi)^{k/2}\det{}^{-\frac{1}{2}}\mathbf{A} \qquad (19)$$

In many cases where the regulariser is not quadratic (for example, entropy–based), this gaussian approximation is still servicable [6]. Thus we can write the log evidence for $\alpha$ and $\beta$ as:

$$\log P(D|\alpha,\beta,\mathcal{A},\mathcal{R}) = -\alpha E_W^{MP} - \beta E_D^{MP} - \frac{1}{2}\log\det\mathbf{A} - \log Z_W(\alpha) - \log Z_D(\beta) + \frac{k}{2}\log 2\pi \qquad (20)$$

The term $\beta E_D^{MP}$ represents the misfit of the interpolant to the data. The three terms $-\alpha E_W^{MP} - \frac{1}{2}\log\det\mathbf{A} - \log Z_W(\alpha)$ constitute the 'Occam factor' penalising over–powerful values of $\alpha$: the ratio $(2\pi)^{k/2}\det{}^{-\frac{1}{2}}\mathbf{A}/Z_W(\alpha)$ is the ratio of the posterior accessible volume in parameter

[6] Note that it is not satisfactory to simply maximise the likelihood over $\mathbf{w}$, $\alpha$ and $\beta$; the likelihood has a skew peak such that the maximum likelihood value for the parameters is not in the same place as most of the posterior probability. To get a feeling for this here is a more familiar problem: examine the posterior probability for the parameters of a gaussian $(\mu,\sigma)$ given $N$ samples: the maximum likelihood value for $\sigma$ is $\sigma_N$, but the most probable value for $\sigma$ is $\sigma_{N-1}$.

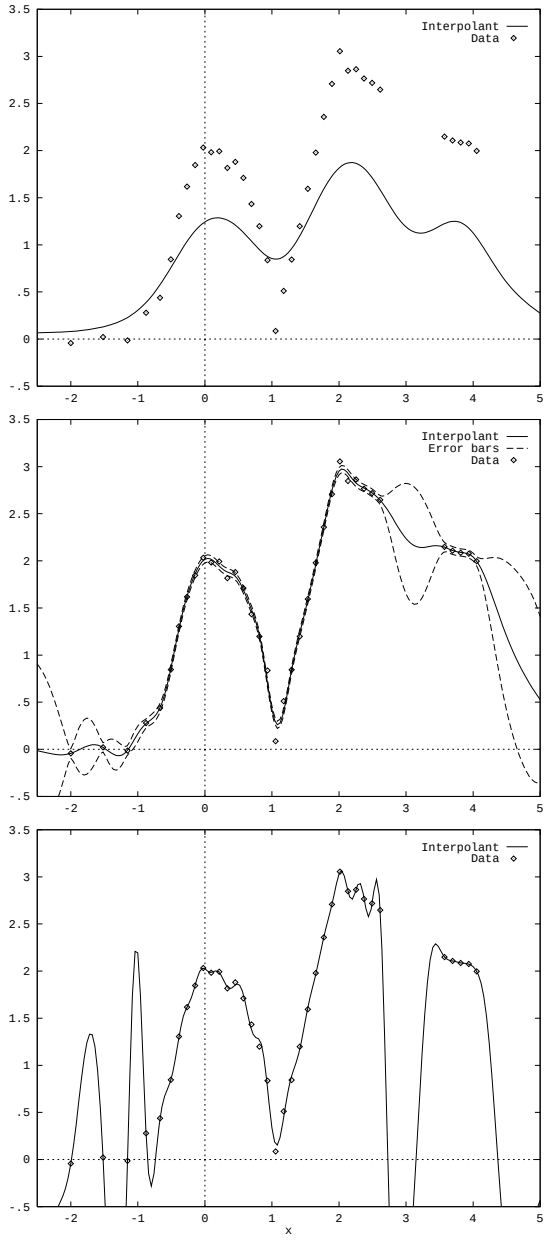


Figure 4: **How the best interpolant depends on $\alpha$**
These figures introduce a data set, 'X,' which is interpolated with a variety of models in this paper. Notice that the density of data points is not uniform on the $x$–axis. In the three figures the data set is interpolated using a radial basis function model with a basis of 60 equally spaced Cauchy functions, all with radius 0.2975. The regulariser is $E_W = \frac{1}{2}\sum w^2$, where $w$ are the coefficients of the basis functions. Each figure shows the most probable interpolant for a different value of $\alpha$: a) 6000; b) 2.5; c) $10^{-7}$. Note at the extreme values how the data is oversmoothed and overfitted respectively. $\alpha = 2.5$ is the *most probable* value of $\alpha$. In b), the most probable interpolant is displayed with its $1\sigma$ error bars, which represent how uncertain we are about the interpolant at each point, under the assumption that the interpolation model and the value of $\alpha$ are correct. Notice how the error bars increase in magnitude where the data is sparse. The error bars do not include the datapoint close to (1,0), because the radial basis function model does not expect sharp discontinuities, so that point is interpreted as an improbable outlier.

probable) interpolant vary. Assume that we are using a prior like the splines prior defined earlier, and imagine that we interpolate at a very large value of $\alpha$; then this will constrain the interpolant to be very smooth and flat, and it will not fit the data at all well (figure 4a). As $\alpha$ is decreased, the interpolant starts to fit the data better (figure 4b). If $\alpha$ is made even smaller, the interpolant oscillates wildly so as to overfit the noise in the data (figure 4c). The choice of the 'best' value of $\alpha$ is our first 'Occam's razor' problem: large values of $\alpha$ correspond to simple hypotheses which make constrained and precise predictions, saying 'the interpolant is expected to not have extreme curvature anywhere;' a tiny value of $\alpha$ corresponds to the more powerful and flexible hypothesis that says 'the interpolant could be anything at all, our prior belief in smoothness is very weak.' The task is to find a value of $\alpha$ which is small enough that the data are fitted but not so small that they are overfitted. For more severely ill–posed problems such as deconvolution, the precise value of the regularising parameter is increasingly important. Orthodox statistics has ways of assigning values to such parameters, based for example on misfit criteria and cross–validation. Gull has demonstrated why the popular use of misfit criteria is incorrect and how Bayes sets these parameters [6]. Cross–validation may be an unreliable technique unless large quantities of data are available. See section 6 and [13] however for further discussion of cross–validation. I will explain the Bayesian method of setting $\alpha$ and $\beta$ after first reviewing some statistics of misfit.

## Misfit, $\chi^2$, and the effect of parameter measurements

For $N$ gaussian variables with mean $\mu$ and standard deviation $\sigma$, the statistic $\chi^2 = \sum(x-\mu)^2/\sigma^2$ is a measure of misfit. If $\mu$ is known *a priori*, $\chi^2$ has expectation $N \pm \sqrt{N}$. However, if $\mu$ is fitted from the data by setting $\mu = \bar{x}$, we 'use up a degree of freedom', and $\chi^2$ has expectation $N-1$. In the second case $\mu$ is a 'well–measured parameter' (measured by the data). When a parameter is determined from the data in this way it is unavoidable that the parameter fits some of the noise as well. That is why the expectation of $\chi^2$ is reduced by one. This is the basis of the distinction between the $\sigma_N$ and $\sigma_{N-1}$ buttons on your calculator. It is common for this distinction to be ignored, but in cases such as interpolation where the number of free parameters is similar to the number of data points, it is essential to find and make the analogous distinction. It will be demonstrated that the Bayesian choice of both $\alpha$ and $\beta$ are most simply expressed in terms of the effective number of well–measured parameters, $\gamma$, to be derived below.

Misfit criteria are 'principles' which set parameters like $\alpha$ and $\beta$ by requiring that $\chi^2$ should have a particular value. The discrepancy principle requires $\chi^2 = N$. Another principle requires $\chi^2 = N - k$, where $k$ is the number of free parameters. We will find that an intuitive misfit criterion