Prior Probability

THE AUSTRALIAN NATIONAL UNIVERSITY

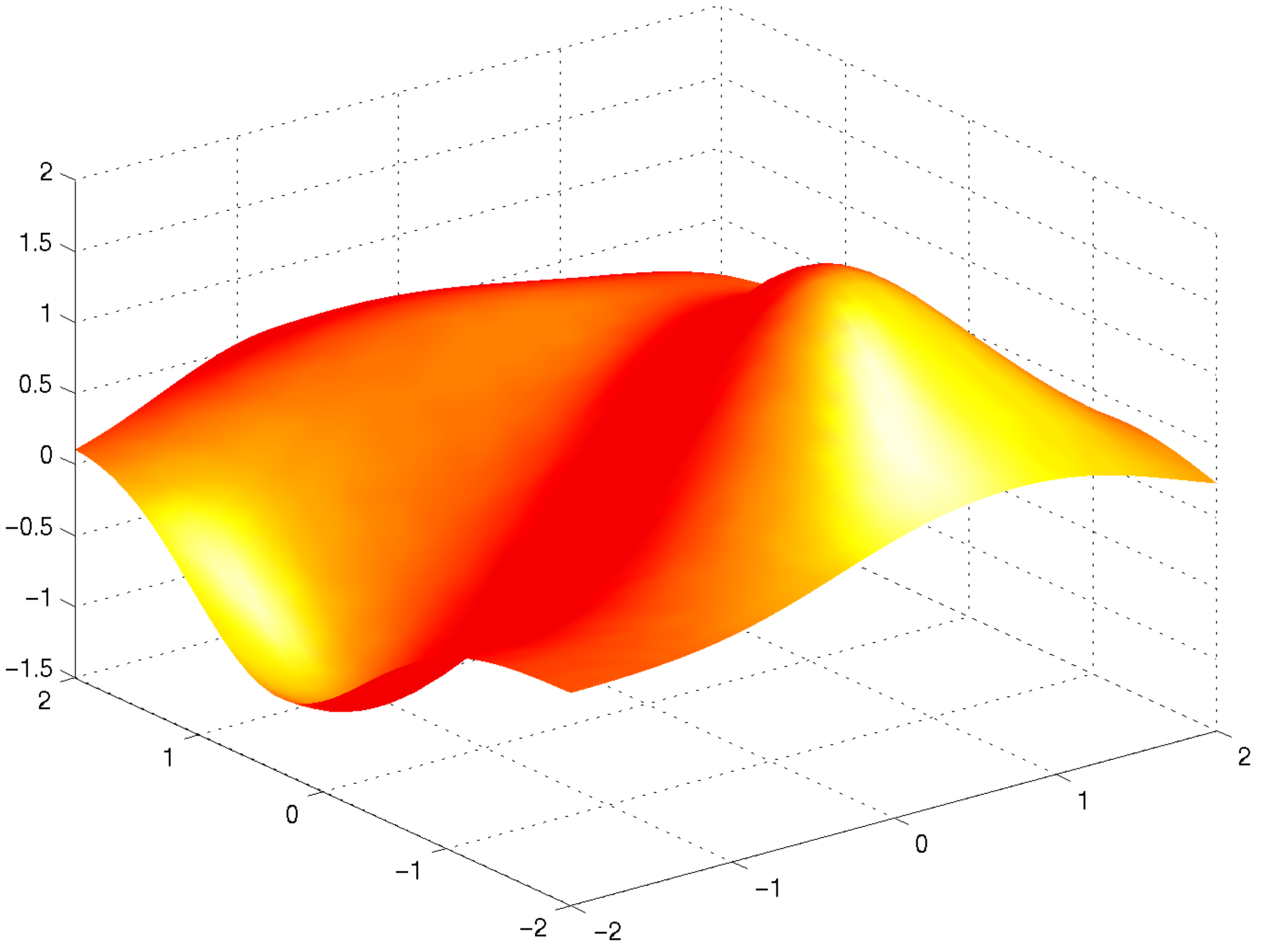
Idea 1

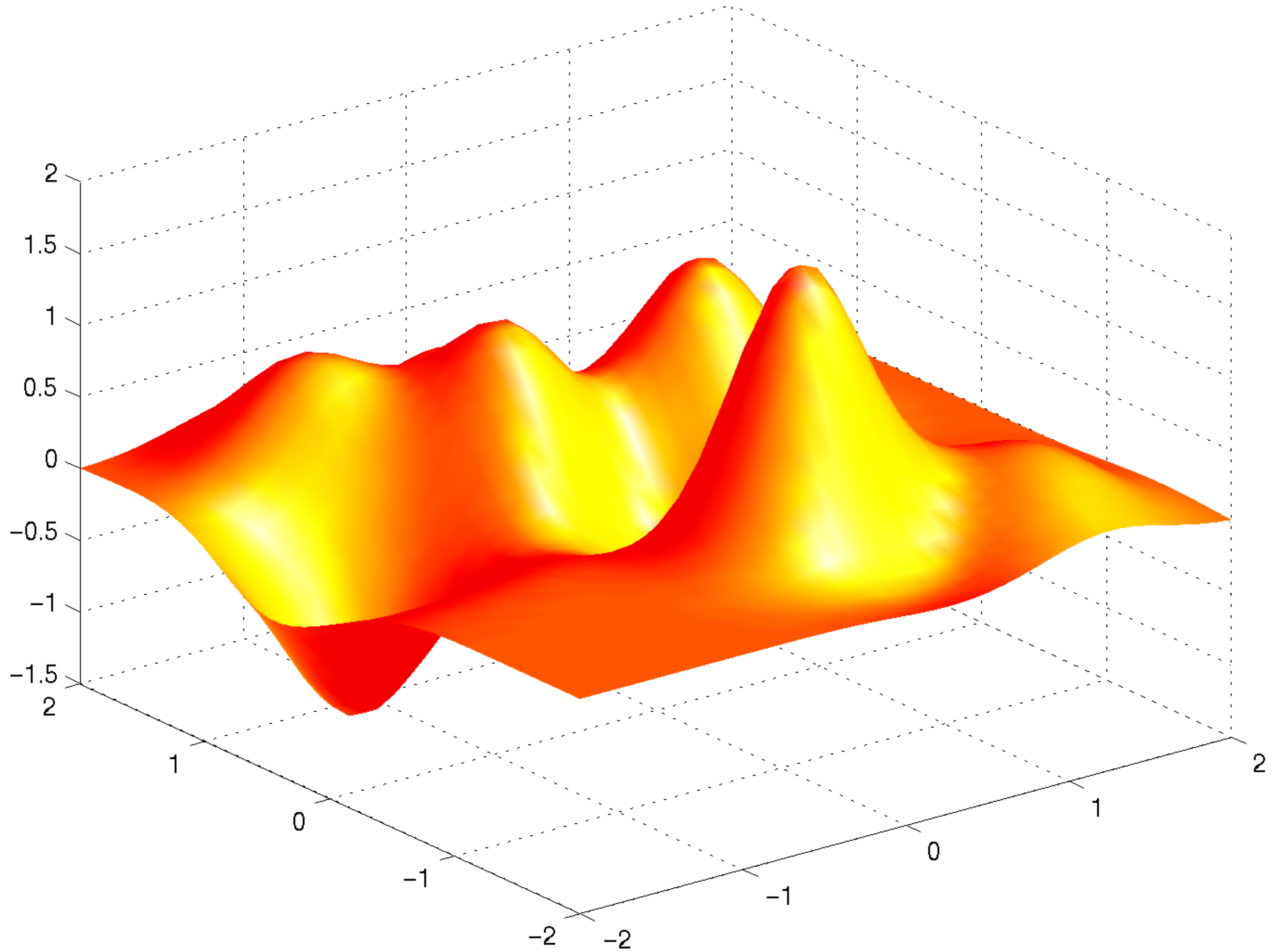
Quite often we have a rough idea of what function we can expect beforehand.

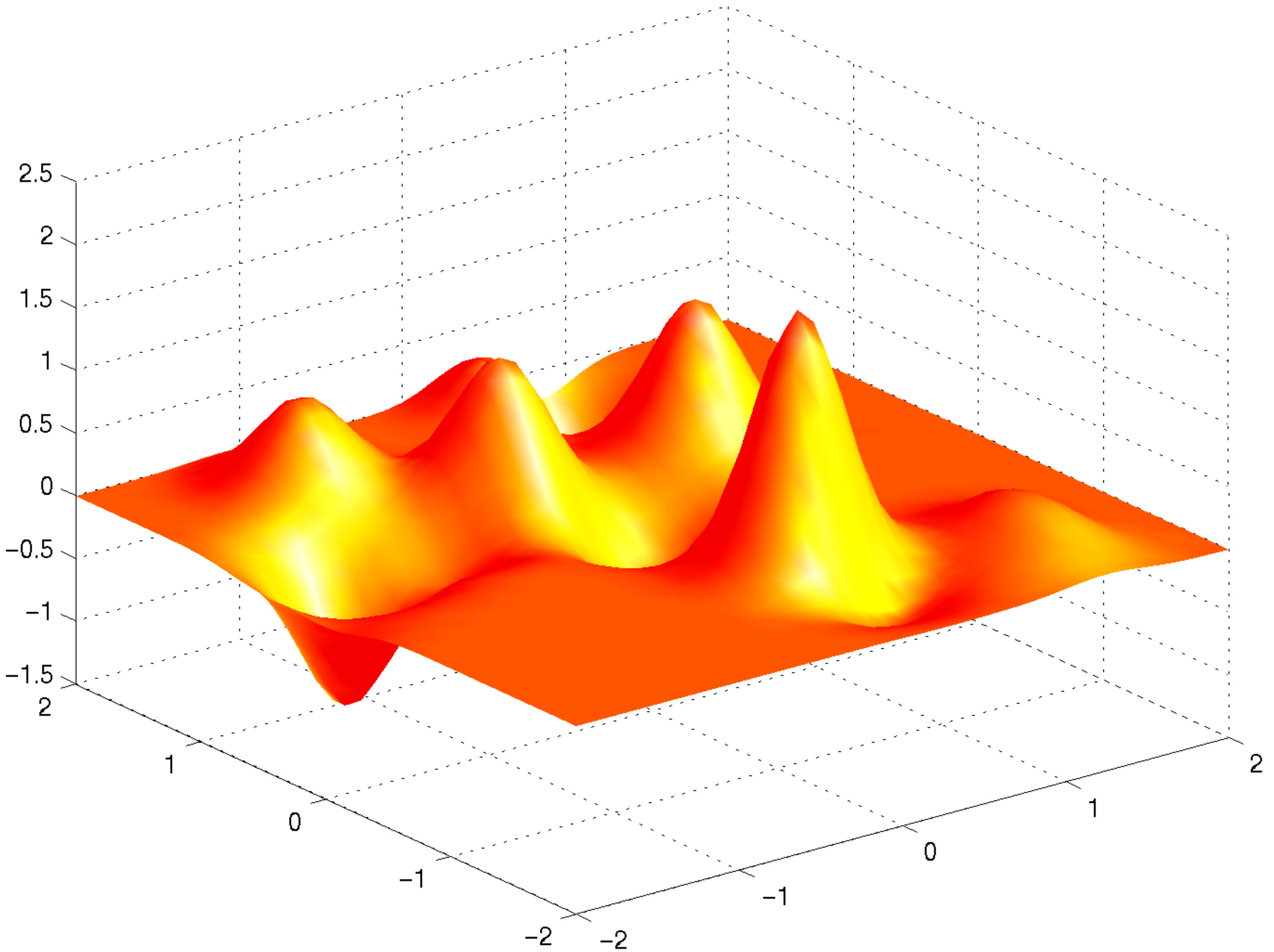
- We observe similar functions in practice.
- We **think** that e.g. smooth functions should be more likely.
- We **would like** a certain type of functions.
- We have **prior knowledge** about specific properties, e.g. vanishing second derivative, etc.

Idea 2

We have to specify somehow, how likely it is to observe a specific function f from an overall class of functions. This is done by assuming some density $p(f)$ describing how likely we are to observe f.


2
1.5
1
0.5
0
−0.5
−1
−1.5
2
1
0
−1
−2
−2
−1
0
1
2


2
1.5
1
0.5
0
−0.5
−1
−1.5
2
1
0
−1
−2
−2
−1
0
1
2


2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
2
1
0
−1
−2
−2
−1
0
1
2

Examples

THE AUSTRALIAN NATIONAL UNIVERSITY

Speech Signal

We know that the signal is bandlimited, hence any signal containing frequency components above 10kHz has density 0.

Parametric Prior

We may know that f is a linear combination of $\sin x, \cos x, \sin 2x$, and $\cos 2x$ and that the coefficients may be chosen from the interval $[-1, 1]$.

$$p(f) = \begin{cases} \frac{1}{16} & \text{if } f = \alpha_1 \sin x + \alpha_2 \cos x + \alpha_3 \sin 2x + \alpha_4 \cos 2x \text{ with } \alpha_i \in [-1, 1] \\ 0 & \text{otherwise} \end{cases}$$

Prior on Function Values

We assume that there is a correlation between the function values f_i at location $f(\mathbf{x}_i)$. There we have

$$p(f_1, f_2, f_3) = \frac{1}{\sqrt{(2\pi)^3 \det K}} \exp\left(-\frac{1}{2}(f_1, f_2, f_3)^\top K^{-1}(f_1, f_2, f_3)\right).$$

Examples

Prior on Function Values

The larger the off diagonal elements K_{ij} are, the more the corresponding function values $f(x_i)$ and $f(x_j)$ are correlated. The main diagonal elements K_{ii} provide the variance of f_i and the off diagonal elements the covariance between pairs f_i and f_j. This is not a prior assumption about the *function* f but only about *its values* $f(x_i)$ at some previously specified locations.

Nonparametric Priors

We may only have the abstract knowledge that smooth functions with small function values are more likely to occur. One possible way of quantifying such a relation is to posit that the prior probability of a function occurring depends only on its L_2 norm and the L_2 norm of its first derivative. This leads to expressions of the form

$$-\ln p(f) = c + \|f\|^2 + \|\partial_x f\|^2.$$

How to use Priors

THE AUSTRALIAN NATIONAL UNIVERSITY

Bayes Rule

We want to infer the probability of f, having observed X, Y. By Bayes' rule we obtain

$$p(f|X, Y) = \frac{p(Y|f, X)p(f|X)}{p(Y|X)}.$$

This is also often called the **posterior probability** of observing f, after that the data X, Y arrived.

Usual Assumption

Typically we assume that X has no influence as to which f we may assume, i.e. $p(f|X) = p(f)$ (X and f are independent random variables).

Likelihood

$p(Y|f, X)$ is the Likelihood term that we used in Maximum Likelihood estimation. All that is happening is a **reweighting** of the likelihood by the prior distribution.

Inference

Goal

We want to infer f, possibly its value at a new location $\mathbf{x}$ via $p(f|X,Y)$.

Trick

The quantity $p(Y|X)$ is usually quite hard to obtain, moreover it is independent of f, therefore we can just treat it as a normalizing factor and we obtain

$$p(f|X,Y) \propto p(Y|f,X)p(f)$$

The normalization constant can be taken care of later.

Prediction

If we want to compute the expected value of $f(\mathbf{x})$ at a new location all we have to do is compute

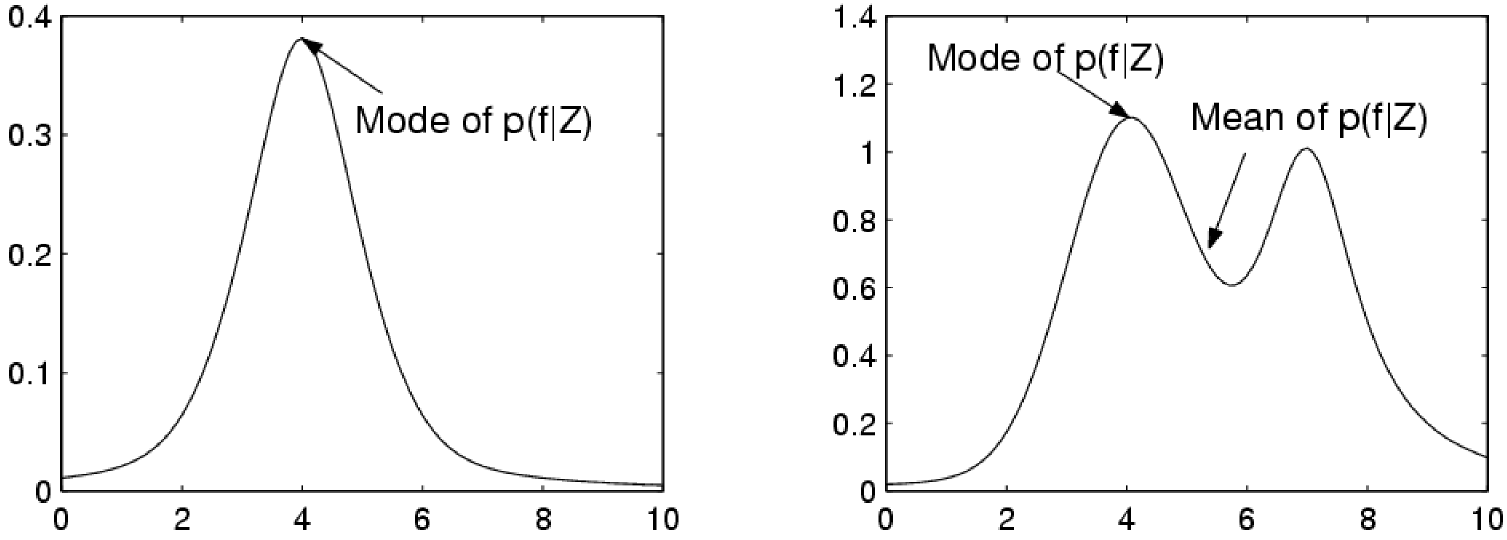
$$\mathbf{E}[f(\mathbf{x})] = \int f(\mathbf{x})p(f|X,Y)df$$

Variance

Likewise, to infer the predictive variance we compute

$$\mathbf{E}\left[(f(\mathbf{x}) - \mathbf{E}[f(\mathbf{x})])^2\right] = \int (f(\mathbf{x}) - \mathbf{E}[f(\mathbf{x})])^2 p(f|X,Y)df$$

This means that we can estimate the variation of $f(\mathbf{x})$, given the data and our prior knowledge about f, as encoded by $p(f)$.



Approximate Inference

Problem

Nobody wants to compute integrals ...

Idea

After all, we are only **averaging**, so replace the mean of the distribution by the mode and hope that it will be ok. This leads to the maximum a posteriori estimate (see next slide).

Lucky Coincidence

For Gaussian distributions (and many others) mode and mean coincide.

Problem 2

For some distributions it does not work well ...

Idea 2

Approximate the posterior $p(f|X, Y)$ by a **parametric** model. This is often referred to as **variational approximation**.

Maximum a Posteriori Estimate

THE AUSTRALIAN NATIONAL UNIVERSITY

Maximizing the Posterior Probability

To find the hypothesis f with the highest posterior probability we have to maximize

$$p(f|X, Y) = \frac{p(Y|f, X)p(f|X)}{p(Y|X)}$$

Lazy Trick

Since we only want f (and $p(Y|X)$ is independent of f), all we have to do is maximize $p(Y|f, X)p(f)$.

Taking Logs

For convenience we get f by minimizing

$$-\log p(Y|f, X)p(f|X) = -\log p(Y|f, X) - \log p(f) = -\log \mathfrak{L} - \log p(f)$$

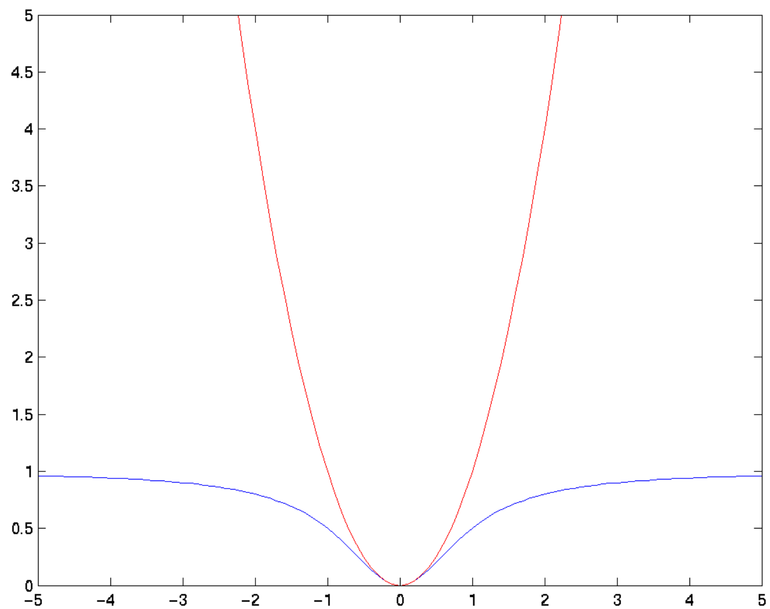
So all we are doing is to **reweight the likelihood** by $-\log p(f)$. This looks suspiciously like the regularization term. We will match up the two terms later.

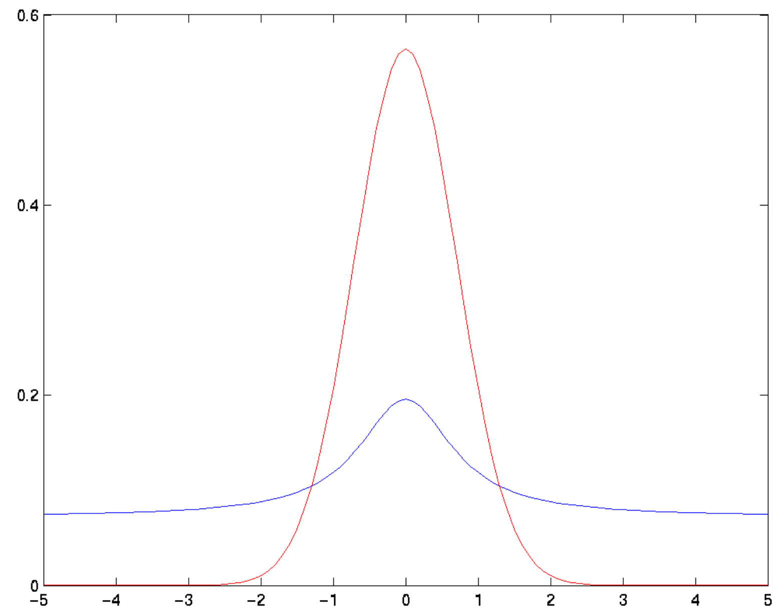
Maximum a Posteriori Estimate, Part II

Variance

Once we found the **mode** f_0 of the distribution, we might as well approximate the variance by approximating $p(f|X,Y)$ with a normal distribution around f_0.

This is done by computing the second order information at f_0, i.e. $\partial_f^2 - \log p(f|X,Y)$.





Connection to Regularized Risk

Recycling of the Likelihood

Match up terms as we did with the likelihood and the loss function. In particular, we recycle these terms:

$$\begin{aligned} c(\mathbf{x}, y, f(\mathbf{x})) &\equiv -\log p(y - f(\mathbf{x})) \\ p(y|f(\mathbf{x}) &\equiv \exp(-c(\mathbf{x}, y, f(\mathbf{x})) \end{aligned}$$

Now all we have to do is take care of $m\lambda\Omega[f]$ and $-\log p(f)$.

Regularizer and Prior

The correspondence

$$m\lambda\Omega[f] + c = -\log p(f) \text{ or equivalently } p(f) \propto \exp(-m\lambda\Omega[f])$$

is the link between regularizer and prior.

Caveat

The translation from regularizer into prior works only to some extent, since the integral over f need not converge.