Bayesian Kernel Methods

THE AUSTRALIAN NATIONAL UNIVERSITY

Unit 1: Bayes Rule, Approximate Inference, Hyperparameters

Unit 2: Gaussian Processes, Covariance Function, Kernel

Unit 3: GP: Regression

Unit 4: GP: Classification

Unit 5: Implementation: Laplace Approximation, Low Rank Methods

Unit 6: Implementation: Low Rank Methods, Bayes Committee Machine

Unit 7: Relevance Vector Machine: Priors on Coefficients

Unit 8: Relevance Vector Machine: Efficient Optimization and Extensions

http://mlg.anu.edu.au/∼smola/summer2002/

Overview of Unit 1: Bayesics

THE AUSTRALIAN NATIONAL UNIVERSITY

01: Parametric Density Models
02: Maximum Likelihood
03: Example: Mean and Variance
04: Bayes' Rule and Conditional Probabilities
05: Example: Unfair Jury
06: Priors
07: Example: Prior on Function Space
08: Bayesian Inference
09: Confidence Intervals
10: Problems with Exact Inference
11: Maximum a Posteriori Approximation
12: Laplace Approximation for Confidence Intervals
13: Relation to Regularized Risk Functional
14: Hyperparameters
15: MAP2 Approximation
16: To integrate or not to integrate

Parametric Density Models

Goal: We want to estimate the density of a random variable, say, $\mathbf{x}$, given a set of observations $X := \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$.

Problem: Without additional knowledge, this is very difficult (and we need lots of data).

"Solution:" Assume **a lot** about $p(\mathbf{x})$ and X.

Assumption 1: The set X has been obtained by drawing **independent identically distributed** samples from $p(\mathbf{x})$.

This assumption will hold throughout the lectures.

$$\text{It follows that } p(X) = p(\mathbf{x}_1, \ldots, \mathbf{x}_m) = \prod_{i=1}^{m} p(\mathbf{x}_i)$$

Assumption 2: The density $p(\mathbf{x})$ can be parameterized by θ, that is $p(\mathbf{x}) = p(\mathbf{x}|\theta)$.

Caution: We should write $p_\theta(\mathbf{x})$ to indicate that $p(\mathbf{x})$ is **parameterized** by θ, rather than the density of $\mathbf{x}$, given θ. But it will be useful later ...

Maximum Likelihood

THE AUSTRALIAN NATIONAL UNIVERSITY

Inference Principle: Find θ such that $p(X|\theta)$ is maximized. This means maximizing

$$p(X|\theta) = \prod_{i=1}^{m} p(\mathbf{x}_i|\theta) \text{ or equivalently } \log p(X|\theta) = \sum_{i=1}^{m} \log p(\mathbf{x}_i|\theta).$$

Likelihood: $p(X|\theta)$ as a **function of** θ is commonly referred to as the likelihood $\mathfrak{L}(\theta)$. Thereby we can find the parameter θ that is most plausible given X by maximizing $\mathfrak{L}(\theta)$.

Numerical Trick: Typically we minimize $-\log \mathfrak{L}$, that is, we minimize $\sum_{i=1}^{m} - \log p(\mathbf{x}_i|\theta)$

Note: Similarity to training error for regularized risk, here the error per observation corresponds to $-\log(x_i|\theta)$.

Problem 1: The maximum value of $\mathfrak{L}$ can be misleading, since $p(\mathbf{x}|\theta)$ may not be the right model (**approximation error**).

Problem 2: We may not have enough data to adjust θ properly, so the maximum value of $\mathfrak{L}$ may be misleadingly **high**.

Example: Mean and Variance

Normal Distribution: Estimate parameters $\theta := (\mu, \sigma^2)$ for a normal distribution

$$p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{\sigma^2}\right)$$

Negative Log-Likelihood:

$$-\log \mathfrak{L}(\mu, \theta) = -\frac{m}{2} \log 2\pi - m \log \sigma + \frac{1}{2\sigma^2} \sum_{i=1}^{m} (x_i - \mu)^2$$

Optimum for μ: (we assume $\sigma^2 \neq 0$)

$$\partial_\mu - \log \mathfrak{L}(\mu, \sigma^2) = \frac{1}{\sigma^2} \sum_{i=1}^{m} x_i - \mu = 0 \Longleftrightarrow \mu = \frac{1}{m} \sum_{i=1}^{m} x_i.$$

Optimum for σ^2: (we assume $\sigma^2 \neq 0$)

$$\partial_\sigma - \log \mathfrak{L}(\mu, \sigma^2) = -\frac{m}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{m} = 0 \Longleftrightarrow \sigma^2 = \frac{1}{m} \sum_{i=1}^{m} (x_i - \mu_i)^2.$$

Bayes' Rule and Conditional Probabilities

THE AUSTRALIAN NATIONAL UNIVERSITY

Joint Probabilitiy: $\Pr(X, Y)$ is the probability of the events X and Y occurring simultaneously.

Conditional Probabilitiy: $\Pr(X|Y)$ is the probability of the event X, given Y.

Bayes Rule: Joint and Conditional Probability are related by $\Pr(X, Y) = \Pr(X|Y)\Pr(Y)$. We may therefore expand $\Pr(X, Y)$ in X and Y to obtain

$$\Pr(X|Y) = \frac{\Pr(Y|X)\Pr(X)}{\Pr(Y)}$$

Joint Density: $\Pr(\mathbf{x}, \mathbf{y})$ is the density of the events $\mathbf{x}$ and $\mathbf{y}$ occurring simultaneously.

Conditional Density: $\Pr(\mathbf{x}|\mathbf{y})$ is the density of the event $\mathbf{x}$, given $\mathbf{y}$.

Bayes Rule: Joint and Conditional Density are related by

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x}|\mathbf{y})p(\mathbf{y}) \text{ and therefore } p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})}.$$

Examples

AIDS-Test:

We want to find out likely it is that a patient *really* has AIDS (denoted by X) if the test is positive (denoted by Y).

Roughly 0.1% of all Australians are infected ($\Pr(X) = 0.001$). The probability that an AIDS test tells us the wrong result is in the order of 1% ($\Pr(Y|\mathfrak{X}\backslash X) = 0.01$) and moreover we assume that it detects all infections ($\Pr(Y|X) = 1$). We have

$$\Pr(X|Y) = \frac{\Pr(Y|X)\Pr(X)}{\Pr(Y)} = \frac{\Pr(Y|X)\Pr(X)}{\Pr(Y|X)\Pr(X) + \Pr(Y|\mathfrak{X}\backslash X)\Pr(\mathfrak{X}\backslash X)}$$

Hence $\Pr(X|Y) = \frac{1 \cdot 0.001}{1 \cdot 0.001 + 0.01 \cdot 0.999} = 0.091$, i.e. the probability of AIDS is 9.1%!

Reliability of Eye-Witness:

Assume that an eye-witness is 90% sure and that there were 20 people at the crime scene, what is the probability that the guy identified committed the crime?

$$\Pr(X|Y) = \frac{0.9 \cdot 0.05}{0.9 \cdot 0.05 + 0.1 \cdot 0.95} = 0.3213 = 32\% \text{ that's a worry } \ldots$$

Priors

THE AUSTRALIAN NATIONAL UNIVERSITY

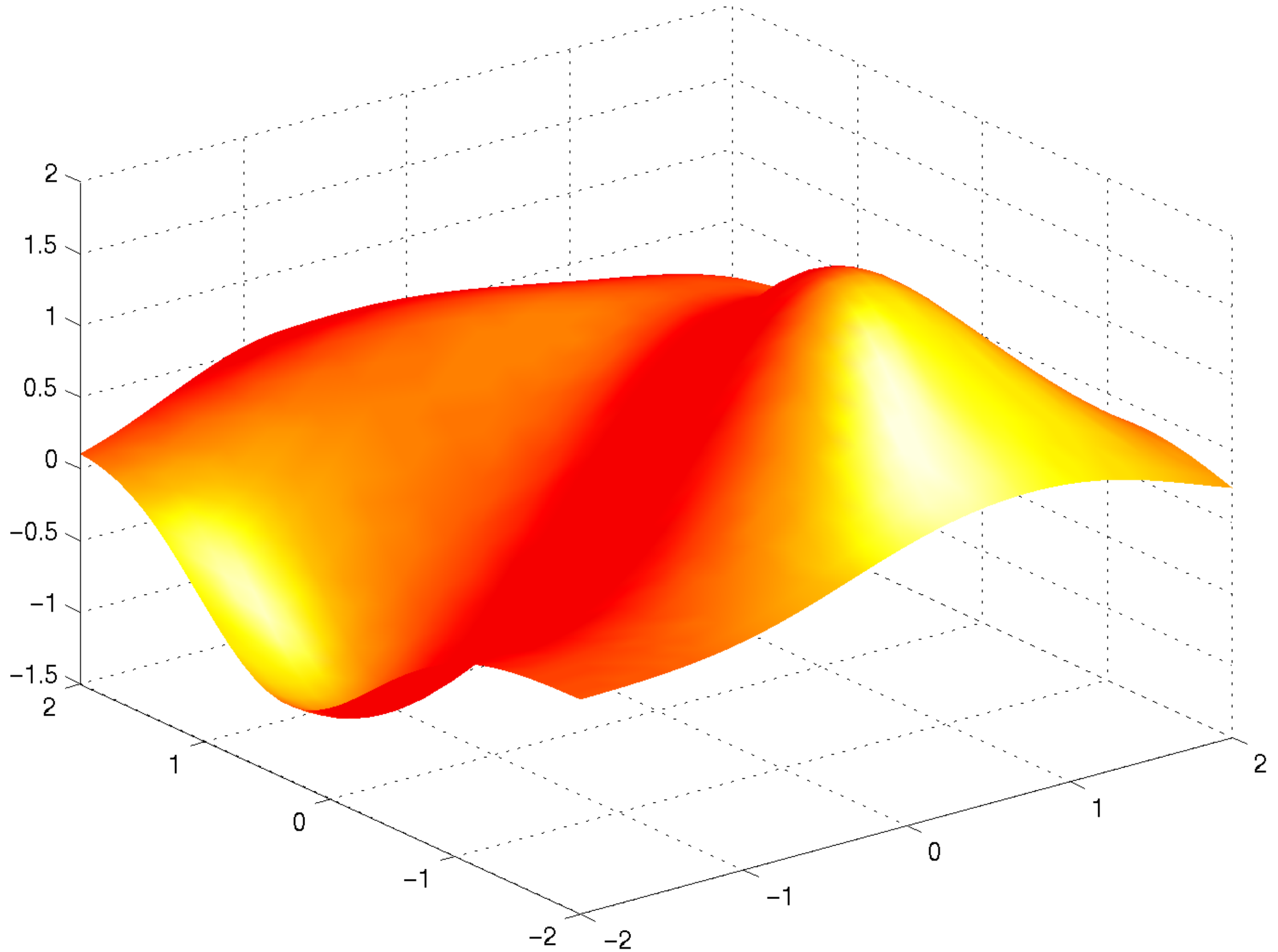
Idea 1

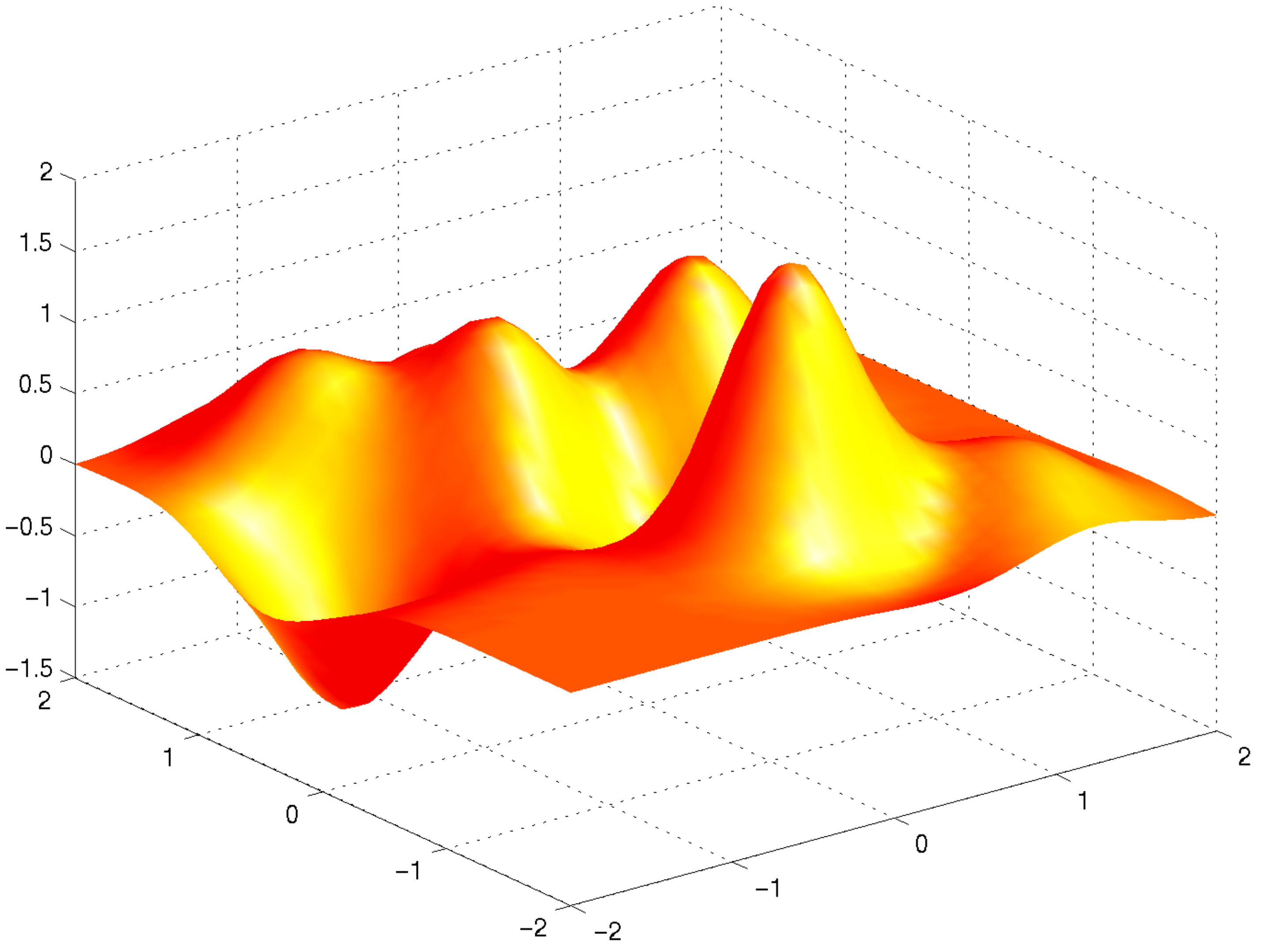
Quite often we have a rough idea of what function we can expect beforehand.

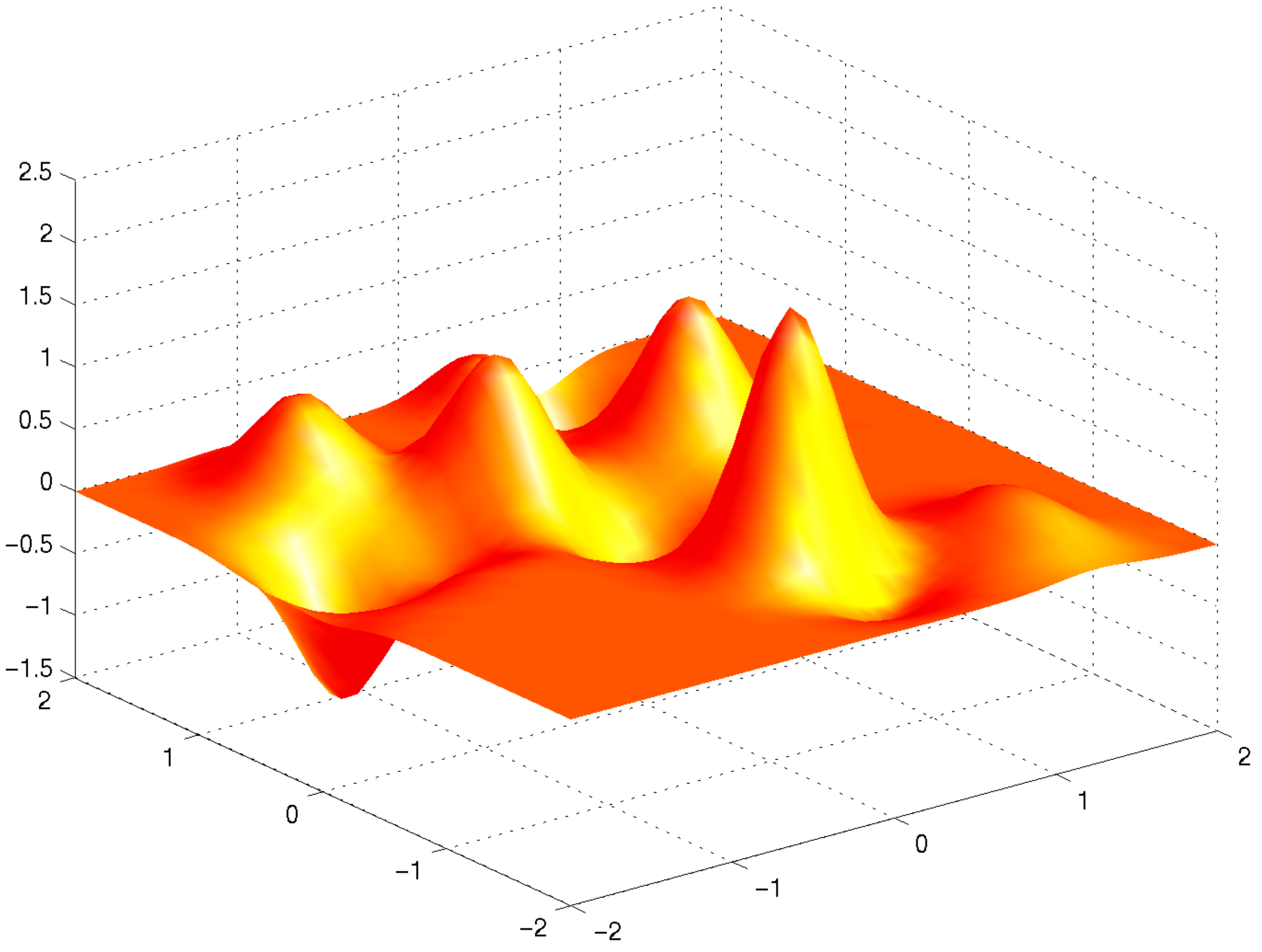
- We observe similar functions in practice.
- We **think** that e.g. smooth functions should be more likely.
- We **would like** a certain type of functions.
- We have **prior knowledge** about specific properties, e.g. vanishing second derivative, etc.

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

Example: Prior on Function Space

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

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

Bayesian Inference

Applying Bayes Rule:

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

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

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).

Prediction: Given $p(f|X,Y)$ we can predict $f(\mathbf{x})$ via

$$\int f(\mathbf{x})p(f|X,Y)df = \frac{1}{Z}\int f(\mathbf{x})p(Y|f,X)dp(f) \text{ where } Z = \int p(Y|f,X)dp(f)$$

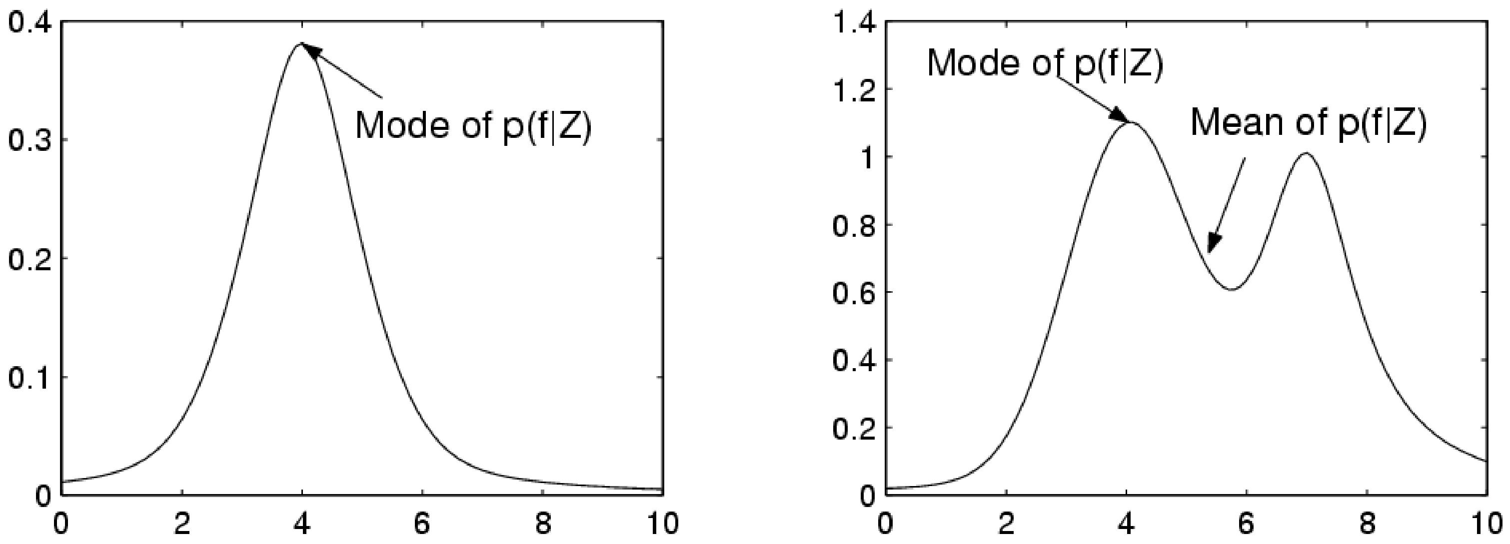
Confidence

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 $p(f)$.



Problems with Exact Inference

THE AUSTRALIAN NATIONAL UNIVERSITY

Problem

Nobody wants to compute integrals, because ...

- Computing integrals is expensive
- No closed form possible
- Not very intuitive for inference

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).

Problem

Error bars are really hard to obtain.

Idea

Approximate $p(f|X, Y)$ by a normal distribution (Laplace Approximation).

Maximum a Posteriori Approximation

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.

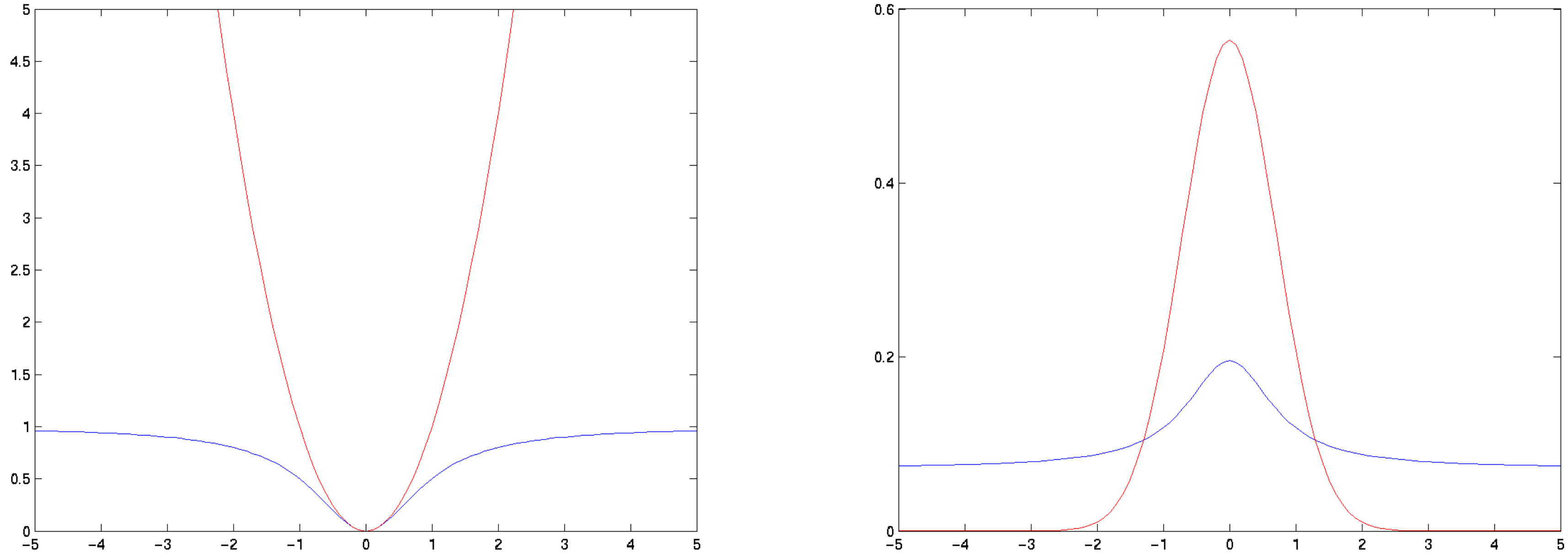
Laplace Approximation for Confidence Intervals

THE AUSTRALIAN NATIONAL UNIVERSITY

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. $\partial_f^2 - \log p(f|X, Y)$.



Relation to Regularized Risk Functional

THE AUSTRALIAN NATIONAL UNIVERSITY

Recycling of the Likelihood

Match up terms between likelihood and loss function $c(\mathbf{x}, y, f(\mathbf{x}))$. 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 the regularizer $m\lambda\Omega[f]$ 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 $\Omega[f]$ and prior $p(f)$.

Caveat

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

Hyperparameters

THE AUSTRALIAN NATIONAL UNIVERSITY

Problem

Sometimes we are not quite sure about the type of prior $p(f)$ we might have, e.g., the variance of some parameters ...

Solution

Put a **prior** on the parameters governing the prior. Instead of $p(f)$ we now have $p(f|\omega)$ and a prior $p(\omega)$ on the **hyperparameter** ω.

Effective Prior: We can obtain the effective prior by integrating out the hyperparameter

$$p(f) = \int p(f|\omega)p(\omega)d\omega$$

Inference

Using the effective prior for $p(f|X,Y)$ (and the assumption $p(f|X) = p(f)$) we obtain $p(f|X,Y) \propto p(Y|f,X)p(f) = p(Y|f,X) \int p(f|X,\omega)p(\omega)d\omega$.

MAP2 Approximation

THE AUSTRALIAN NATIONAL UNIVERSITY

Problem: Nobody wants to compute integrals, because ...

- Computing integrals is expensive
- No closed form possible
- Not very intuitive for inference

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 on the hyperparameter**.

Result

$$\underset{f,\omega}{\text{maximize}}\; p(f|X,Y) \propto p(Y|f,X)p(f|\omega)p(\omega)$$

Practical Trick

$$\underset{f,\omega}{\text{minimize}}\; -\log \underbrace{p(Y|f,X)}_{\text{Likelihood}} - \log \underbrace{p(f|\omega)}_{\text{Prior}} - \log \underbrace{p(\omega)}_{\text{Hyperprior}}$$

To integrate or not to integrate

THE AUSTRALIAN NATIONAL UNIVERSITY

Integrate

- This is what you need to do for proper inference
- Fewer Parameters
- $p(f)$ may be of a simpler functional form than $p(f|\omega)p(\omega)$, e.g.,

  $p(a|\omega) = (2\pi\omega^2)^{-\frac{1}{2}}e^{-\frac{a^2}{2\omega^2}}$ and $p(\omega) = (2\pi)^{-\frac{1}{2}}e^{-\frac{\omega^2}{2}}$ hence $p(a) = \frac{1}{2\pi}\mathrm{BesselK}(0, |a|)$.

Don't Integrate

- Sometimes easier to optimize (convex optimization problem or simple one-dimensional minimization which can be solved explicitly).
- MAP1 part may become exact (for fixed hyperparameter we have a Gaussian posterior).
- $p(f)$ may be of a simpler functional form than $p(f|\omega)p(\omega)$, e.g., if in the example above $p(\omega) = \frac{1}{2}\exp(-|\omega|)$, then $p(f)$ is really complicated ...

To integrate or not to integrate