Bayesian Inference

Mohammad Emtiyaz Khan
AIP (RIKEN), Tokyo

http://emtiyaz.github.io

emtiyaz.khan@riken.jp

June 28-29, 2018

Contents

1 Introduction

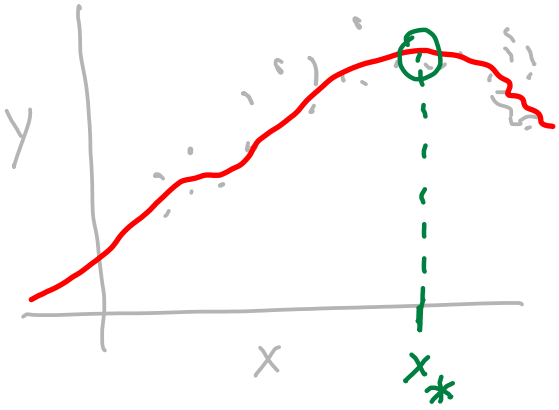
We will consider Bayesian inference in the context of supervised learning. Bayesian inference involves computation of posterior distribution, which is fundamentally different from the maximum-likelihood principle. We will demonstrate this on four models: linear regression, logistic regression, Neural networks, and Gaussian process.

By using the posterior distribution, Bayesian inference can reduce overfitting, represent uncertainty, and perform model selection.

Unfortunately, Bayesian inference involves a difficult integral which involves computing an average over all possible explanations of the data. We will learn about some of the reasons behind this difficulty. Approximate Bayesian inference addresses this problem by finding approximations to the integral

2 Regression and Classification

Regression/classification is to relate input variables to the output variable, to predict outputs for new inputs and/or to understand the effect of the input on the output.



$y_n, \underline{x}_n$

$y_n \approx f(\underline{x}_n)$

Dataset for regression

The data, denoted by $\mathcal{D}$, consist of pairs $(\mathbf{x}_n, y_n)$, where $\mathbf{x}_n$ is a vector of D inputs and y_n is the n'th output. Number of pairs N is the data-size and D is the dimensionality.

Prediction

In prediction, we wish to predict the output for a new input vector, i.e., find a regression function that approximates the output "well enough" given inputs.

$$y_n \approx f_w(\mathbf{x}_n), \text{ for all } n$$

where $\mathbf{w}$ is the parameter of the regression model.

3 Maximum Likelihood

Assume y_n to be independent samples from an *exponential-family distribution*, whose *expectation parameter* is equal to $f_w(\mathbf{x}_n)$:

$$p(\mathcal{D}|\mathbf{w}) := \prod_{n=1}^{N} p(y_n|f_w(\mathbf{x}_n)).$$

The function $p(\mathcal{D}|\mathbf{w})$ is the likelihood, which can be maximized to obtain a "good enough" $\mathbf{w}$,

$$\mathcal{L}_{ML}(\mathbf{w}) := \log p(\mathcal{D}|\mathbf{w}).$$

This is known as the maximum-likelihood estimation.

$p(y_n|f_w(x_n))$

$\prod_{n=1}^{N} \mathcal{N}(y_n|f_w(x_n),1)$

log

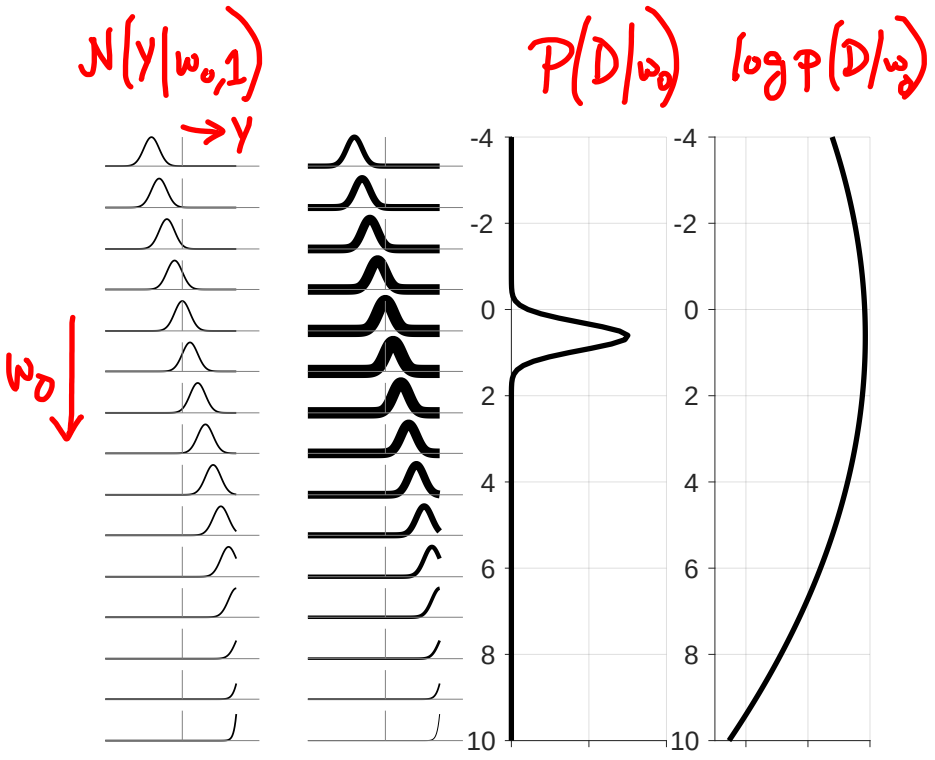
$\sum_{n=1}^{N} (y_n - f_w(x_n))^2$

1 parameter model

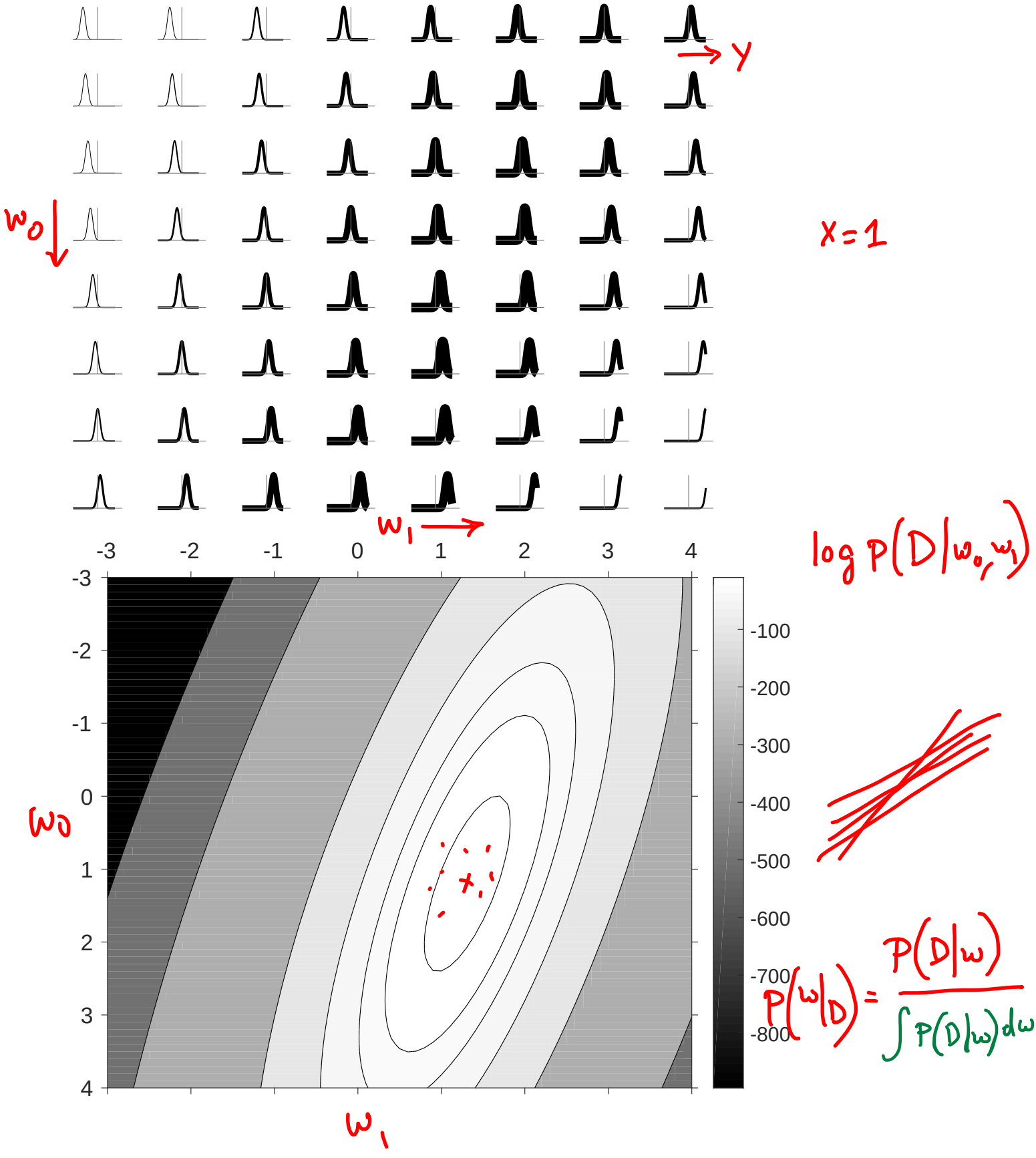
$y_n \approx w_0$

$\mathcal{N}(y_n|w_0,1)$

$\sum_{n=1}^{N} \log \mathcal{N}(y_n|w_0,1)$



Two-parameter model. $y = w_0 + w_1 x$



4 Maximum A Posteriori (MAP)

To avoid overfitting, we can use a regularizer $\log p(\mathbf{w})$ to perform maximum a posteriori estimation,

$$\mathcal{L}_{MAP}(\mathbf{w}) := \log p(\mathcal{D}|\mathbf{w}) + \log p(\mathbf{w}).$$

Note that not all regularizers correspond to a probability distribution.

We can view $p(\mathbf{w})$ as a prior distribution to get the joint distribution:

$$\begin{aligned} p(\mathcal{D}, \mathbf{w}) &= p(\mathcal{D}|\mathbf{w})p(\mathbf{w}), \\ &= p(w|\mathcal{D})p(\mathcal{D}). \end{aligned}$$

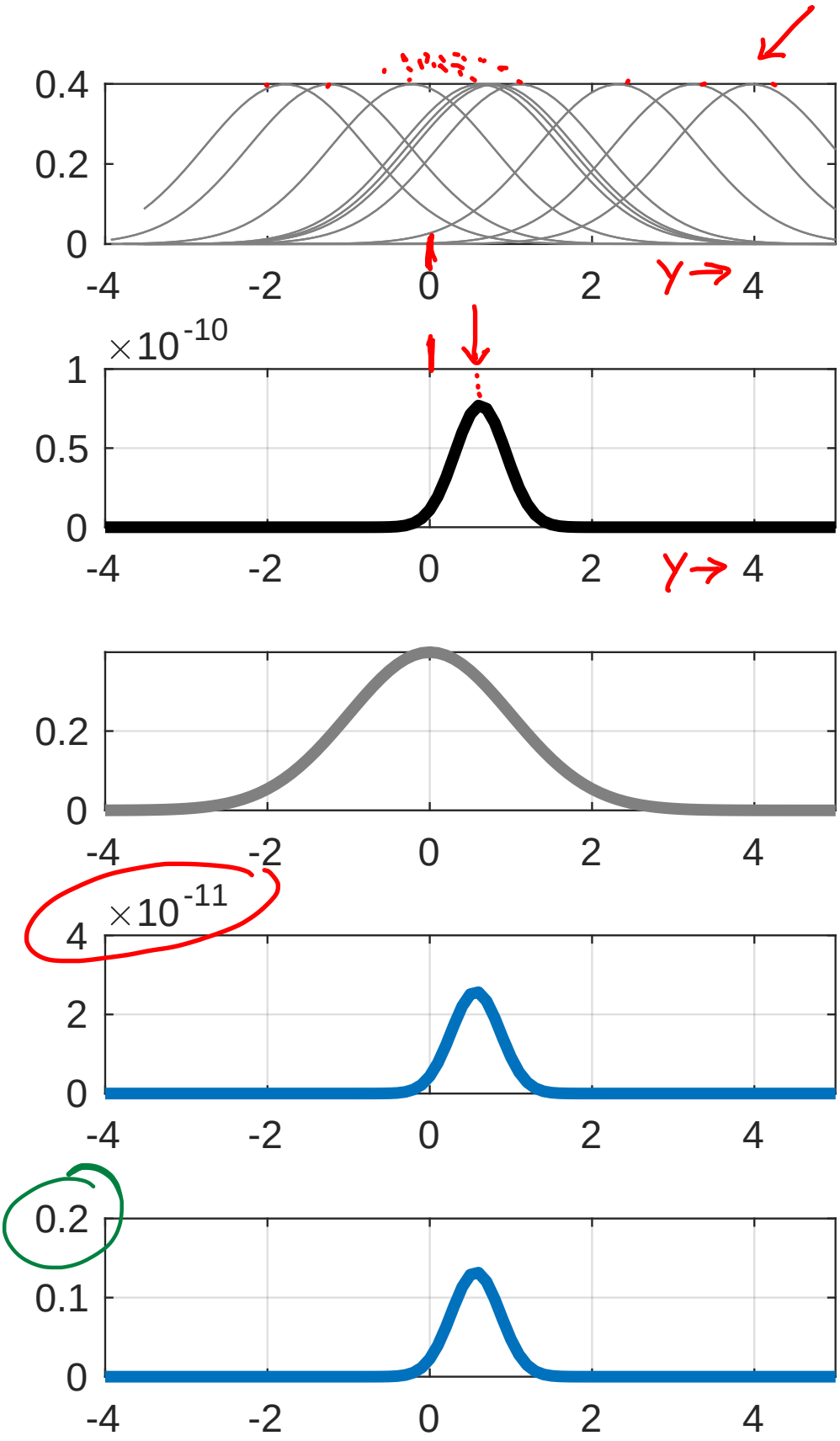
5 The Posterior Distribution

The posterior distribution is defined using the Bayes' rule:

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{\int p(\mathcal{D}|\mathbf{w})p(\mathbf{w})d\mathbf{w}}.$$

The integral $p(\mathcal{D})$ is the normalizing constant, also known as, the marginal likelihood. Without it, we do not know the ~~true~~ *spread* of the distribution, which gives us a notion of uncertainty or confidence.

Example: One-Parameter Model



$y_n \simeq w_0$

$N(y_n | w_0, 1)$

lik

$\prod_{n=1}^{N} N(y_n | w_0, 1)$

Prior

$N(w_0 | 0, 1)$

$\frac{lik(w_0)\, Prior(w_0)}{\sum_{w_0} lik(w_0)\, Prior(w_0)}$

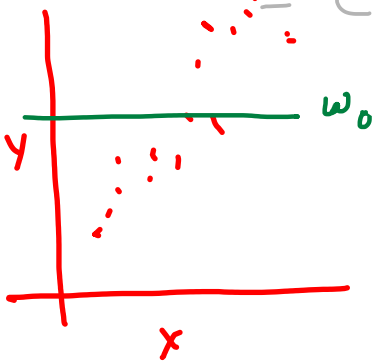
$P(w_0 | D)$

Posterior of a One-Parameter Model

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{\int p(\mathcal{D}|\mathbf{w})p(\mathbf{w})d\mathbf{w}},$$

$$p(w_0|\mathcal{D}) = \frac{\left[\prod_{n=1}^{N}\mathcal{N}(y_n|w_0,1)\right]\mathcal{N}(w_0|0,1)}{\int\left[\prod_{n=1}^{N}\mathcal{N}(y_n|w_0,1)\right]\mathcal{N}(w_0|0,1)dw_0}.$$

Question:Derive the posterior distribution and the marginal likelihood.



$p(\mathcal{D}) =$

$p(w_0|\mathcal{D}) =$

Example: Two-Parameter Model

$y_n \approx w_0 + w_1 x_n$

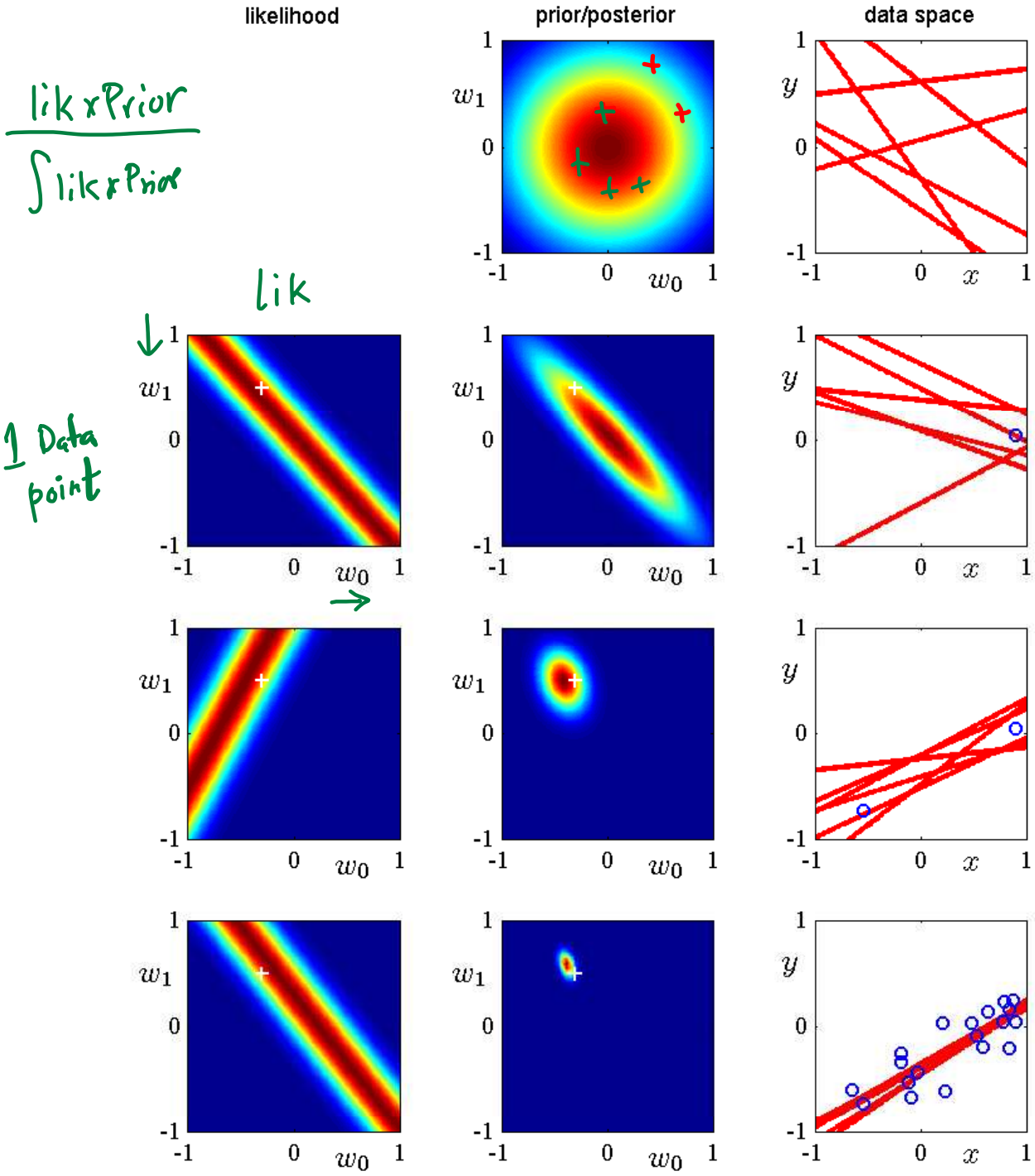


Figure taken from [Bishop, 2006]

Example: Two-Parameter Model

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{\int p(\mathcal{D}|\mathbf{w})p(\mathbf{w})d\mathbf{w}},$$

$$p(\mathbf{w}|\mathcal{D}) = \frac{\left[\prod_{n=1}^{N} \mathcal{N}(y_n|w_0 + w_1 x_n, 1)\right] \mathcal{N}(\mathbf{w}|0, \mathbf{I})}{\int \left[\prod_{n=1}^{N} \mathcal{N}(y_n|w_0 + w_1 x_n, 1)\right] \mathcal{N}(\mathbf{w}|0, \mathbf{I}) d\mathbf{w}}.$$

Question:Derive the posterior distribution and the marginal likelihood.

$[w_0 \; w_1] \begin{bmatrix} 1 \\ x_n \end{bmatrix}$

$$-\frac{1}{2}\left[\sum_{n=1}^{N}(y_n - w^T\tilde{x}_n)^T(y_n - w^T\tilde{x}_n) + w^T w\right]$$

$$\sum_n \left[y_n y_n - 2y_n x_n^T w + w^T x_n x_n^T w\right] + w^T w$$

$$w^T\left(I + X^T X\right)w - 2\,y^T X\, w + \text{const}$$

V^{-1} $\qquad$ $V^{-1}m$

$p(\mathcal{D}) =$

$p(w_0, w_1|\mathcal{D}) =$

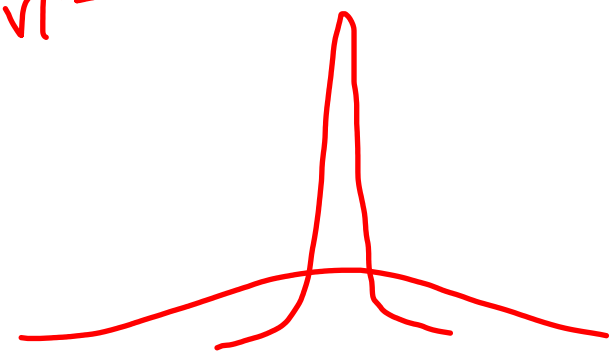
6 Bayesian Linear Regression

Consider $f_w(\mathbf{x}_n) := \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n)$, (where $\boldsymbol{\phi}(\cdot)$ is a nonlinear function) with the following joint distribution,

$$p(\mathcal{D}, \mathbf{w}) = \left[\prod_{n=1}^{N} \mathcal{N}(y_n | f_w(\mathbf{x}_n), 1)\right] \mathcal{N}(\mathbf{w}|0, \lambda\mathbf{I}).$$

Question:Derive the posterior distribution and the marginal likelihood.

$$p(\mathbf{w}|\mathcal{D}) = \mathcal{N}(\mathbf{w}|\mathbf{m}, \mathbf{V}),$$
$$\mathbf{m} = (\lambda\mathbf{I} + \boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top\mathbf{y}$$
$$\mathbf{V} = (\lambda\mathbf{I} + \boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}$$
$$p(\mathcal{D}) = \mathcal{N}(\mathbf{y}|\mathbf{0}, \boldsymbol{\Phi}\boldsymbol{\Phi}^\top + \mathbf{I}/\lambda)$$



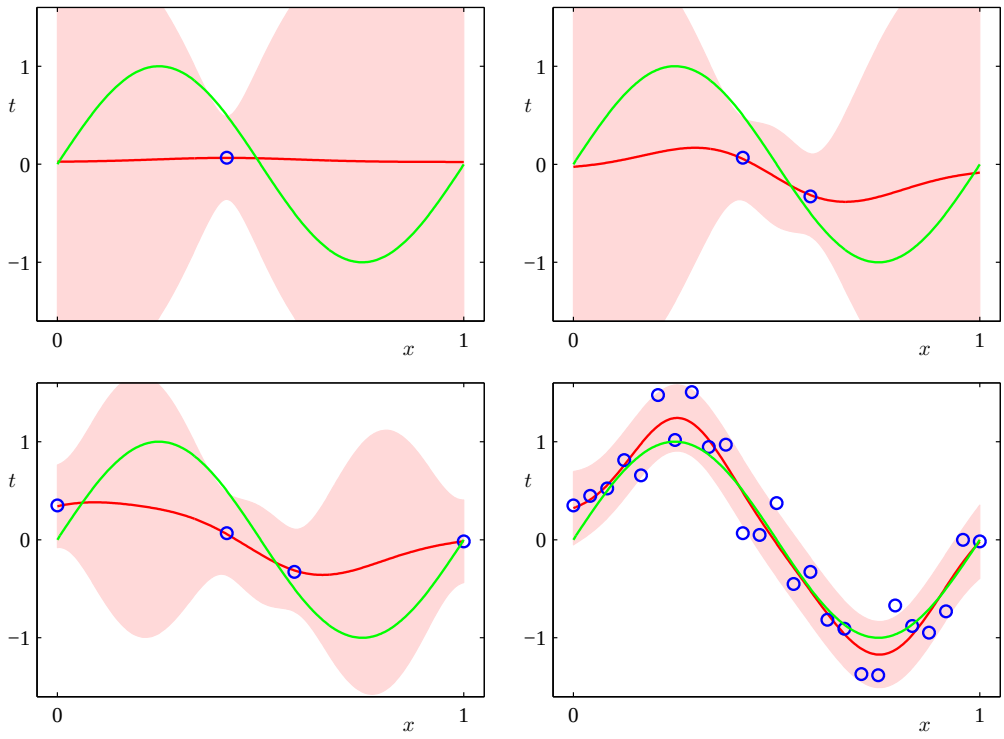
Question:When is the computation of the posterior distribution difficult?

The predictive distribution is,

$$p(y_*|\mathbf{x}_*, \mathcal{D}) = \int \mathcal{N}(y_*|f_w(\mathbf{x}_*), 1)p(\mathbf{w}|\mathcal{D})d\mathbf{w},$$
$$= \mathcal{N}\left(y_*|\mathbf{m}^\top\boldsymbol{\phi}(\mathbf{x}_*), 1 + \boldsymbol{\phi}(\mathbf{x}_*)^\top\mathbf{V}\boldsymbol{\phi}(\mathbf{x}_*)\right).$$

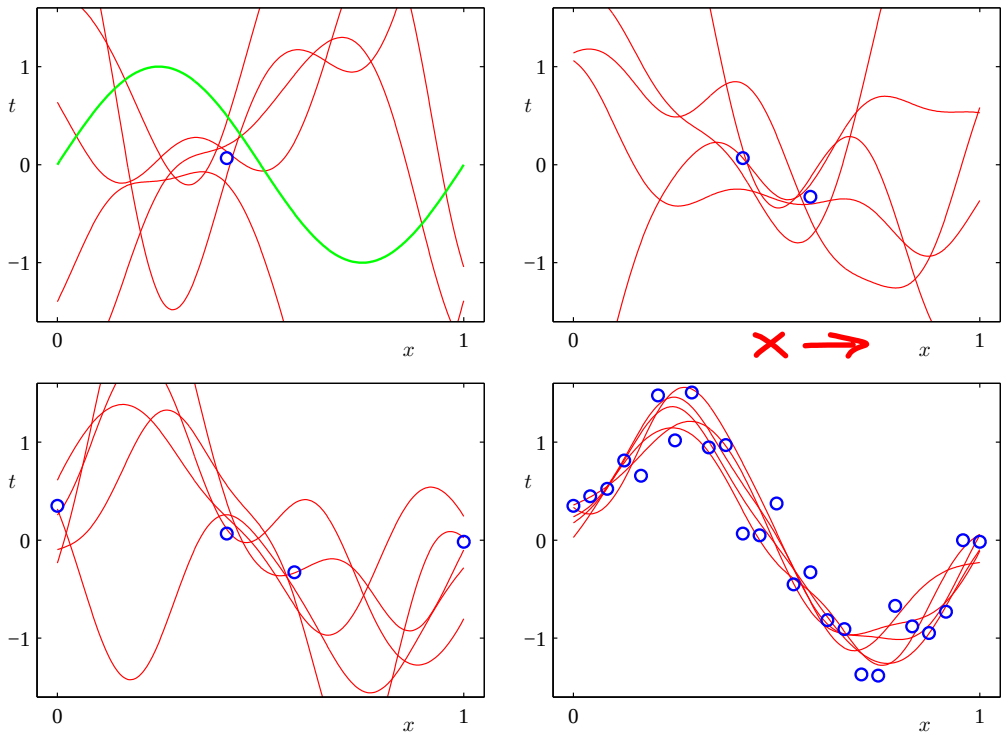
Example of Predictive Distribution

From [Bishop, 2006] Figure 3.8.



Sampled Prediction Functions

From [Bishop, 2006] Figure 3.9.



$f_w(x)$

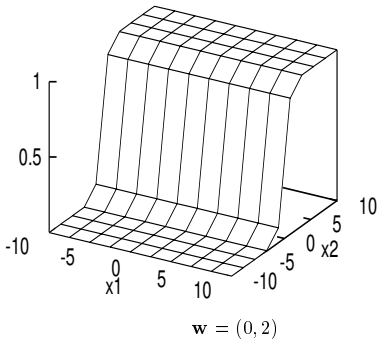
7 Bayesian Logistic Regression

When $y_n \in \{0, 1\}$, we can use a Bernoulli distribution,

$$p(\mathcal{D}, \mathbf{w}) = \left[\prod_{n=1}^{N} \text{Ber}(y_n | f_w(\mathbf{x}_n))\right] \mathcal{N}(\mathbf{w}|0, \lambda\mathbf{I}),$$

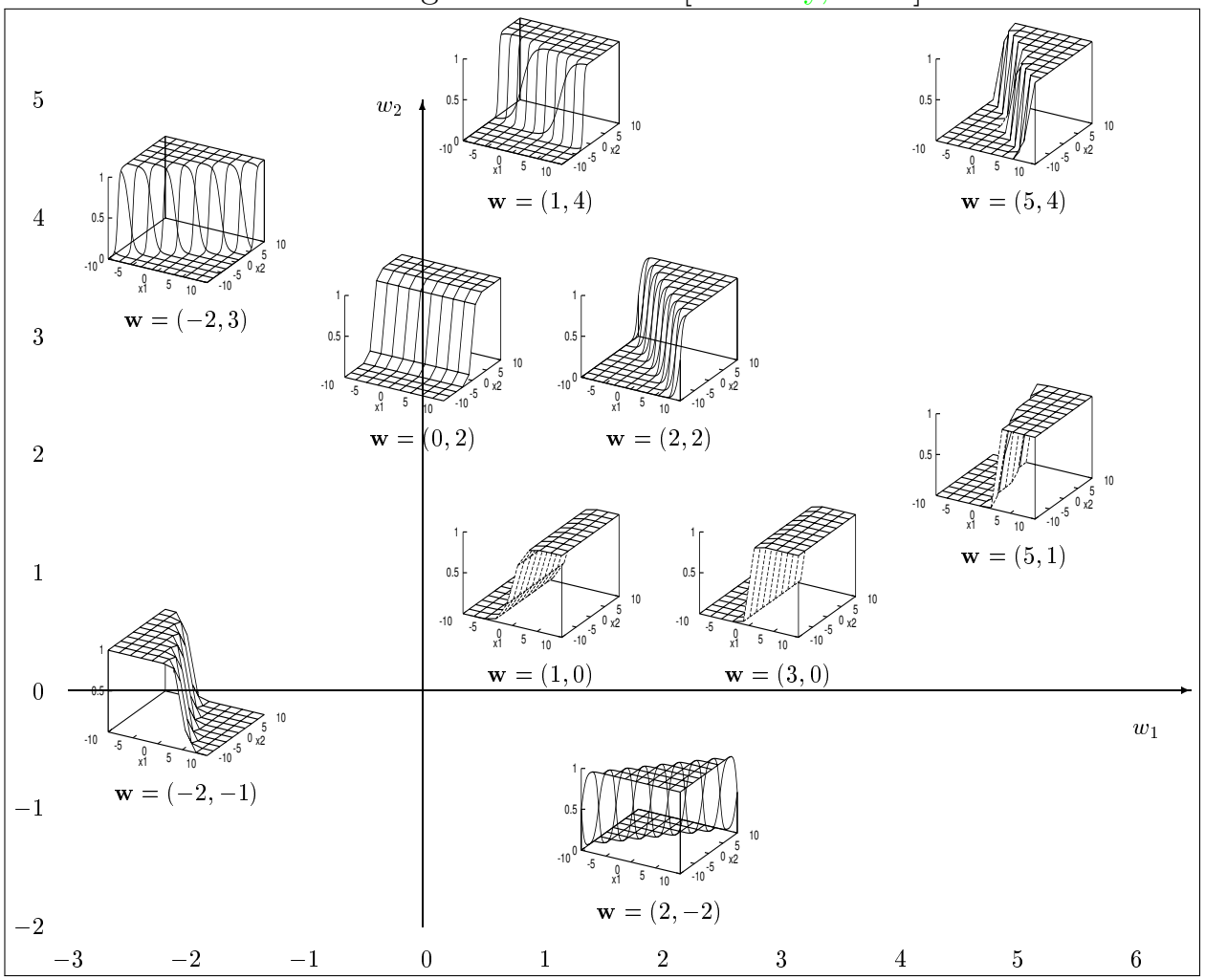
$$p(y_n = 1|f_w(\mathbf{x}_n)) = \frac{1}{1 + e^{\mathbf{w}^\top \mathbf{x}_n}}.$$

$w_1 x_{n1} + w_2 x_{n2}$

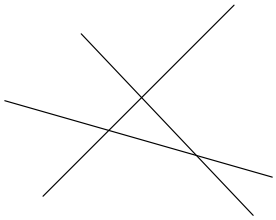


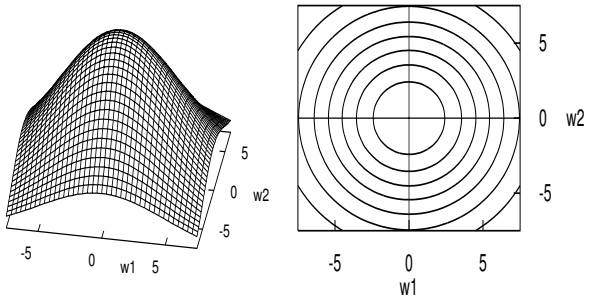
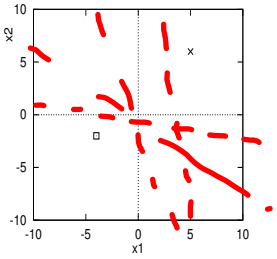
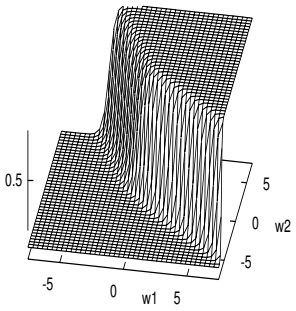
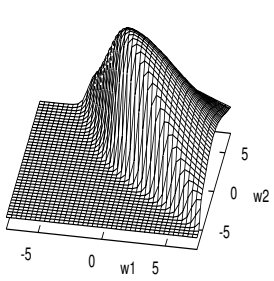
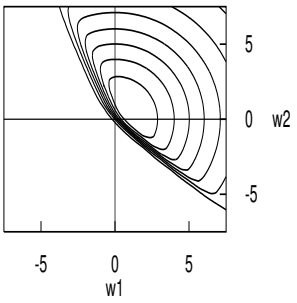
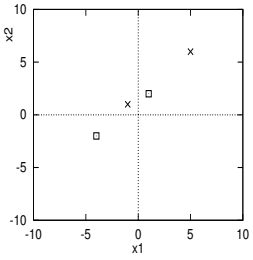
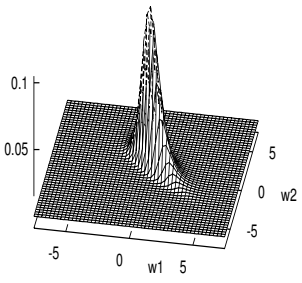
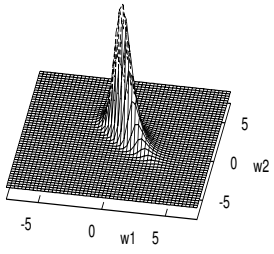
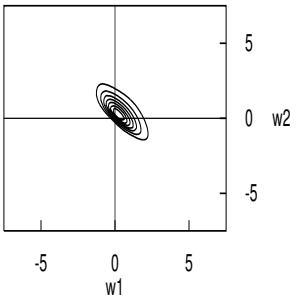
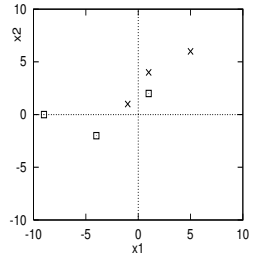
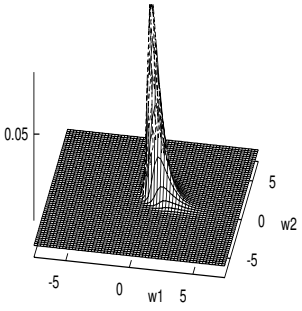
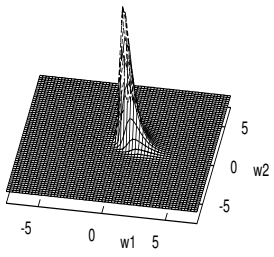
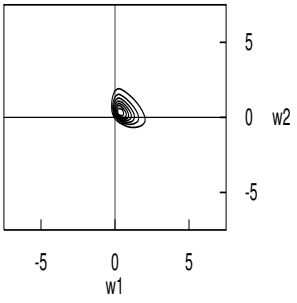
$\mathbf{w} = (0, 2)$

Figure taken from [MacKay, 2003]



2D Example



Data set	Likelihood	Probability of parameters
$N = 0$	*(constant)*	
$N = 2$		  
$N = 4$		  
$N = 6$		  

Posterior Distribution

The posterior distribution is,

$$p(\mathbf{w}|\mathcal{D}) =$$

The marginal-likelihood is,

$$p(\mathcal{D}) = \int \left\{ \prod_{n=1}^{N} \left[\frac{1}{1+e^{\mathbf{w}^T \mathbf{x}_n}} \right]^{y_n} \left[1 - \frac{1}{1+e^{\mathbf{w}^T \mathbf{x}_n}} \right]^{1-y_n} \right\} \mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{I}) d\mathbf{w}$$

$$e^{-\frac{1}{2}\mathbf{w}^T \mathbf{w}}$$

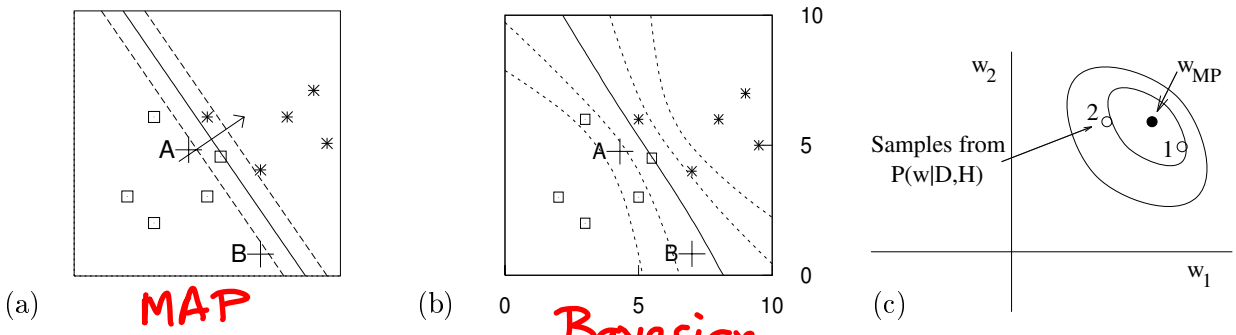
There is no closed-form solution. This is because the likelihood is not conjugate to the Gaussian prior. That is, with respect to $\mathbf{w}$, the likelihood cannot be expressed in the same form as the prior.

Predictive Distribution

The predictive distribution is intractable too,

$$p(y_*|\mathbf{x}_*, \mathcal{D}) = \int \text{Ber}(y_*|f_w(\mathbf{x}_*)) p(\mathbf{w}|\mathcal{D}) d\mathbf{w}.$$

Figure taken from [MacKay, 2003]

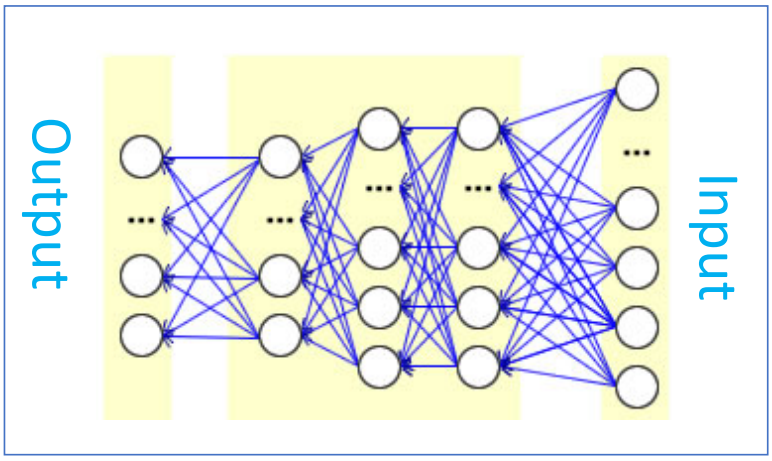


The prediction uncertainty is useful to avoid "overconfident" decision boundary found by a MAP method.

8 Deep Neural Networks

These difficult become much worse with deep models.

$f_w(\mathbf{x}_n) = f_1(\mathbf{w}_1 f_2(\mathbf{w}_2 \dots f_L(\mathbf{w}_L \mathbf{x}_n)))$



Nonconvexity translates to a *multi-modal* posterior distribution, whose marginal-likelihood is a massive, intractable integral.

Three computational challenges:
1) Too many factors (large N)
2) Too many dimensions (large D)
3) Nonconjugacy

9 Gaussian Processes

Instead of assigning a Gaussian prior to $\mathbf{w}$, we can directly apply an informative prior on $f_w(\mathbf{x})$ using Gaussian process, which is defined using a kernel matrix $\mathbf{K}_\theta$ with entries $k_{ij} := k(\mathbf{x}_i, \mathbf{x}_j)$ where $\boldsymbol{\theta}$ are kernel hyperparameters, e.g.,

$$k_{ij} := \boldsymbol{\phi}(\mathbf{x}_i)^T \boldsymbol{\phi}(\mathbf{x}_j).$$

We can then directly sample N function values $\mathbf{f} = [f_1, f_2, \ldots, f_N]$ from a Gaussian-process prior:

$$p(\mathbf{f}|\mathbf{X}, \boldsymbol{\theta}) := \mathcal{N}(\mathbf{f}|0, \mathbf{K}_\theta).$$

No need for $\mathbf{w}$! Combining it with a likelihood, we get GP models.

A good text is [Rasmussen and Williams, 2006]

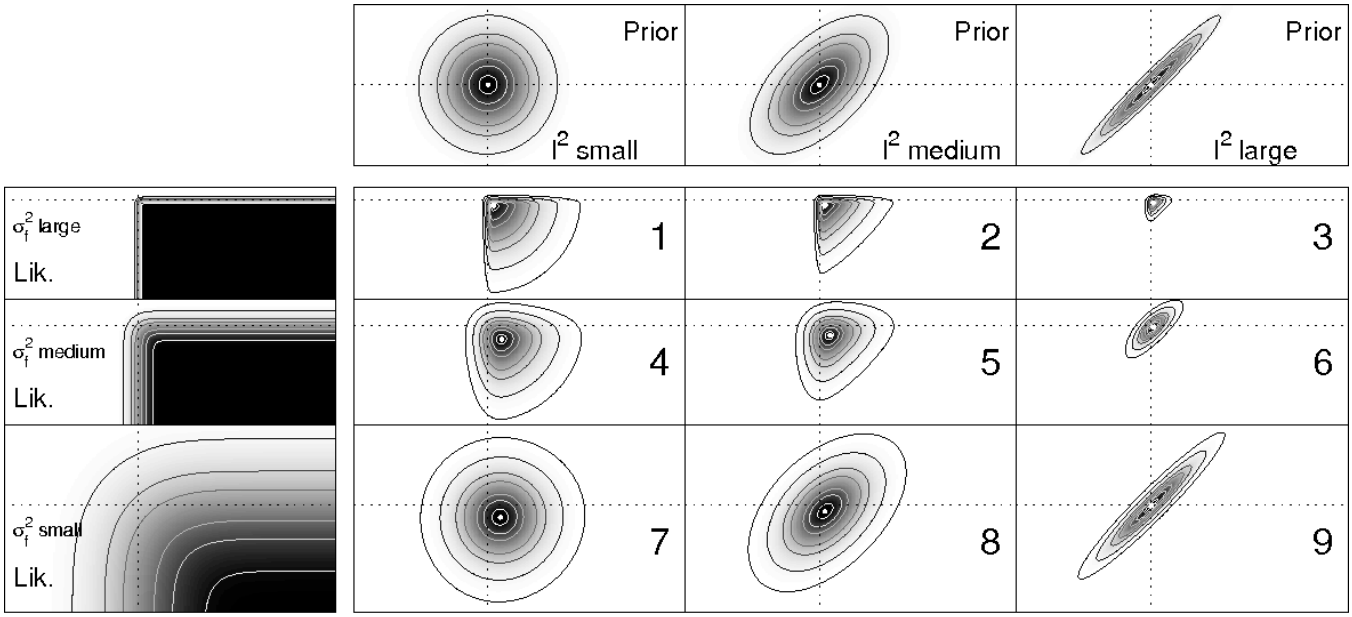
2D Example

Consider square-exponential kernel:

$$k_{ij} := \sigma_f^2 \exp\left(-\frac{1}{2l}\|\mathbf{x}_i - \mathbf{x}_j\|^2\right),$$

along with a Bernoulli likelihood:

$$p(\mathcal{D}, \mathbf{w}) = \left[\prod_{n=1}^{N} \text{Ber}\left(y_n \middle| \frac{1}{1+\exp(f_n)}\right)\right] \mathcal{N}(\mathbf{f}|0, \mathbf{K}_\theta).$$



Taken from [Nickisch and Rasmussen, 2008]

This is again a nonconjugate model, which results in an intractable integral. Could you think of a lower bound on the computation?

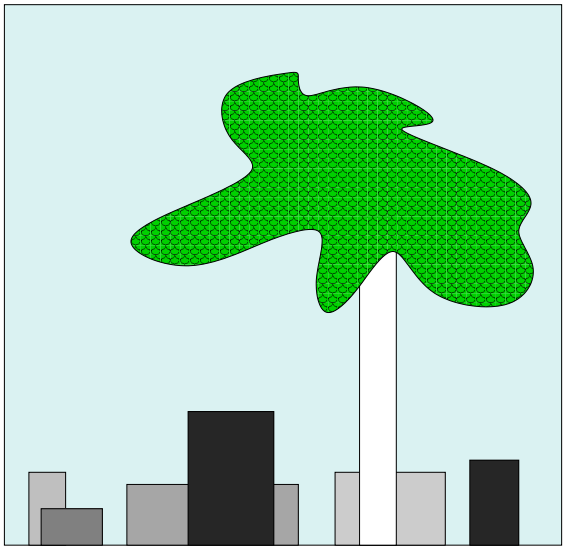
Hint: Use a Gaussian likelihood.

10 Benefits of Bayesian Inference

There are multiple benefits:
1) Posterior distribution gives an estimate of uncertainty.
2) Averaging with respect to it can reduce overfitting.
3) Posterior distribution enables data-generation.
4) Marginal likelihood enables model-selection.
We give more details about the last point now.

Occam's Razor

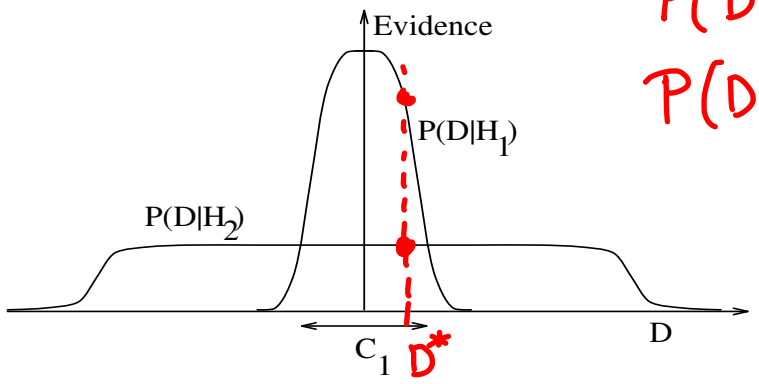
What is behind the tree?



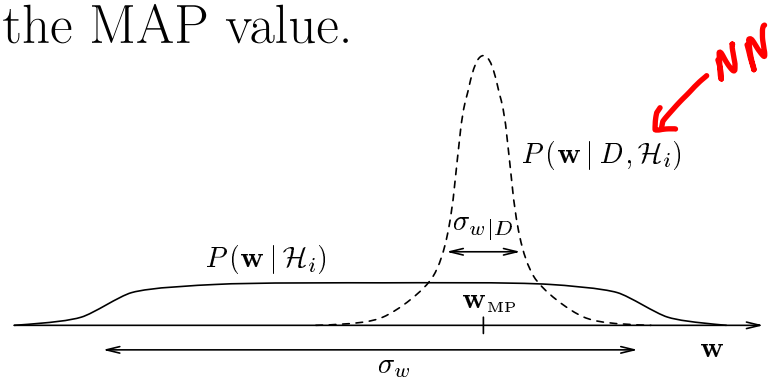
From [MacKay, 2003].
Simpler explanations are better – in absence of certainty. Bayesian inference naturally incorporates this principle.

Bayesian Razor

A simple model $\mathcal{H}_1$ can make only limited range of predictions well, while a complex models $\mathcal{H}_2$ can cover more range. However, this means that $\mathcal{H}_2$ does not predict the data set in region C_1 as strongly as $\mathcal{H}_1$. Suppose the two models have equal prior probabilities, then $\mathcal{H}_1$ will be more probable than $\mathcal{H}_2$ in the region C_1.



In a similar fashion, at the parameter level the posterior distribution $p(\mathbf{w}|\mathcal{D}, \mathcal{H}_i)$ incorporates the Occam factor due to the "spread" around the MAP value.

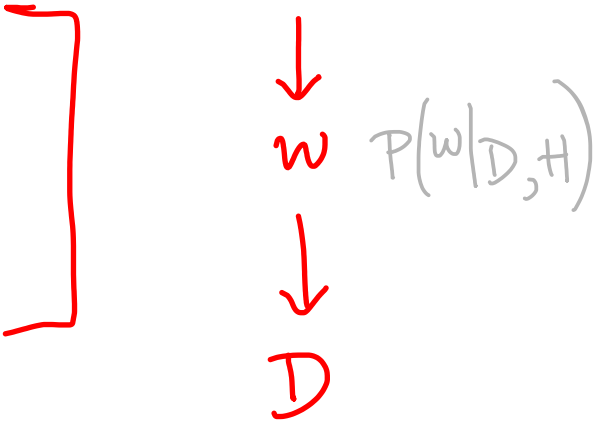


From [MacKay, 2003].

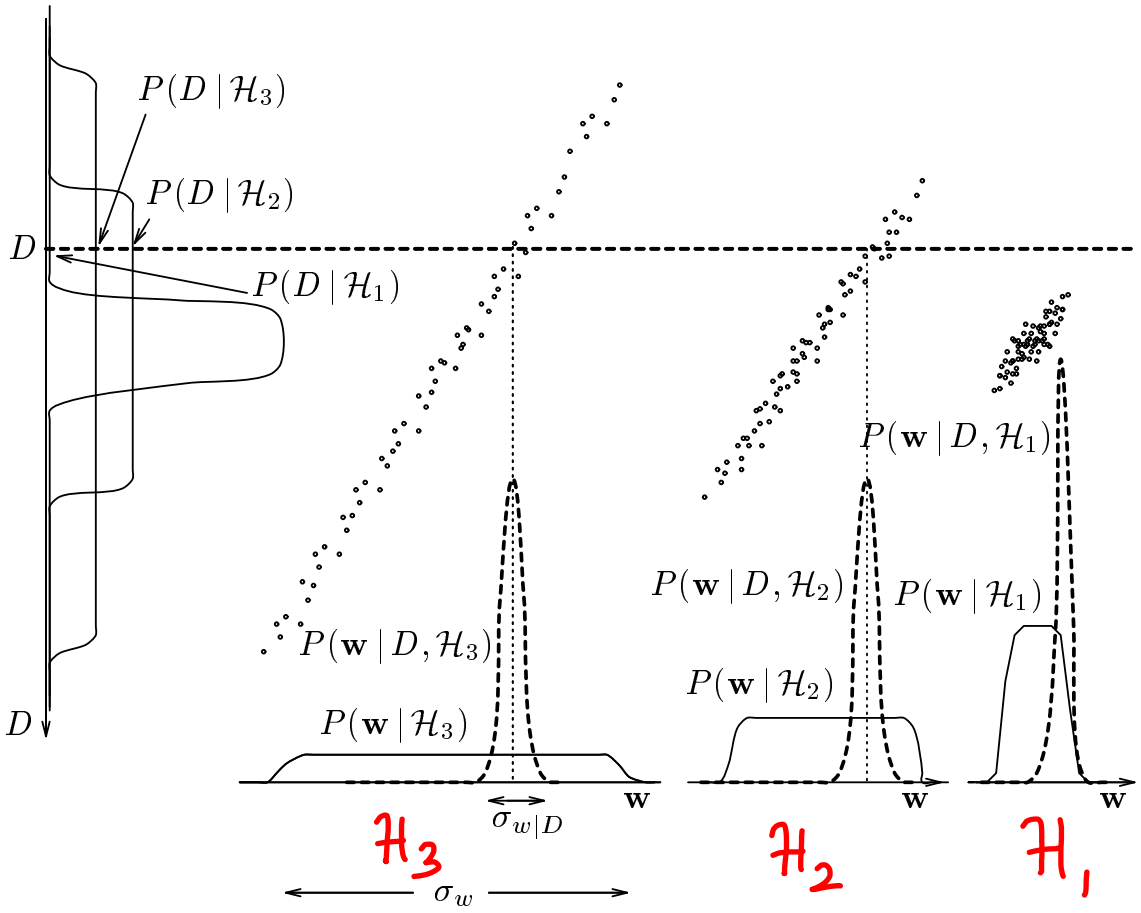
Bayesian inference performs such selection at both model and parameter levels:

$$p(\mathbf{w}|\mathcal{D}, \mathcal{H}_i) = \frac{p(\mathcal{D}|\mathbf{w}, \mathcal{H}_i)p(\mathbf{w}|\mathcal{H}_i)}{p(\mathcal{D}|\mathcal{H}_i)},$$

$$p(\mathcal{H}_i|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{H}_i)p(\mathcal{H}_i)}{p(\mathcal{D})}.$$



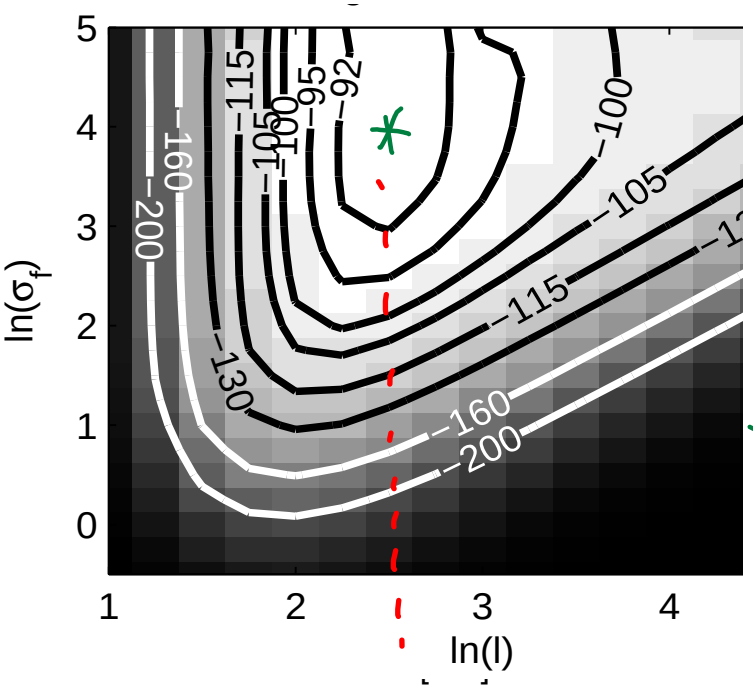
This beautiful illustration by [MacKay, 2003] demonstrates this point for a simple model $\mathcal{D} = w + \text{noise}$. Complexity of the model varies as $\mathcal{H}_3 > \mathcal{H}_2 > \mathcal{H}_1$, but all models have equal prior probabilities.

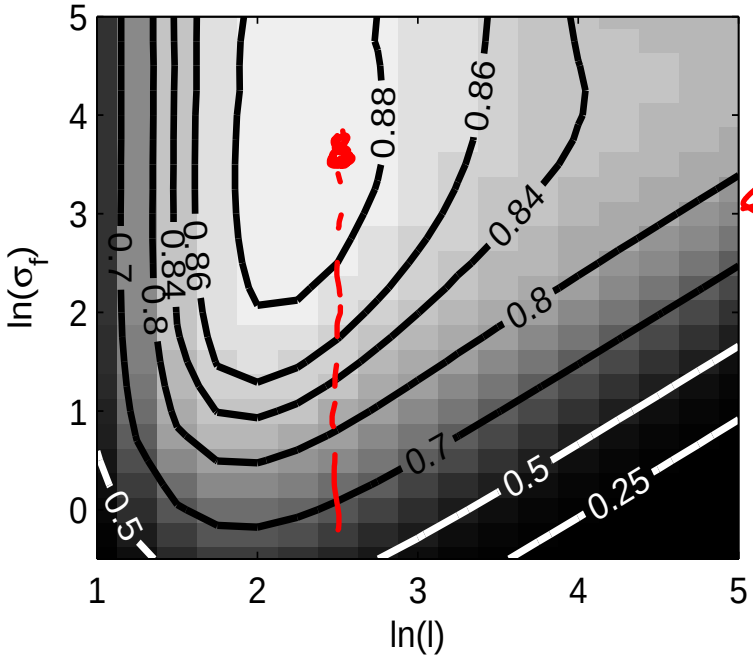


From [MacKay, 2003].

Demonstration: GP Classification

Marginal likelihood on the training data (left) reflects the shape of generalization error evaluated on test data (right).

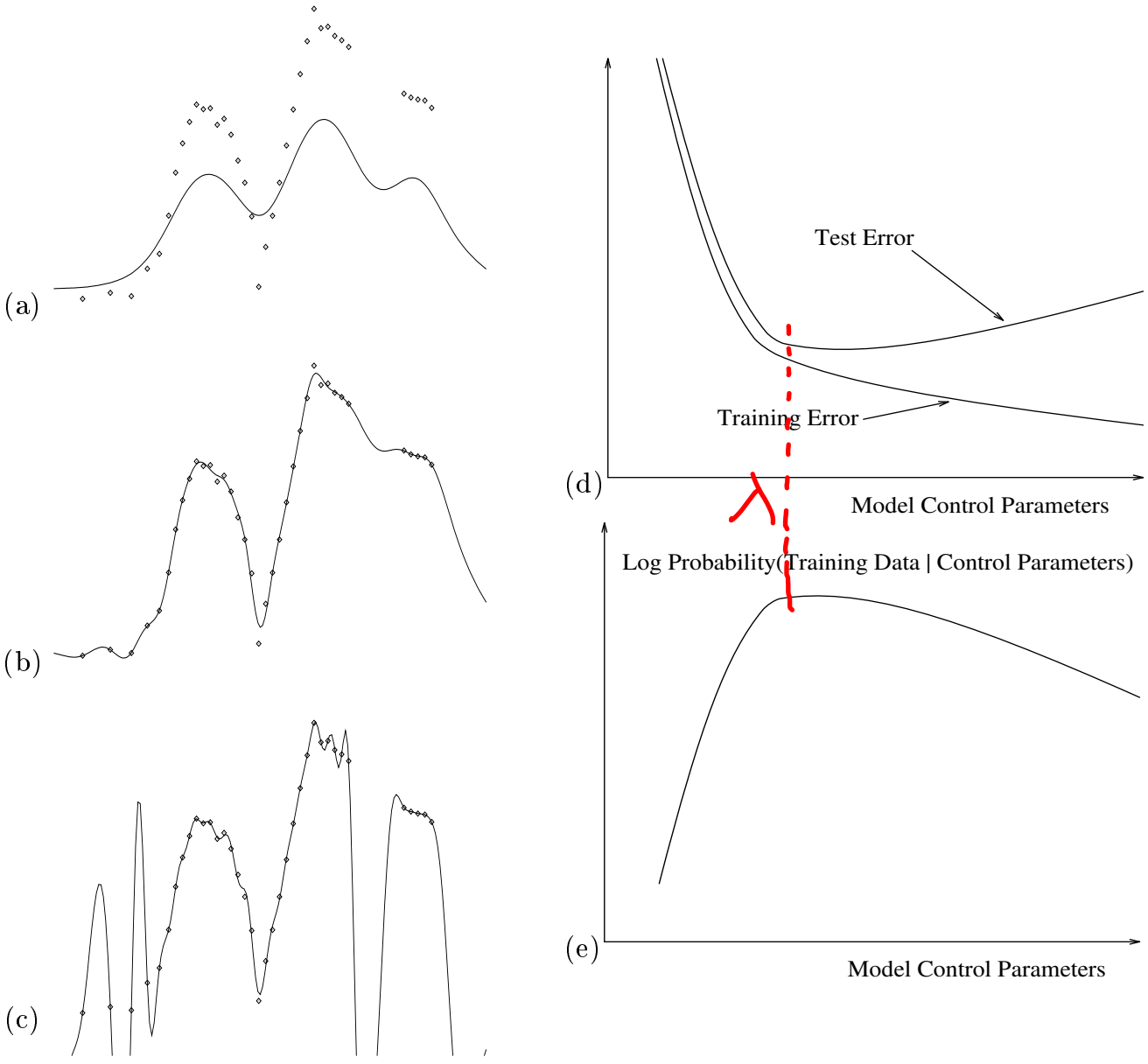




From [Rasmussen and Williams, 2006].

Demonstration: Neural Networks

When increasing the model complexity, the neural network easily overfits on the test data, but the marginal likelihood on training data reflects the generalization error.



From [MacKay, 2003].

11 Summary

The main computational issue is that we need to average over all possible $\mathbf{w}$, which is very difficult to do exactly.

$$p(\mathcal{D}) = \int \left[\prod_{n=1}^{N} p(y_n | f_w(\mathbf{x}_n)) \right] p(\mathbf{w}) d\mathbf{w}.$$

There are at-least three computational challenges:
1) Too many factors (large N)
2) Too many dimensions (large D)
3) Nonconjugacy (e.g., non Gaussian likelihood with a Gaussian prior)

References

[Bishop, 2006] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*.

[Khan, 2009] Khan, M. E. (2009). Gaussian likelihood with Gaussian prior on its mean. https://emtiyaz.github.io/Writings/normNormPdf.pdf [Online; accessed 21-July-2018].

[MacKay, 2003] MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

[Nickisch and Rasmussen, 2008] Nickisch, H. and Rasmussen, C. (2008). Approximations for binary Gaussian process classification. *Journal of Machine Learning Research*, 9(10).

[Rasmussen and Williams, 2006] Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

List of concepts

(Notes)

(Notes)