# *Introduction to Statistical Machine Learning*


Christfried Webers

Statistical Machine Learning Group
NICTA
and
College of Engineering and Computer Science
The Australian National University


Canberra
February – June 2013

(Many figures from C. M. Bishop, "Pattern Recognition and Machine Learning")



*Outlines*

# Part VI

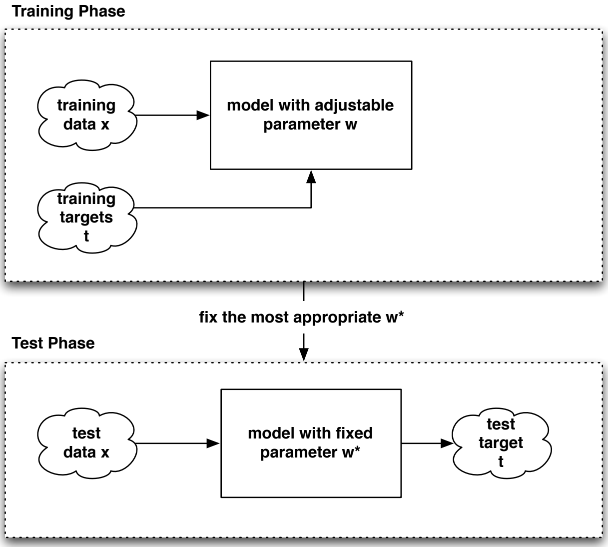
# *Linear Regression 2*



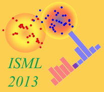


ISML 2013

# Bayesian Regression





# Bayesian Regression

- Bayes Theorem

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{normalisation}} \qquad p(\mathbf{w} \,|\, \mathbf{t}) = \frac{p(\mathbf{t} \,|\, \mathbf{w})\, p(\mathbf{w})}{p(\mathbf{t})}$$

- likelihood for i.i.d. data

$$\begin{aligned} p(\mathbf{t} \,|\, \mathbf{w}) &= \prod_{n=1}^{N} \mathcal{N}(t_n \,|\, y(\mathbf{x}_n, \mathbf{w}), \beta^{-1}) \\ &= \prod_{n=1}^{N} \mathcal{N}(t_n \,|\, \mathbf{w}^T \phi(\mathbf{x}_n), \beta^{-1}) \\ &= \text{const} \times \exp\{-\beta \frac{1}{2} (\mathbf{t} - \mathbf{\Phi} \mathbf{w})^T (\mathbf{t} - \mathbf{\Phi} \mathbf{w})\} \end{aligned}$$

where we left out the conditioning on $\mathbf{x}$ (always assumed), and $\beta$, which is assumed to be constant.

# How to choose a prior?

conjugate priors are preferred.
Normal * Normal = Normal distribution

- Can we find a prior for the given likelihood which
  - makes sense for the problem at hand
  - allows us to find a posterior in a 'nice' form

An answer to the second question:

*Definition ( Conjugate Prior)*

A class of prior probability distributions $p(w)$ is conjugate to a class of likelihood functions $p(x \mid w)$ if the resulting posterior distributions $p(w \mid x)$ are in the same family as $p(w)$.

# Examples of Conjugate Prior Distributions

*Table :* Discrete likelihood distributions

| Likelihood | Conjugate Prior |
|---|---|
| Bernoulli | Beta |
| Binomial | Beta |
| Poisson | Gamma |
| Multinomial | Dirichlet |

*Table :* Continuous likelihood distributions

| Likelihood | Conjugate Prior |
|---|---|
| Uniform | Pareto |
| Exponential | Gamma |
| Normal | Normal |
| Multivariate normal | Multivariate normal |

# Conjugate Prior to a Gaussian Distribution

Bernouli, Beta, Gaussian all have this property

- Example : The Gaussian family is conjugate to itself with respect to a Gaussian likelihood function: if the likelihood function is Gaussian, choosing a Gaussian prior will ensure that the posterior distribution is also Gaussian.
- Given a marginal distribution for $\mathbf{x}$ and a conditional Gaussian distribution for $\mathbf{y}$ given $\mathbf{x}$ in the form

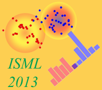
$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x} \,|\, \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$$
$$p(\mathbf{y} \,|\, \mathbf{x}) = \mathcal{N}(\mathbf{y} \,|\, \boldsymbol{A}x + \mathbf{b}, \boldsymbol{L}^{-1})$$

- we get

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y} \,|\, \boldsymbol{A}\boldsymbol{\mu} + \mathbf{b}, \boldsymbol{L}^{-1} + \boldsymbol{A}\boldsymbol{\Lambda}^{-1}\boldsymbol{A}^T)$$
$$p(\mathbf{x} \,|\, \mathbf{y}) = \mathcal{N}(\mathbf{x} \,|\, \boldsymbol{\Sigma}\{\boldsymbol{A}^T\boldsymbol{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma})$$

where $\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \boldsymbol{A}^T\boldsymbol{L}\boldsymbol{A})^{-1}$.





# Bayesian Regression

- Choose a Gaussian prior with mean $\mathbf{m}_0$ and covariance $\mathbf{S}_0$

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w} \,|\, \mathbf{m}_0, \mathbf{S}_0)$$

- After having seen $N$ training data pairs $(\mathbf{x}_n, t_n)$, the posterior for the given likelihood is now

$$p(\mathbf{w} \,|\, \mathbf{t}) = \mathcal{N}(\mathbf{w} \,|\, \mathbf{m}_N, \mathbf{S}_N)$$

  where

$$\begin{aligned} \mathbf{m}_N &= \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta \mathbf{\Phi}^T \mathbf{t}) \\ \mathbf{S}_N^{-1} &= \mathbf{S}_0^{-1} + \beta \mathbf{\Phi}^T \mathbf{\Phi} \end{aligned}$$

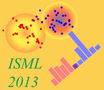
- The posterior is Gaussian, therefore mode $=$ mean.
- The maximum posterior weight vector $\mathbf{w}_{MAP} = \mathbf{m}_N$.
- Assume infinitely broad prior $\mathbf{S}_0 = \alpha^{-1}\mathbf{I}$ with $\alpha \to 0$, the mean reduces to the maximum likelihood $\mathbf{w}_{ML}$.

# Bayesian Regression

- If we have not yet seen any data point ( $N = 0$), the posterior is equal to the prior.
- Sequential arrival of data points : Each posterior distribution calculated after the arrival of a data point and target value, acts as the prior distribution for the subsequent data point.
- Nicely fits a sequential learning framework.







# Bayesian Regression

- Special simplified prior in the remainder, $\mathbf{m}_0 = 0$ and $\mathbf{S}_0 = \alpha^{-1}\mathbf{I}$,

$$p(\mathbf{x} \mid \alpha) = \mathcal{N}(\mathbf{x} \mid 0, \alpha^{-1}\mathbf{I}) \tag{1}$$

- The parameters of the posterior distribution $p(\mathbf{w} \mid \mathbf{t}) = \mathcal{N}(\mathbf{w} \mid \mathbf{m}_n, \mathbf{S}_N)$ are now

$$\mathbf{m}_N = \beta \mathbf{S}_N \boldsymbol{\Phi}^T \mathbf{t}$$

$$\mathbf{S}_N^{-1} = \alpha \mathbf{I} + \beta \boldsymbol{\Phi}^T \boldsymbol{\Phi}$$

- For $\alpha \to 0$ we get

$$\mathbf{m}_N \to \mathbf{w}_{ML} = (\boldsymbol{\Phi}^T \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^T \mathbf{t}$$

- Log of posterior is sum of log likelihood and log of prior

$$\ln p(\mathbf{w} \mid \mathbf{t}) = -\frac{\beta}{2}(\mathbf{t} - \boldsymbol{\Phi}\mathbf{w})^T(\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}) - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} + \text{const}$$




Christfried Webers
NICTA
The Australian National University

# Bayesian Regression

- Log of posterior is sum of log likelihood and log of prior

$$\ln p(\mathbf{w} \mid \mathbf{t}) = -\beta \underbrace{\frac{1}{2}(\mathbf{t} - \boldsymbol{\Phi}\mathbf{w})^T(\mathbf{t} - \boldsymbol{\Phi}\mathbf{w})}_{\text{sum-of-squares-error}} - \frac{\alpha}{2} \underbrace{\mathbf{w}^T\mathbf{w}}_{\text{quadr. regulariser}} + \text{const}$$

- Maximising the posterior distribution with respect to $\mathbf{w}$ corresponds to minimising the sum-of-squares error function with the addition of a quadratic regularisation term $\lambda = \alpha/\beta$.

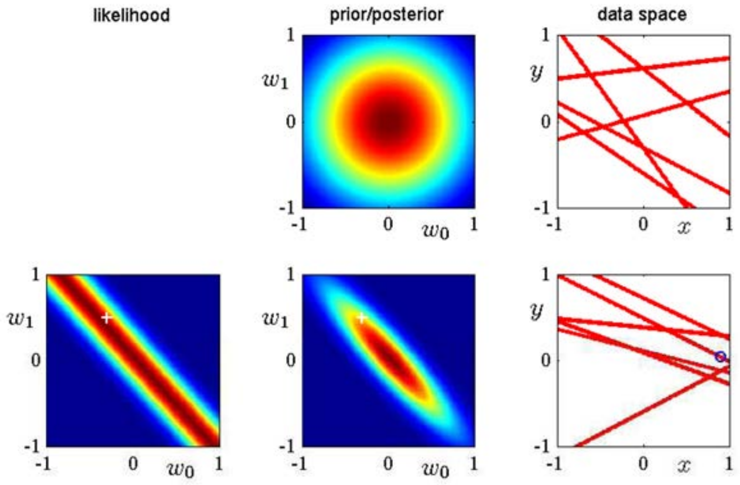
# Sequential Update of the Posterior

- Example of a linear basis function model
- Single input $x$, single output $t$
- Linear model $y(x, \mathbf{w}) = w_0 + w_1 x$.
- Data creation
  1. Choose an $x_n$ from the uniform distribution $\mathcal{U}(x \mid -1, 1)$.
  2. Calculate $f(x_n, \mathbf{a}) = a_0 + a_1 x_n$, where $a_0 = -0.3$, $a_1 = 0.5$.
  3. Add Gaussian noise with standard deviation $\sigma = 0.2$,

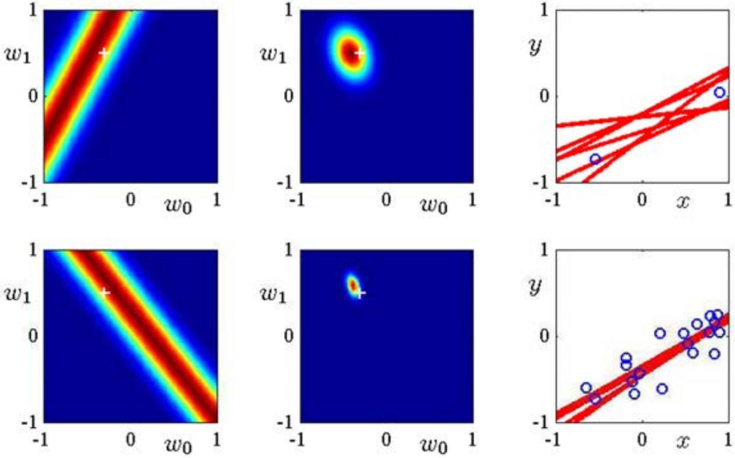
$$t_n = \mathcal{N}(x_n \,|\, f(x_n, \mathbf{a}), 0.04)$$

- Set the precision of the uniform prior to $\alpha = 2.0$.

# Sequential Update of the Posterior



# Sequential Update of the Posterior



# Predictive Distribution

prediction phase: using learned model
produced in the learning phase

- In the training phase, data $\mathbf{x}$ and targets $\mathbf{t}$ are provided
- In the test phase, a new data value $x$ is given and the corresponding target value $t$ is asked for
- Bayesian approach: Find the probability of the test target $t$ given the test data $x$, the training data $\mathbf{x}$ and the training targets $\mathbf{t}$

$$p(t \mid x, \mathbf{x}, \mathbf{t})$$

- This is the Predictive Distribution.

# How to calculate the Predictive Distribution?

- Introduce the model parameter $\mathbf{w}$ via the sum rule

$$p(t \mid x, \mathbf{x}, \mathbf{t}) = \int p(t, \mathbf{w} \mid x, \mathbf{x}, \mathbf{t}) d\mathbf{w}$$

$$= \int p(t \mid \mathbf{w}, x, \mathbf{x}, \mathbf{t}) p(\mathbf{w} \mid x, \mathbf{x}, \mathbf{t}) d\mathbf{w}$$

- The test target $t$ depends only on the test data $x$ and the model parameter $\mathbf{w}$, but not on the training data and the training targets

$$p(t \mid \mathbf{w}, x, \mathbf{x}, \mathbf{t}) = p(t \mid \mathbf{w}, x)$$

- The model parameter $\mathbf{w}$ are learned with the training data $\mathbf{x}$ and the training targets $\mathbf{t}$ only

$$p(\mathbf{w} \mid x, \mathbf{x}, \mathbf{t}) = p(\mathbf{w} \mid \mathbf{x}, \mathbf{t})$$

- Predictive Distribution

$$p(t \mid x, \mathbf{x}, \mathbf{t}) = \int p(t \mid \mathbf{w}, x) p(\mathbf{w} \mid \mathbf{x}, \mathbf{t}) d\mathbf{w}$$

# *Proof of the Predictive Distribution*

- How to prove the Predictive Distribution in the general form?

$$p(t \,|\, x, \mathbf{x}, \mathbf{t}) = \int p(t \,|\, \mathbf{w}, x, \mathbf{x}, \mathbf{t}) p(\mathbf{w} \,|\, x, \mathbf{x}, \mathbf{t}) d\mathbf{w}$$

- Convert each conditional probability on the right-hand-side into a joint probability.

$$\begin{aligned} & \int p(t \,|\, \mathbf{w}, x, \mathbf{x}, \mathbf{t}) p(\mathbf{w} \,|\, x, \mathbf{x}, \mathbf{t}) d\mathbf{w} \\ &= \int \frac{p(t, \mathbf{w}, x, \mathbf{x}, \mathbf{t})}{p(\mathbf{w}, x, \mathbf{x}, \mathbf{t})} \frac{p(\mathbf{w}, x, \mathbf{x}, \mathbf{t})}{p(x, \mathbf{x}, \mathbf{t})} d\mathbf{w} \\ &= \int \frac{p(t, \mathbf{w}, x, \mathbf{x}, \mathbf{t})}{p(x, \mathbf{x}, \mathbf{t})} d\mathbf{w} \\ &= \frac{p(t, x, \mathbf{x}, \mathbf{t})}{p(x, \mathbf{x}, \mathbf{t})} \\ &= p(t \,|\, x, \mathbf{x}, \mathbf{t}) \end{aligned}$$

# Predictive Distribution with Simplified Prior

- Find the predictive distribution

$$p(t \,|\, \mathbf{t}, \alpha, \beta) = \int p(t \,|\, \mathbf{w}, \beta)\, p(\mathbf{w} \,|\, \mathbf{t}, \alpha, \beta) d\mathbf{w}$$

(remember : The conditioning on the input variables $\mathbf{x}$ is often suppressed to simplify the notation.)

- Now we know

$$p(t \,|\, \mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n \,|\, \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1})$$

- and the posterior was

$$p(\mathbf{w} \,|\, \mathbf{t}, \alpha, \beta) = \mathcal{N}(\mathbf{w} \,|\, \mathbf{m}_N, \mathbf{S}_N)$$

where

$$\begin{aligned} \mathbf{m}_N &= \beta \mathbf{S}_N \boldsymbol{\Phi}^T \mathbf{t} \\ \mathbf{S}_N^{-1} &= \alpha \mathbf{I} + \beta \boldsymbol{\Phi}^T \boldsymbol{\Phi} \end{aligned}$$

# Predictive Distribution with Simplified Prior







- If we do the convolution of the two Gaussians, we get for the predictive distribution

$$p(t \mid \mathbf{x}, \mathbf{t}, \alpha, \beta) = \mathcal{N}(t \mid \mathbf{m}_N^T \boldsymbol{\phi}(\mathbf{x}), \sigma_N^2(\mathbf{x}))$$

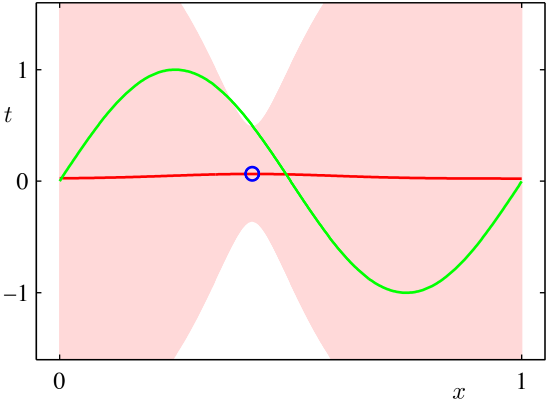
where the variance $\sigma_N^2(\mathbf{x})$ is given by

$$\sigma_N^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^T \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}).$$

# Predictive Distribution with Simplified Prior



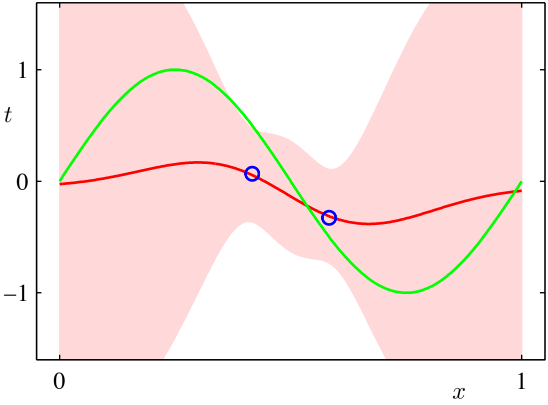
Example with artificial sinusoidal data from $\sin(2\pi x)$ (green) and added noise. Number of data points $N = 1$.




Mean of the predictive distribution (red) and regions of one standard deviation from mean (red shaded).

# Predictive Distribution with Simplified Prior

Example with artificial sinusoidal data from $\sin(2\pi x)$ (green) and added noise. Number of data points $N = 2$.




Mean of the predictive distribution (red) and regions of one standard deviation from mean (red shaded).

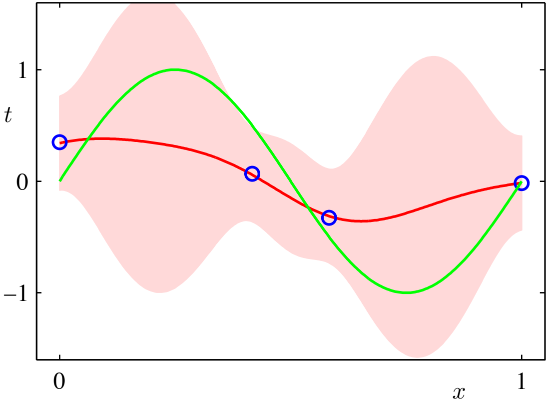
# Predictive Distribution with Simplified Prior







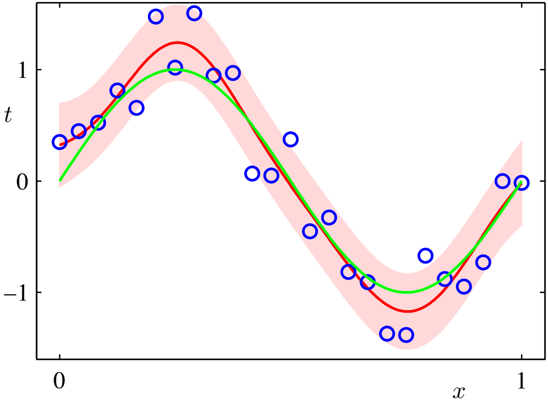
Example with artificial sinusoidal data from $\sin(2\pi x)$ (green) and added noise. Number of data points $N = 4$.




Mean of the predictive distribution (red) and regions of one standard deviation from mean (red shaded).

# Predictive Distribution with Simplified Prior

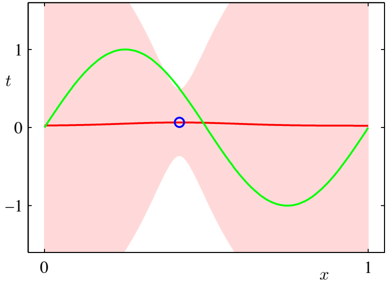
Example with artificial sinusoidal data from $\sin(2\pi x)$ (green) and added noise. Number of data points $N = 25$.

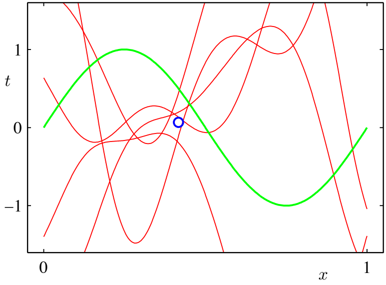



Mean of the predictive distribution (red) and regions of one standard deviation from mean (red shaded).

# Predictive Distribution with Simplified Prior

Plots of the function $y(x, \mathbf{w})$ using samples from the posterior distribution over $\mathbf{w}$. Number of data points $N = 1$.



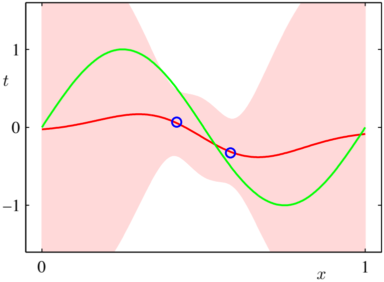



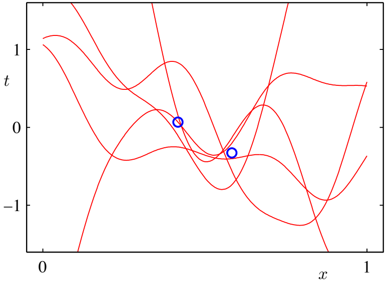
# Predictive Distribution with Simplified Prior





Plots of the function $y(x, \mathbf{w})$ using samples from the posterior distribution over $\mathbf{w}$. Number of data points $N = 2$.



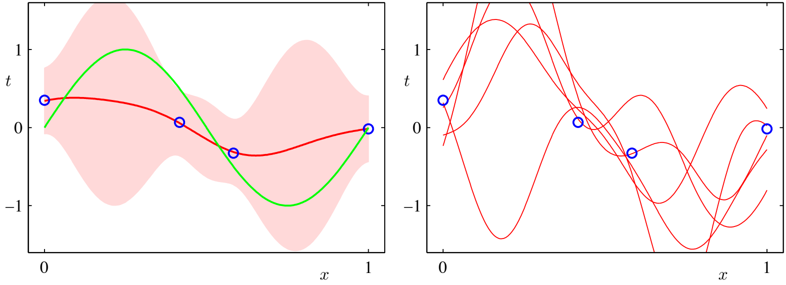



# Predictive Distribution with Simplified Prior





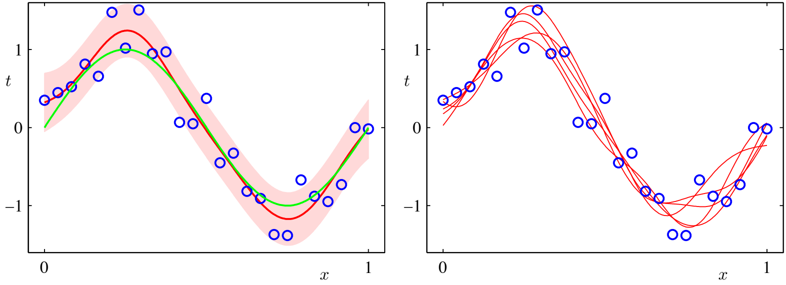
Plots of the function $y(x, \mathbf{w})$ using samples from the posterior distribution over $\mathbf{w}$. Number of data points $N = 4$.



# Predictive Distribution with Simplified Prior



Plots of the function $y(x, \mathbf{w})$ using samples from the posterior distribution over $\mathbf{w}$. Number of data points $N = 25$.



# Limitations of Linear Basis Function Models

- Basis function $\phi_j(\mathbf{x})$ are fixed before the training data set is observed.
- Curse of dimensionality : Number of basis function grows rapidly, often exponentially, with the dimensionality $D$.
- But typical data sets have two nice properties which can be exploited if the basis functions are not fixed :
  - Data lie close to a nonlinear manifold with intrinsic dimension much smaller than $D$. Need algorithms which place basis functions only where data are (e.g. radial basis function networks, support vector machines, relevance vector machines, neural networks).
  - Target variables may only depend on a few significant directions within the data manifold. Need algorithms which can exploit this property (Neural networks).

# *Curse of Dimensionality*

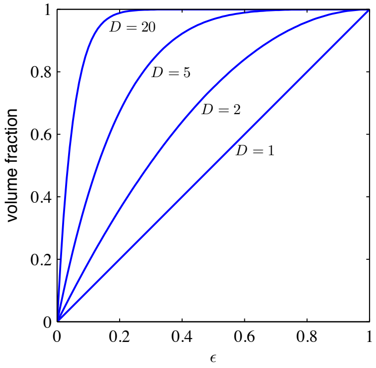
- Linear Algebra allows us to operate in $n$-dimensional vector spaces using the intution from our $3$-dimensional world as a vector space. No surprises as long as $n$ is finite.
- If we add more structure to a vector space (e.g. inner product, metric), our intution gained from the $3$-dimensional world around us may be wrong.
- Example: Sphere of radius $r = 1$. What is the fraction of the volume of the sphere in a $D$-dimensional space which lies between radius $r = 1$ and $r = 1 - \epsilon$ ?
- Volume scales like $r^D$, therefore the formula for the volume of a sphere is $V_D(r) = K_D r^D$.

$$\frac{V_D(1) - V_D(1-\epsilon)}{V_D(1)} = 1 - (1-\epsilon)^D$$

# Curse of Dimensionality

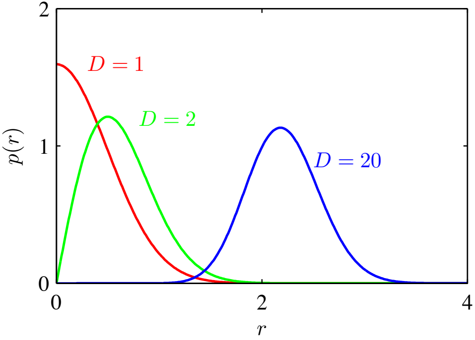
- Fraction of the volume of the sphere in a $D$-dimensional space which lies between radius $r = 1$ and $r = 1 - \epsilon$

$$\frac{V_D(1) - V_D(1-\epsilon)}{V_D(1)} = 1 - (1-\epsilon)^D$$



# Curse of Dimensionality

- Probability density with respect to radius $r$ of a Gaussian distribution for various values of the dimensionality $D$.



# *Curse of Dimensionality*

- Probability density with respect to radius $r$ of a Gaussian distribution for various values of the dimensionality $D$.
- Example: $D = 2$; assume $\mu = 0, \Sigma = I$

$$\mathcal{N}(x \,|\, 0, I) = \frac{1}{2\pi} \exp\left\{ -\frac{1}{2} x^T x \right\} = \frac{1}{2\pi} \exp\left\{ -\frac{1}{2} (x_1^2 + x_2^2) \right\}$$

- Coordinate transformation

$$x_1 = r \cos(\phi) \qquad x_2 = r \sin(\phi)$$

- Probability in the new coordinates

$$p(r, \phi \,|\, 0, I) = \mathcal{N}(r(x), \phi(x) \,|\, 0, I) \,|\, J \,|$$

where $|\, J \,| = r$ is the determinant of the Jacobian for the given coordinate transformation.

$$p(r, \phi \,|\, 0, I) = \frac{1}{2\pi} r \exp\left\{ -\frac{1}{2} r^2 \right\}$$

# Curse of Dimensionality

- Probability density with respect to radius $r$ of a Gaussian distribution for $D = 2$ (and $\mu = 0, \Sigma = I$)

$$p(r, \phi \,|\, 0, I) = \frac{1}{2\pi} r \exp\left\{-\frac{1}{2} r^2\right\}$$

- Integrate over all angles $\phi$

$$p(r \,|\, 0, I) = \int_0^{2\pi} \frac{1}{2\pi} r \exp\left\{-\frac{1}{2} r^2\right\} d\phi = r \exp\left\{-\frac{1}{2} r^2\right\}$$