Bayesian Linear Regression

Sargur Srihari
srihari@cedar.buffalo.edu

Motivation of Bayesian Approach

- We have seen Linear Regression with M basis functions:

$$y(\mathrm{x}, \mathrm{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathrm{x}) = \mathrm{w}^T \phi(\mathrm{x})$$

- The maximum likelihood solution is

$$\mathrm{w}_{ML} = (\Phi^T \Phi)^{-1} \Phi^T \mathrm{t}$$

$$\Phi = \begin{pmatrix} \phi_0(\mathrm{x}_1) & \phi_1(\mathrm{x}_1) & \ldots & \phi_{M-1}(\mathrm{x}_1) \\ \phi_0(\mathrm{x}_2) & & & \\ & & & \\ \phi_0(\mathrm{x}_N) & & & \phi_{M-1}(\mathrm{x}_N) \end{pmatrix}$$

Shortcomings of MLE

- M.L.E. of parameters $\mathbf{w}$ does not address
  - M (Model complexity: how many basis functions?
  - It is controlled by data size N
    - More data allows better fit without overfitting
- Regularization also controls overfit (λ controls its effect) $E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{ t_n - \mathbf{w}^T\phi(\mathbf{x}_n) \}^2$ where $E(\mathbf{w}) = E_D(\mathbf{w}) + \lambda E_W(\mathbf{w})$ $E_W(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w}$
    - But M and choice of ϕ_j are still important
      - M can be determined by holdout, but wasteful of data
- Model complexity and over-fitting better handled using Bayesian approach

Bayesian Approach

- Posterior is proportional to Likelihood x Prior

$$p(\mathrm{w}\,|\,\mathbf{t}) = \frac{p(\mathbf{t}\,|\,\mathrm{w})p(\mathrm{w})}{p(\mathbf{t})}$$

  - where $p(\mathbf{t}|\mathrm{w})$ is the likelihood of observed data
- We will look at:
  - A normal distribution for prior $\mathrm{p}(\mathrm{w})$
  - Likelihood $p(\mathrm{t}|\mathrm{w})$ is a product of Gaussians based on the noise model

Prior Distribution for Parameters

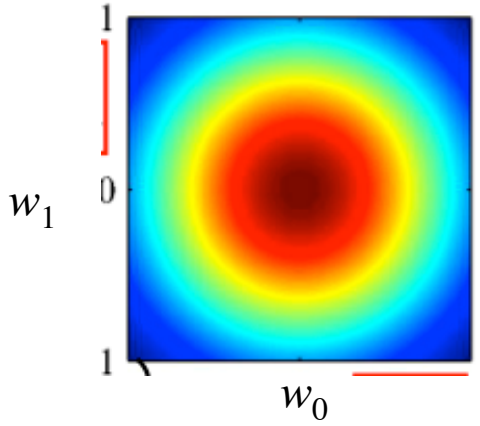
Assume multivariate Gaussian prior for $\mathbf{w}$ (which has components $w_0,..,w_{M-1}$)

$$p(\mathbf{w}) = N(\mathbf{w}|\mathbf{m}_0, S_0)$$

with mean $\mathbf{m}_0$ and covariance matrix S_0

If we choose $S_0 = \alpha^{-1}\mathbf{I}$ it means that the variances of the weights are all equal to α^{-1} and covariances are zero

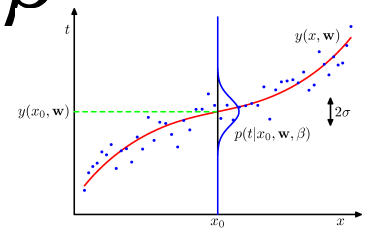
$p(\mathbf{w})$ with zero mean ($\mathbf{m}_0=\mathbf{0}$)
and isotropic over weights (*same variances*)



Likelihood of Data

- Assume noise precision parameter β

$$t = y(\mathrm{x},\mathrm{w})+\varepsilon$$
$$p(t|\mathrm{x},\mathrm{w},\beta)=N(t|y(\mathrm{x},\mathrm{w}),\beta^{-1})$$



- Likelihood $p(\mathbf{t}|\mathrm{w})$ with Gaussian noise has an exponential form

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

This is the probability of data t given the parameters w and input X

Posterior Distribution of Parameters

- Posterior is proportional to Likelihood x Prior

$$p(\mathrm{w}\,|\,\mathrm{t}) = \frac{p(\mathrm{t}\,|\,\mathrm{w})p(\mathrm{w})}{p(\mathrm{t})}$$

  - where likelihood function is

$$p(\mathrm{t}\,|\,X,\mathrm{w},\beta) = \prod_{n=1}^{N} N\left(t_n\,|\,\mathrm{w}^T\phi(\mathrm{x}_n),\beta^{-1}\right)$$

Product of Gaussians

- Multiplying by Gaussian prior, $p(\mathrm{w})=N(\mathrm{w}|\mathrm{m}_0,S_0)$ posterior is also Gaussian, written directly as

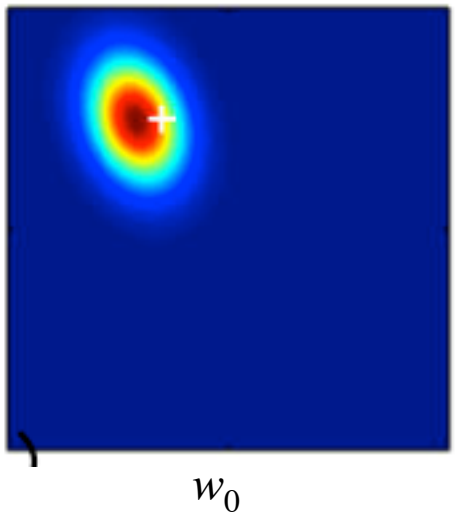
$$p(\mathrm{w}|\mathrm{t})=N(\mathrm{w}|\mathrm{m}_N,S_N)$$

  - Where m_N is the mean of the posterior

    given by $\mathrm{m}_N = S_N(S_0^{-1}\mathrm{m}_0 + \beta\Phi^T\mathrm{t})$

  - And S_N is the covariance matrix of posterior

    given by $S_N^{-1} = S_0^{-1} + \beta\Phi^T\Phi$



Properties of Posterior

1. Since posterior $p(\mathrm{w}|\mathrm{t})=N(\mathrm{w}|\mathrm{m}_N,S_N)$ is Gaussian its mode coincides with its mean
   - Thus maximum posterior weight vector is $\mathrm{w}_{\mathrm{MAP}}=\mathrm{m}_{\mathrm{N}}$
2. For infinitely broad prior $S_0=\alpha^{-1}\mathrm{I}$ with $\alpha->0$
   - Then mean m_{N} reduces to the maximum likelihood value

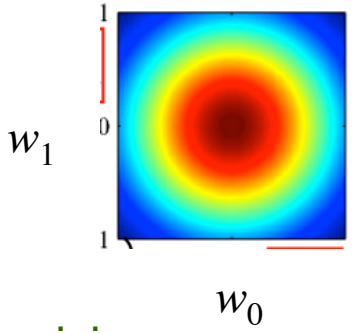
$$\mathrm{w}_{ML}=(\Phi^T\Phi)^{-1}\Phi^T\mathrm{t}$$

3. If $N=0,$ posterior reverts to the prior
4. If data points arrive sequentially, then posterior to any stage acts as prior distribution for subsequent data points

Choose a simple Gaussian prior $p(\mathrm{w})$

- Zero mean ($\mathrm{m}_0=\mathbf{0}$) isotropic
- (*same variances*) Gaussian

$$p(\mathrm{w}\,|\,\alpha)=N(\mathrm{w}\,|\,\mathbf{0},\alpha^{-1}I)$$



Single precision parameter

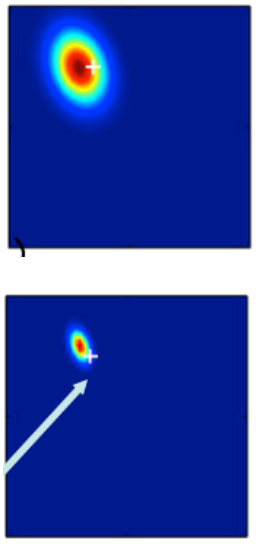
- Corresponding posterior distribution is

$$p(\mathrm{w}|\mathrm{t})=N(\mathrm{w}|\mathrm{m}_N,S_N)$$

where

$$\mathrm{m}_N=\beta S_N\Phi^T\mathrm{t} \quad \text{and} \quad S_N^{-1}=\alpha I+\beta\Phi^T\Phi$$

Note:
β is noise precision and
α is variance of parameter w in prior

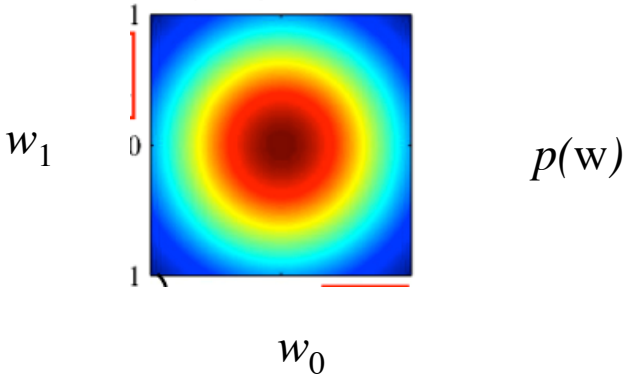


Point Estimate with infinite samples

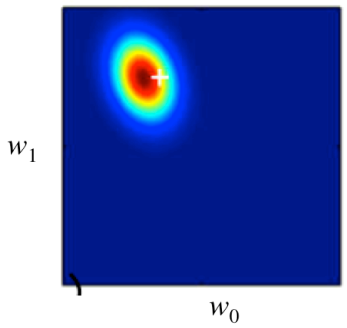
Sampling $p(\mathrm{w})$ and $p(\mathrm{w/t})$

- Each sample represents a straight line in data space

Distribution

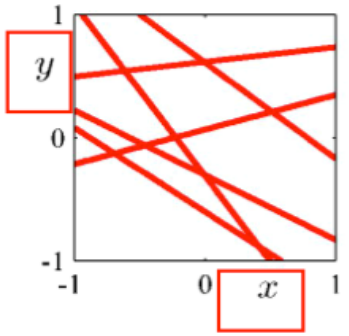


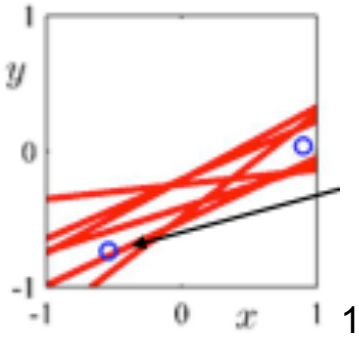
$p(\mathrm{w})$



$p(\mathrm{w|t})$

Six samples





Goal of Bayesian Linear Regression:
Determine $p(\mathrm{w/t})$

Interesting Relationship to MLE

- Since $p(\mathrm{w}\,|\,\mathrm{t}) = \prod_{n=1}^{N} N\left(t_n \,|\, \mathrm{w}^T \phi(\mathrm{x}_n), \beta^{-1}\right) N(\mathrm{w}\,|\,0, \alpha^{-1}\mathrm{I})$
- Log of Posterior is

$$\ln p(\mathrm{w}\,|\,\mathrm{t}) = -\frac{\beta}{2}\sum_{n=1}^{N}\left\{t_n - \mathrm{w}^T\phi(\mathrm{x}_n)\right\}^2 - \frac{\alpha}{2}\mathrm{w}^T\mathrm{w} + const$$

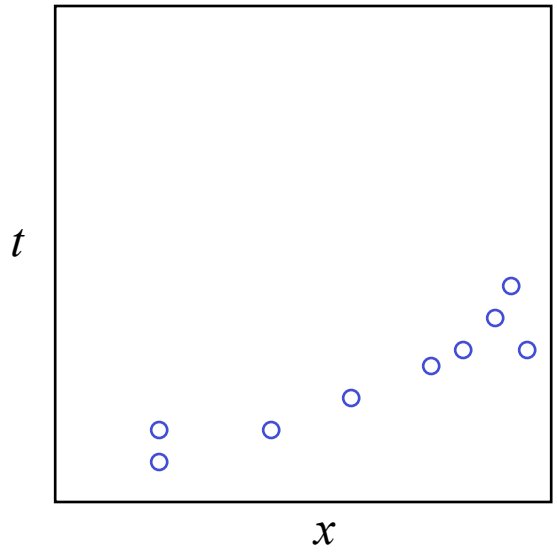
- Thus Maximization of posterior distribution wrt w is equivalent to minimization of sum-of-squares error function

$$E(\mathrm{w}) = \frac{1}{2}\sum_{n=1}^{N}\left\{\, t_n - \mathrm{w}^T\phi(x_n)\,\right\}^2 + \frac{\lambda}{2}\mathrm{w}^T\mathrm{w}$$

with addition of quadratic regularization term $\mathrm{w}^{\mathrm{T}}\mathrm{w}$ with $\lambda = \alpha/\beta$

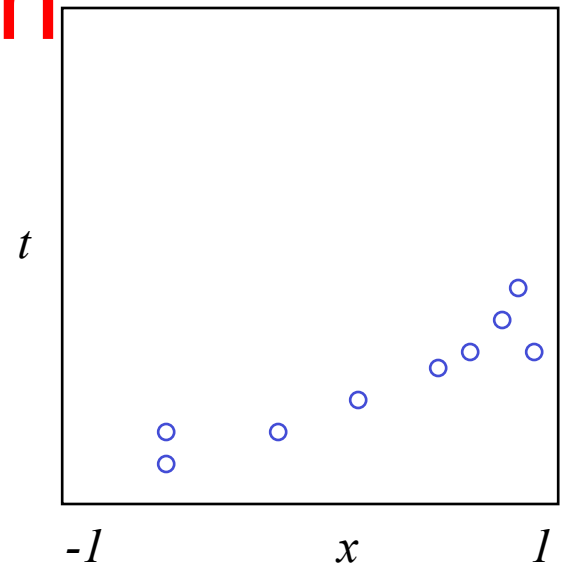
Bayesian Linear Regression Example (Straight Line Fit)

- Single input variable x
- Single target variable t
- Goal is to fit
  - Linear model $y(x,\mathrm{w}) = w_0 + w_1 x$
- Goal of Linear Regression is to recover $\mathrm{w}=[w_0, w_1]$ given the samples



Data Generation

- Synthetic data generated from $f(x,\mathrm{w})=w_0+w_1x$ with parameter values $w_0=-0.3$ and $w_1=0.5$



  - First choose x_n from $U(x|-1,1)$, Then evaluate $f(x_n,\mathrm{w})$
  - Add Gaussian noise with st dev 0.2 to obtain target t_n
    - Precision parameter $\beta=(1/0.2)^2=25$

- For prior over w we choose $\alpha=2$

$$p(\mathrm{w}\,|\,\alpha)=N(\mathrm{w}\,|\,\mathbf{0},\alpha^{-1}I)$$

Sequential Bayesian Learning

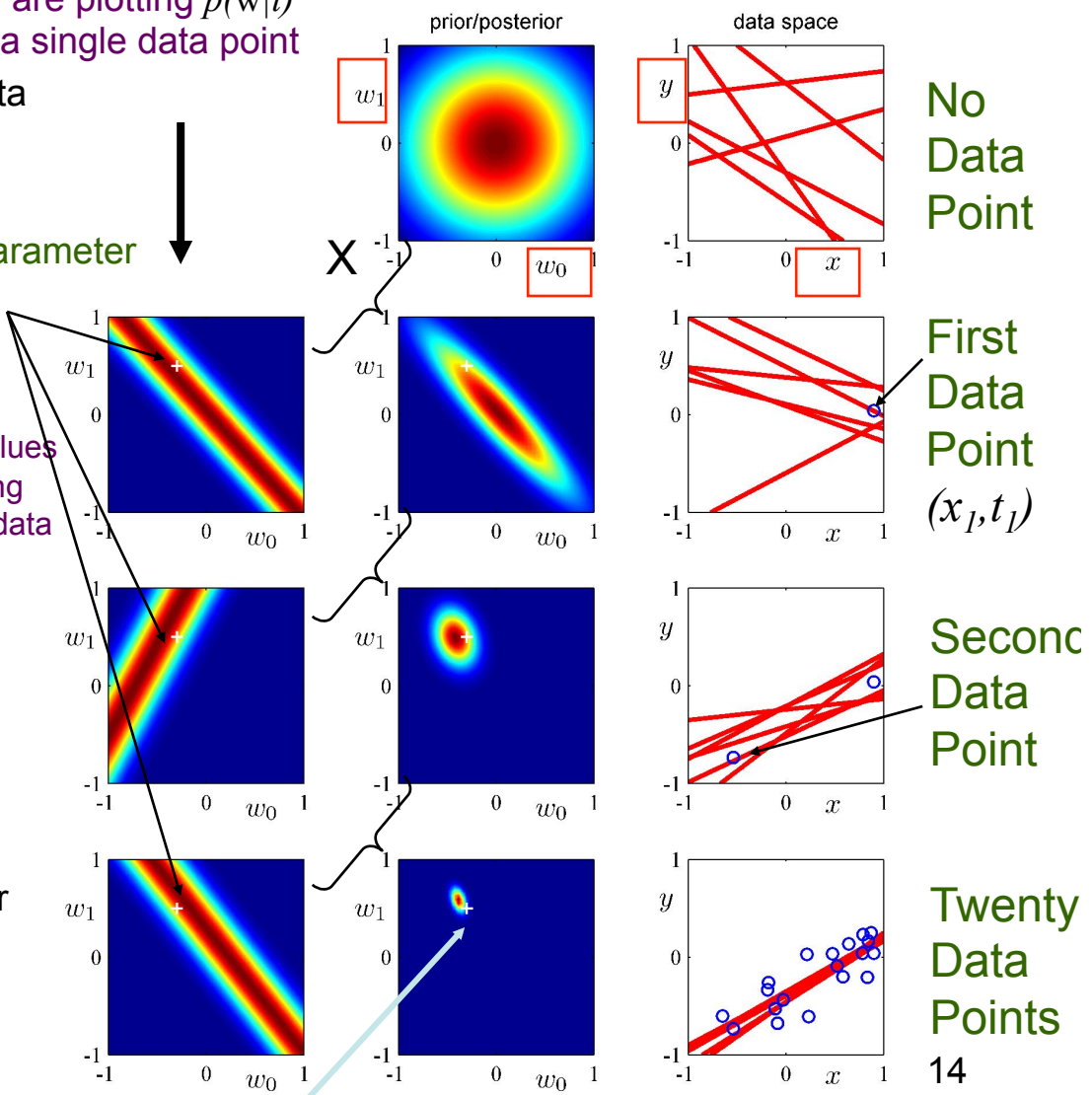
- Since there are only two parameters
  - We can plot prior and posterior distributions in parameter space
- We look at sequential update of posterior

Likelihood $p(t|x.\mathrm{w})$ as function of w

Prior/ Posterior $p(\mathrm{w})$ gives $p(\mathrm{w}|t)$

Six samples (regression functions) corresponding to $y(x,\mathrm{w})$ with w drawn from posterior

We are plotting $p(\mathrm{w}|t)$ for a single data point



With infinite points posterior is a delta function centered at true parameters (white cross)

Generalization of Gaussian prior

- The Gaussian prior over parameters is

$$p(\mathrm{w}\,|\,\alpha) = N(\mathrm{w}\,|\,\mathbf{0}, \alpha^{-1} I)$$

Maximization of posterior ln $p(\mathrm{w}|\mathbf{t})$ is equivalent to minimization of sum of squares error

$$E(\mathrm{w}) = \frac{1}{2}\sum_{n=1}^{N}\left\{ t_n - \mathrm{w}^T \phi(x_n) \right\}^2 + \frac{\lambda}{2}\mathrm{w}^T\mathrm{w}$$

- Generalization of Gaussian prior

$$p(\mathrm{w}\,|\,\alpha) = \left[\frac{q}{2}\left(\frac{\alpha}{2}\right)^{1/q}\frac{1}{\Gamma(1/q)}\right]^M \exp\left(-\frac{\alpha}{2}\sum_{j=1}^{M}|\,w_j\,|^q\right)$$

  - $q=2$ corresponds to Gaussian
  - Corresponds to minimization of regularized error function

$$\frac{1}{2}\sum_{n=1}^{N}\left\{ t_n - \mathrm{w}^T \phi(\mathrm{x}_n) \right\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|\,w_j\,|^q$$

Predictive Distribution

- Usually not interested in the value of $\mathbf{w}$ itself
- But predicting t for a new value of $\mathbf{x}$

$$p(t|\mathbf{t},\mathbf{X},\mathbf{x}) \text{ or}$$

$$p(t|\mathbf{t})$$

  - Leaving out conditioning variables $\mathbf{X}$ and $\mathbf{x}$ for convenience
- Marginalizing over parameter variable $\mathbf{w}$, is the standard Bayesian approach
  - From sum rule

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

Predictive Distribution with α, β

- We can predict t for a new value of x using

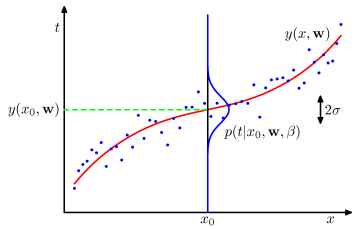
$$p(t \mid \mathbf{t}) = \int p(t|\mathrm{w})\, p(\mathrm{w}|\mathbf{t})\, d\mathrm{w}$$

- Introducing α characterizing prior, and β characterizing noise

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

conditional distribution of target variable

$$p(t \mid \mathrm{x},\mathrm{w},\beta) = N(t \mid y(\mathrm{x},\mathrm{w}),\beta^{-1})$$



posterior weight distribution

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

$$\mathrm{m}_N = \beta S_N \boldsymbol{\Phi}^T \mathrm{t}$$
$$S_N^{-1} = \alpha I + \beta \boldsymbol{\Phi}^T \boldsymbol{\Phi}$$

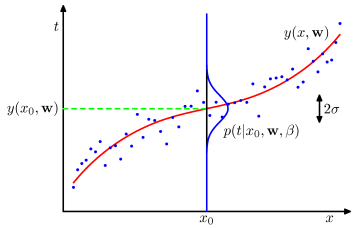
- RHS is convolution of two Gaussian distributions
  - whose result is a Gaussian

Variance of Predictive Distribution

- Predictive distribution takes the form

$$p(t \mid \mathrm{x}, \mathbf{t}, \alpha, \beta) = N(t \mid m_N^T \varphi(\mathrm{x}), \sigma_N^2(\mathrm{x}))$$

$$\text{where} \quad \sigma_N^2(\mathrm{x}) = \underbrace{\frac{1}{\beta}}_{\text{Noise in data}} + \underbrace{\varphi(\mathrm{x})^T S_N \varphi(\mathrm{x})}_{\text{Uncertainty associated with parameters w}}$$



$$S_N^{-1} = \alpha I + \beta \Phi^T \Phi$$

Properties of Variance of Predictive Distribution

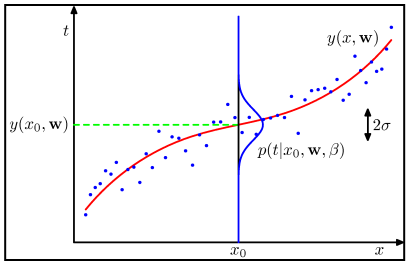
$$\sigma_N^2(\mathrm{x}) = \frac{1}{\beta} + \varphi(\mathrm{x})^T S_N \varphi(\mathrm{x})$$

- It can be shown that $\sigma_{N+1}^2(\mathrm{x}) \leq \sigma_N^2(\mathrm{x})$
- In the limit as $N \to \infty$,
  - The second term goes to zero
  - Variance is solely due to additive noise

Example of Predictive Distribution

- Data generated from $sin(2\pi x)$
- Model: nine Gaussian basis fns

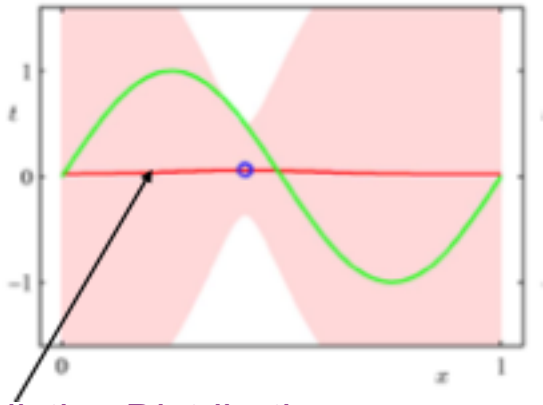
$$y(\mathrm{x},\mathrm{w}) = \sum_{j=0}^{8} w_j \phi_j(\mathrm{x}) = \mathrm{w}^T \phi(\mathrm{x})$$



- Predictive distribution

$$p(t \mid \mathrm{x}, \mathbf{t}, \alpha, \beta) = N(t \mid m_N^T \varphi(\mathrm{x}), \sigma_N^2(\mathrm{x})) \quad \text{where} \quad \sigma_N^2(\mathrm{x}) = \frac{1}{\beta} + \varphi(\mathrm{x})^T S_N \varphi(\mathrm{x})$$

Plot of $p(t|x)$
for one data point
showing mean (red)
and one std dev (pink)



Mean of Predictive Distribution

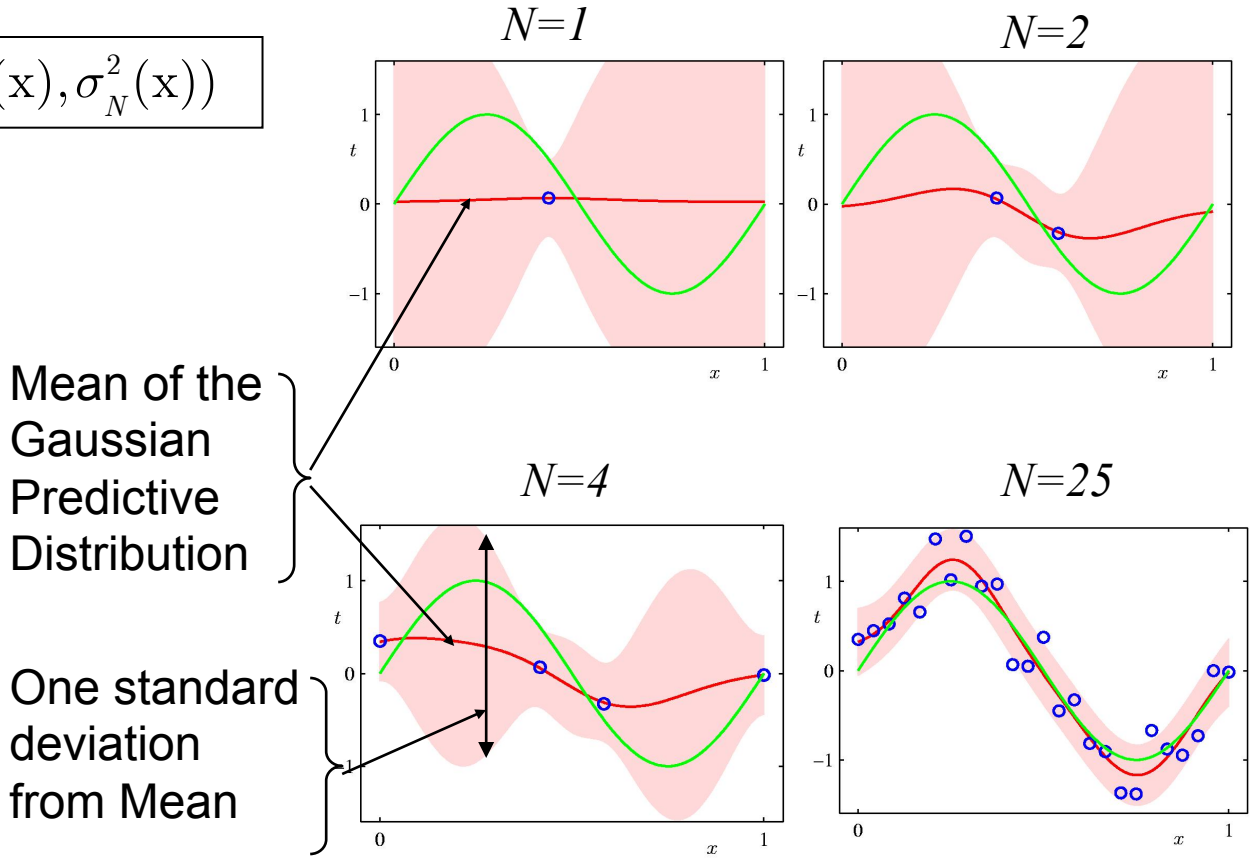
$$\mathrm{m}_N = \beta S_N \Phi^T \mathrm{t}$$
$$S_N^{-1} = \alpha I + \beta \Phi^T \Phi$$

$$\Phi = \begin{pmatrix} \phi_0(\mathrm{x}_1) & \phi_1(\mathrm{x}_1) & \ldots & \phi_{M-1}(\mathrm{x}_1) \\ \phi_0(\mathrm{x}_2) & & & \\ & & & \\ \phi_0(\mathrm{x}_N) & & & \phi_{M-1}(\mathrm{x}_N) \end{pmatrix}$$

Predictive Distribution (Sinusoidal Data)

$$p(t \mid \mathrm{x}, \mathbf{t}, \alpha, \beta) = N(t \mid m_N^T \varphi(\mathrm{x}), \sigma_N^2(\mathrm{x}))$$

- Std dev of t is smallest in neighborhood of data points
- Uncertainty decreases as more data points are observed



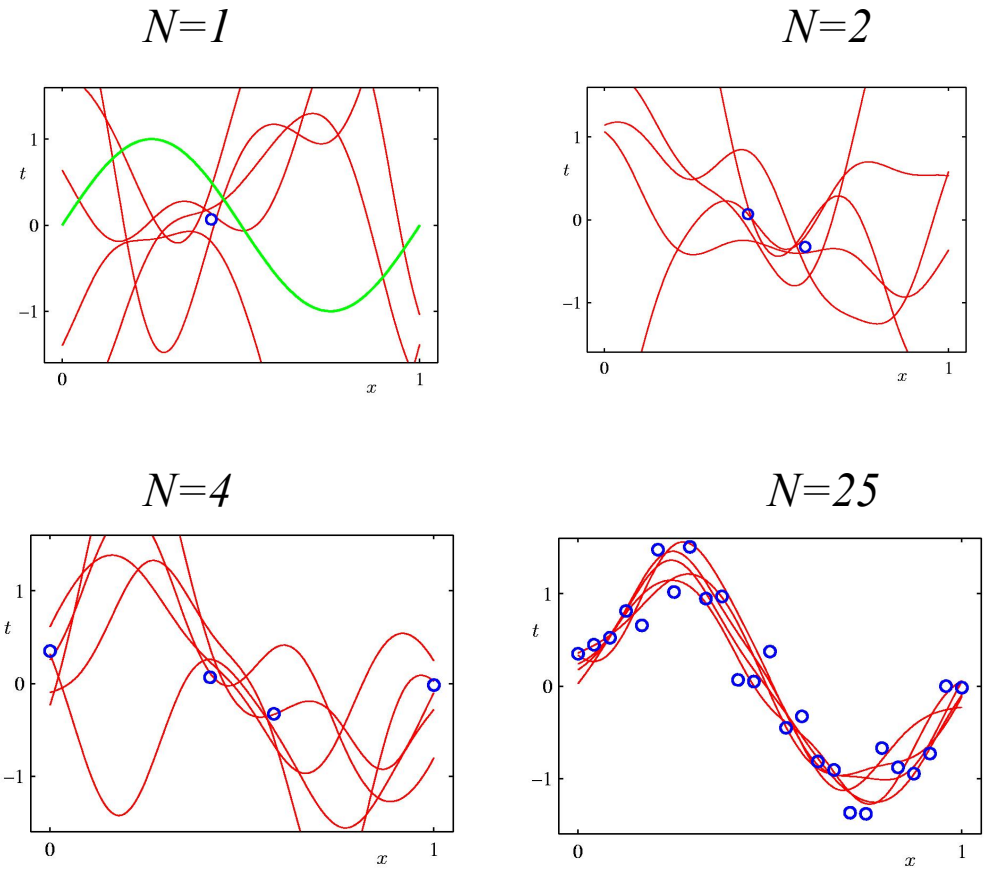
Plot only shows point-wise predictive variance
To show covariance between predictions at different values of x draw samples from posterior distribution over w $p(\mathrm{w}|t)$ and draw samples of $y(x,\mathrm{w})$

Plots of function $y(x,\mathrm{w})$

- Draw samples from from posterior distribution $p(\mathrm{w}/\mathrm{t})$

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

- Plots of samples from $y(x,\mathrm{w})$ corresponding to predictive distributions in previous slide



These curves represent distribution of the regression function

Kernel Interpretation

- Regression function is:

$$y(\mathrm{x},\mathrm{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathrm{x}) = \mathrm{w}^T \phi(\mathrm{x})$$

- Bayesian Soln. with Gaussian prior $p(\mathrm{w}) = N(\mathrm{w}|\mathrm{m}_0, S_0)$
  - Posterior $p(\mathrm{w}|\mathrm{t}) = N(\mathrm{w}|\mathrm{m}_N, S_N)$ where

$$\mathrm{m}_N = S_N (S_0^{-1}\mathrm{m}_0 + \beta \Phi^T \mathrm{t})$$

$$S_N^{-1} = S_0^{-1} + \beta \Phi^T \Phi$$

  - With zero mean isotropic $p(\mathrm{w}|\alpha) = N(\mathrm{w}|0, \alpha^{-1} I)$

$$\mathrm{m}_N = \beta S_N \Phi^T \mathrm{t},$$

$$S_N^{-1} = \alpha I + \beta \Phi^T \Phi$$

- Posterior mean $\beta S_N \Phi^T \mathrm{t}$ has a kernel interpretation
  - Sets stage for kernel methods and Gaussian processes

Equivalent Kernel

- Posterior mean of w is $\mathrm{m}_N = \beta S_N \Phi^T \mathrm{t}$
  - where S_N is the design matrix
- Substitute mean value into Regression function

$$y(\mathrm{x}, \mathrm{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathrm{x}) = \mathrm{w}^T \phi(\mathrm{x})$$

- Mean of predictive distribution at point x

$$\begin{aligned} y(\mathrm{x}, \mathrm{m}_N) &= \mathrm{m}_N^T \phi(\mathrm{x}) = \beta \phi(\mathrm{x})^T S_N \Phi^T \mathrm{t} \\ &= \sum_{n=1}^{N} \beta \phi(\mathrm{x})^T S_N \phi(\mathrm{x}_n) t_n \\ &= \sum_{n=1}^{N} k(\mathrm{x}, \mathrm{x}_n) t_n \end{aligned}$$

  - Where $k(\mathrm{x}, \mathrm{x}') = \beta \phi(\mathrm{x})^T S_N \phi(\mathrm{x}')$ is the *equivalent kernel*
- Mean of predictive distribution is a linear combination of training set target variables

Kernel Function

- Regression functions such as

$$y(\mathrm{x}, \mathrm{m}_N) = \sum_{n=1}^{N} k(\mathrm{x}, \mathrm{x}_n) t_n$$

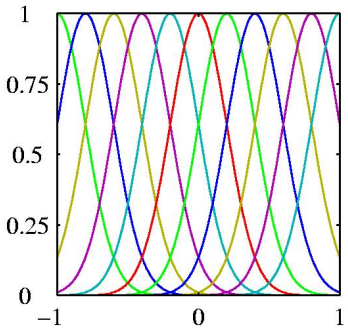
$$k(\mathrm{x},\mathrm{x}') = \beta \phi(\mathrm{x})^T S_N \phi(\mathrm{x}')$$

- That take a linear combination of the training set target values are known as *linear smoothers*
- They depend on the input values x_n from the data set since they appear in the definition of S_N

Equivalent kernel for Gaussian Basis

Gaussian Basis $\phi(\mathrm{x})$

$$\phi_j(x) = \exp\left(\frac{(x-\mu_j)^2}{2s^2}\right)$$

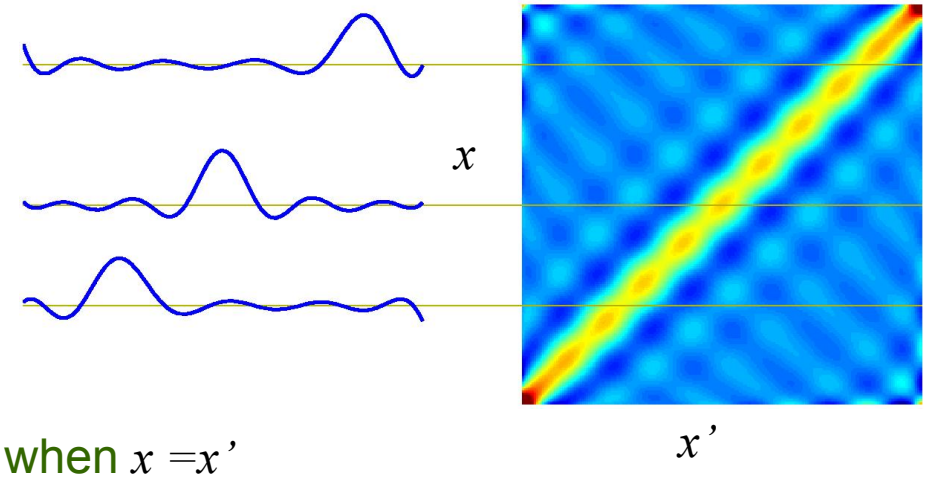


Equivalent Kernel

$$k(\mathrm{x},\mathrm{x}') = \phi(\mathrm{x})^T S_N \phi(\mathrm{x}')$$

Plot of $k(x,x')$ shown as a function of x and x'
Peaks when $x=x'$

For three values of x the behavior of $k(x,x')$ is shown as a slice



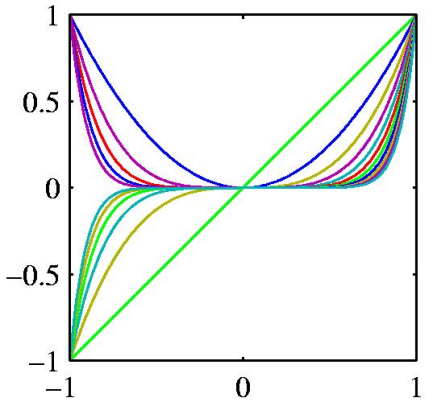
They are localized around x, i.e., peaks when $x = x'$
Local evidence is weighted more than distant evidence

Kernel used directly in regression

$$y(\mathrm{x},\mathrm{m}_N) = \sum_{n=1}^{N} k(\mathrm{x},\mathrm{x}_n) t_n$$

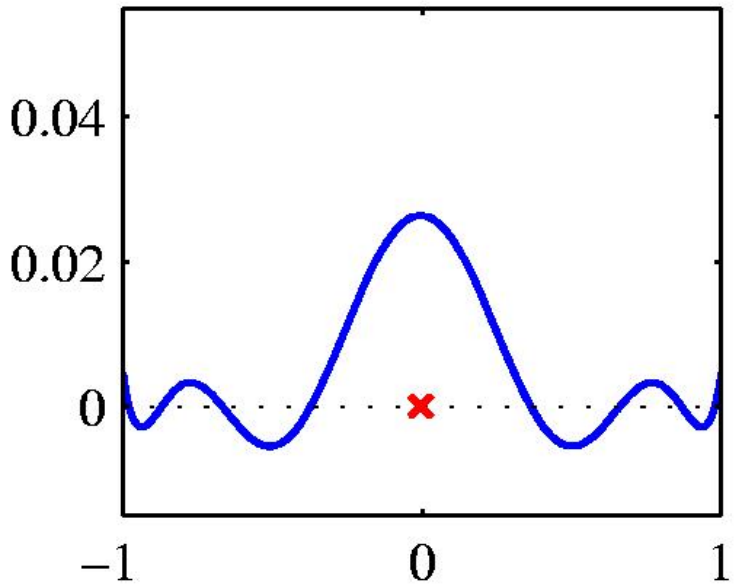
Data set used to generate kernel were 200 values of x equally spaced in (-1,1)

Equivalent Kernel for Polynomial Basis Function



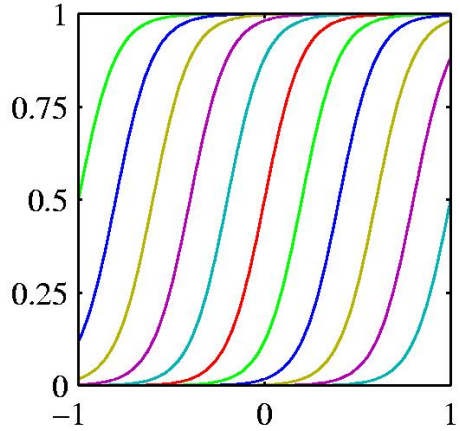
$$\phi_j(x) = x^j$$

$$k(\mathrm{x},\mathrm{x}') = \beta\phi(\mathrm{x})^T S_N \phi(\mathrm{x}')$$



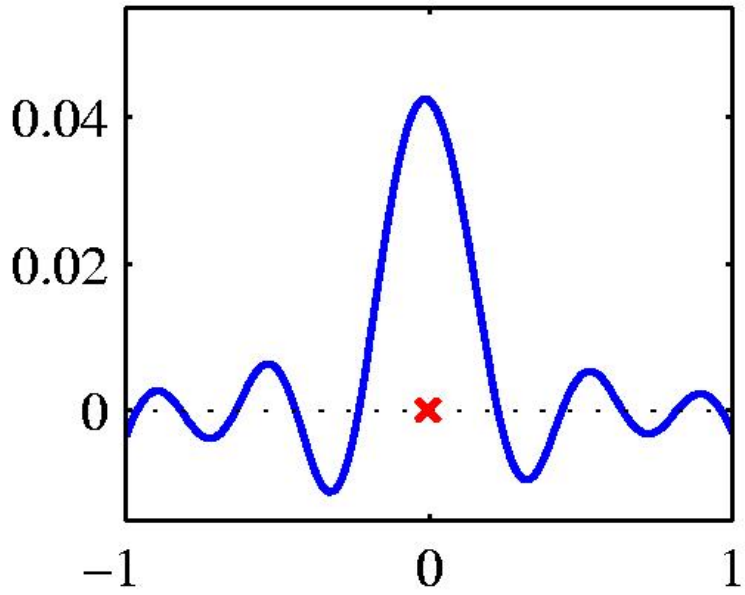
Localized function of x' even though corresponding basis function is nonlocal

Equivalent Kernel for Sigmoidal Basis Function



$$\varphi_j(x) = \sigma\left(\frac{x-\mu_j}{s}\right) \quad \text{where} \quad \sigma(a) = \frac{1}{1+\exp(-a)}$$

$$k(\mathrm{x},\mathrm{x}') = \beta\phi(\mathrm{x})^T S_N \phi(\mathrm{x}')$$



Localized function of x' even though corresponding basis function is nonlocal

Covariance between $y(\mathrm{x})$ and $y(\mathrm{x}')$

$$\mathrm{cov}\,[y(\mathrm{x}), y(\mathrm{x}')] = \mathrm{cov}[\phi(\mathrm{x})^{\mathrm{T}}\mathrm{w}, \mathrm{w}^{\mathrm{T}}\phi(\mathrm{x}')]$$
$$= \phi(\mathrm{x})^{\mathrm{T}}\mathrm{S}_{\mathrm{N}}\phi(\mathrm{x}')$$
$$= \beta^{-1}k(\mathrm{x},\mathrm{x}')$$

The kernel captures the covariance

From the form of the equivalent kernel,

the predictive mean at nearby points $y(\mathrm{x})$, $y(\mathrm{x}')$ will be highly correlated

For more distant pairs correlation is smaller

Kernel Function formulation

- Formulation of Linear Regression in terms of kernel function suggests:
  - Directly define kernel functions and use to make predictions for input x
- Leads to *Gaussian Processes* for Regression and Classification
- Can be shown that $\sum_{n=1}^{N} k(\mathrm{x},\mathrm{x}_n) = 1$
- Equivalent kernel satisfies important property:

$$k(\mathrm{x},\mathrm{z}) = \psi(\mathrm{x})^T \psi(\mathrm{z})$$

  - where $\psi(\mathrm{x}) = \beta^{1/2} S_N^{1/2} \phi(\mathrm{x})$

Probabilistic Linear Regression Revisited

- Re-derive the predictive distribution by working in terms of distribution over functions $y(\mathrm{x},\mathrm{w})$
  - It will provide a specific example of a Gaussian Process
- Consider model with M fixed basis functions

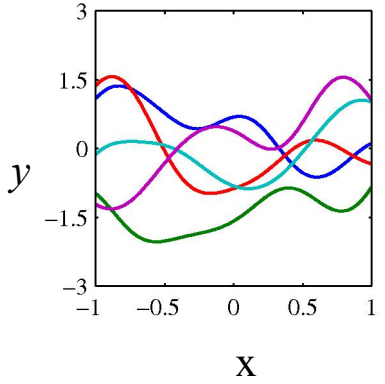
$$y(\mathrm{x})=\mathrm{w}^{\mathrm{T}}\phi(\mathrm{x})$$

where x is the input vector and w is the M-dimensional weight vector

- Assume a Gaussian distribution of weight w

$$p(\mathrm{w}) = \mathcal{N}(\mathrm{w}|0,\alpha^{-1}\mathrm{I})$$

- Probability distribution over w induces a probability distribution over functions $y(\mathrm{x})$



Probabilistic Linear Regression is a GP

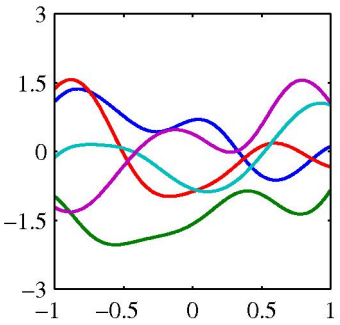
- We wish to evaluate $y(\mathrm{x})$ at training points $\mathrm{x}_1,..,\mathrm{x}_N$
- Our interest is the joint distribution of values $\mathbf{y}=[y(\mathrm{x}_1),..y(\mathrm{x}_N)]$ Since $y(\mathrm{x})=\mathrm{w}^{\mathrm{T}}\phi(\mathrm{x})$ we can write $\mathbf{y}=\Phi\mathrm{w}$
  - where Φ is the design matrix with elements $\Phi_{nk}=\phi_k(\mathrm{x}_n)$

N x M times M x 1 yields a N x 1

- Since y is a linear combination of elements of w which are Gaussian distributed as $p(\mathrm{w})=\mathcal{N}(\mathrm{w}|0,\alpha^{-1}\mathrm{I})$

Covariance expressed by Kernel function

y is itself Gaussian with



mean: $\mathrm{E}[\mathbf{y}]=\Phi\mathrm{E}[\mathrm{w}]=0$ and

variance: $\mathrm{Cov}[\mathbf{y}]=\mathrm{E}[\mathbf{y}\mathbf{y}^{\mathrm{T}}]=\Phi\mathrm{E}[\mathrm{w}\mathrm{w}^{\mathrm{T}}]\Phi^{\mathrm{T}}=(1/\alpha)\Phi\Phi^{\mathrm{T}}=\mathrm{K}$

where K is the N x N Gram Matrix with elements

$$\mathrm{K}_{nm}=k(\mathrm{x}_n,\mathrm{x}_m)=(1/\alpha)\phi(\mathrm{x}_n)^{\mathrm{T}}\phi(\mathrm{x}_m)$$

and $k(\mathrm{x},\mathrm{x}')$ is the kernel function

Thus Gaussian distributed weight vector induces a Gaussian joint distribution over training samples

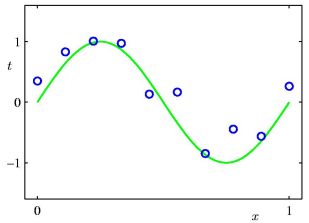
General definition of GP

- We saw a particular example of a Gaussian process
  - Linear regression using the parametric model $y(\mathbf{x})=\mathbf{w}^{\mathsf{T}}\boldsymbol{\phi}(\mathbf{x})$
  - With samples $\mathbf{y}= [y(\mathbf{x}_1),..y(\mathbf{x}_N)]$
  - Assume $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|0,\alpha^{-1}\mathbf{I})$
  - Then $\mathrm{E}[\mathbf{y}] =0$ and $\mathrm{Cov}[\mathbf{y}]=\mathbf{K}$, the Gram matrix which is equivalent to pairwise kernel values
- More generally, a Gaussian process is a probability distribution over functions $y(\mathbf{x})$
  - Such that the set of values of $y(\mathbf{x})$ evaluated at arbitrary points $\mathbf{x}_1,..,\mathbf{x}_N$ jointly have a multivariate Gaussian distribution
    - For a single input $\mathbf{x}_1$, output y_1 is univariate Gaussian. For two inputs $\mathbf{x}_1, \mathbf{x}_2$ the output y_1, y_2 is bivariate Gaussian, etc

Stochastic Process

- A *stochastic process* $y(\mathbf{x})$ is specified by giving the joint probability distribution for any finite set of values $y(\mathbf{x}_1),\ldots,y(\mathbf{x}_N)$ in a consistent manner
  - The random variables typically develop over time
- A Gaussian process is a stochastic process which is Gaussian
- When input $\mathbf{x}$ is 2-D it is also known as a *Gaussian Random Field*

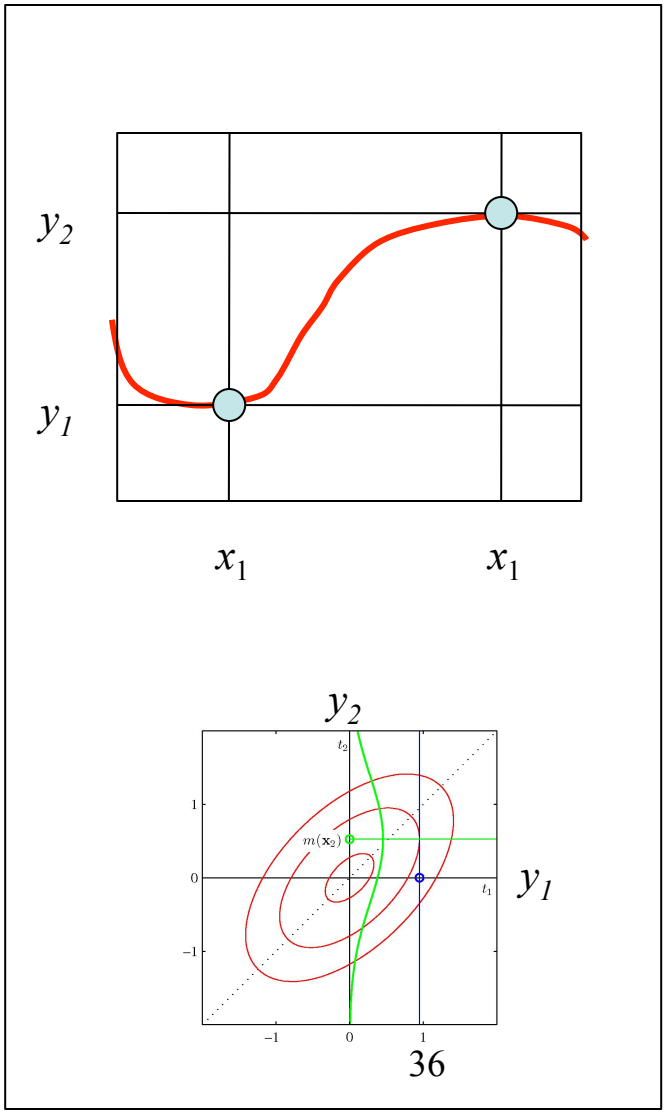
Parametric Regression vs GP Regression



- In parametric regression we have several samples from which we learn the parameters
- In Gaussian processes we we view the samples as one huge input that has a Gaussian distribution
  - with a mean and a covariance matrix
- A **Gaussian process** is a stochastic process X_t, $t \in T$, for which any finite linear combination of samples has a joint Gaussian Distribution
  - any linear functional applied to the sample function X_t will give a normally distributed result. Notation-wise, one can write $X \sim \mathrm{GP}(m,K)$, meaning the random function X is distributed as a GP with mean function m and covariance function K.

Gaussian Process with Two Samples

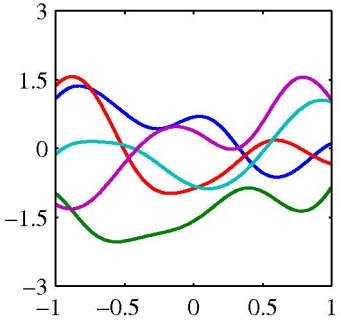
- Let y be a function (curve)
  - of a one-dimensional variable x
- We take two samples y_1 and y_2 corresponding to x_1 and x_2
- Assume they have a bivariate Gaussian distribution $p(y_1,y_2)$
- Each point from this distribution
  - has an associated probability
  - It also defines a function $y(x)$
    - Assuming that two points are enough to define a curve
- More than two points will be needed to define a curve
  - Which leads to a higher dimensional probability distribution



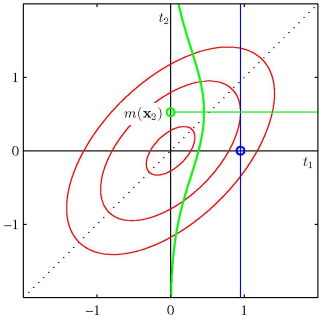
Gaussian Process with N samples

- A process that generates samples over time $\{y_1,..y_N\}$ is a GP iff every set of samples $Y=[y_1,..,y_N]$ is a vector-valued Gaussian random variable
- We define a distribution over all functions
  - Constrained by the samples
- The samples have a joint Gaussian distribution in N-dimensional space

Regression: Possible functions constrained by N samples



Case of two samples t_1 and t_2 that are bivariate Gaussian



Specifying a Gaussian Process

- Key point about Gaussian Stochastic Processes
  - Joint distribution over N variables $y_1,..,y_N$ is completely specified by the second-order statistics,
    - i.e., mean and covariance
- With mean zero, it is completely specified by covariance of $y(\mathrm{x})$ evaluated at any two values of x which is given by a kernel function

$$\mathrm{E}[y(\mathrm{x}_n)\, y(\mathrm{x}_m)] = k(\mathrm{x}_n,\mathrm{x}_m)$$

- For the Gaussian Process defined by the linear regression model $y(\mathrm{x},\mathrm{w})=\mathrm{w}^{\mathrm{T}}\phi(\mathrm{x})$ with prior $p(\mathrm{w}) = \mathcal{N}(\mathrm{w}|0,\alpha^{-1}\mathrm{I})$ the kernel function is

$$\mathrm{K}_{\mathrm{nm}} = k(\mathrm{x}_\mathrm{n},\mathrm{x}_\mathrm{m}) = (1/\alpha)\,\phi(\mathrm{x}_\mathrm{n})^{\mathrm{T}}\phi(\mathrm{x}_\mathrm{m})$$

Defining a Kernel Directly

- GP can also be specified directly by choice of kernel function (instead of indirectly by basis function)
- Samples of functions drawn from Gaussian processes for two different choices of kernel functions are shown next

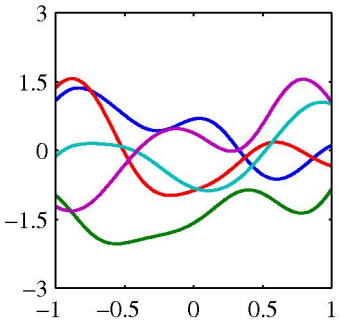
Samples from Gaussian Processes for two kernels

Functions are drawn from Gaussian processes

Note that each sample is a function

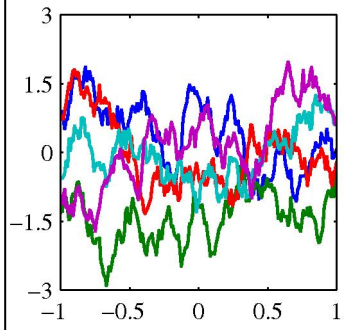
Gaussian Kernel

$$k(\mathrm{x},\mathrm{x}') = \exp\left(-\|\mathrm{x}-\mathrm{x}'\|^2 / 2\sigma^2\right)$$



Exponential Kernel

$$k(\mathrm{x},\mathrm{x}') = \exp\left(-\theta\,|\mathrm{x}-\mathrm{x}'|\right)$$



Ornstein-Uhlenbeck process for Brownian motion

Prob Distributions and Stochastic Processes

- A probability distribution describes the distributions of scalars (x) and vectors (x)
- A stochastic process describes distribution of functions $f(x)$
  - Can think of a function as a very long vector
    - Each entry in the vector is $f(x)$ which is the value of function at x
- A Gaussian process is a generalization of Gaussian distributions where it is describing the distribution of functions
- Since the vector is infinite-dimensional we constrain it to only those points x corresponding to training and test points

GP for Regression (Direct Approach)

- We specify Gaussian Process directly over functions
  - Abandon approach of defining a distribution over weights $\mathbf{w}$
- Take into account noise on observed target values as

$$t_n = y_n + \varepsilon_n \quad \text{where } y_n = y(\mathbf{x}_n)$$

  - Noise process has a Gaussian distribution $p(t_n|y_n)=N(t_n|y_n,\beta^{-1})$
- Note that target t_n is output y_n corrupted by noise
- Defining $\mathbf{t}=(t_1,..,t_N)^T$ our goal is to determine a distribution $p(\mathbf{t})$
  - Which is a distribution over functions

GP for Regression (Direct Approach)

- Assuming noise is independent for each data point

  joint distribution *of* $\mathbf{t} = (t_1,..,t_N)^T$ on values $\mathbf{y} = (y_1,..,y_N)^T$ is

  $$p(\mathbf{t}|\mathbf{y}) = N(\mathbf{t}|\mathbf{y}, \beta^{-1} I_N)$$

- From definition of GP, marginal distribution of $\mathbf{y}$ is given by

  a Gaussian with zero mean, covariance matrix given by Gram matrix $\mathbf{K}$

  $$p(\mathbf{y}) = \mathrm{N}(\mathbf{y}|0,\mathbf{K})$$

- Where kernel function that determines $\mathbf{K}$ is chosen to express:

  property that for points x_n, x_m that are similar corresponding values $y(\mathrm{x}_n)$, $y(\mathrm{x}_m)$ will be more strongly correlated than for dissimilar points

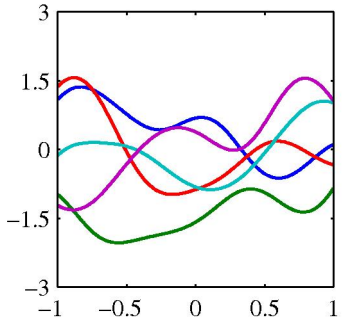
  $\mathbf{K}$ depends on application

Regression: Marginal Distribution

- From distributions $p(\mathrm{t}|\mathrm{y})$ and $p(\mathrm{y})$ we can get marginal $p(\mathrm{t})$ conditioned on input values $\mathrm{x}_1,..,\mathrm{x}_N$
- Applying Sum rule and product rule

$$p(\mathrm{t}) = \int p(\mathrm{t}\,|\,\mathrm{y})p(\mathrm{y})d\mathrm{y}$$
$$= N(\mathrm{t}\,|\,0,C)$$

From the result that when $p(\mathrm{y})$ and $p(\mathrm{t}|\mathrm{y})$ are Gaussian $p(\mathrm{t})$ is also Gaussian



where covariance matrix C has the elements

$$C(\mathrm{x}_n,\mathrm{x}_m) = \underbrace{k(\mathrm{x}_n,\mathrm{x}_m)}_{\text{Due to } \mathrm{y}(\mathrm{x})} + \underbrace{\beta^{-1}\delta_{nm}}_{\text{Due to } \varepsilon}$$

- The two Gaussian sources of randomness, $y(\mathrm{x})$ and ε are independent, so their covariances simply add

δ specified by ε

Widely used kernel function for GP

- Exponential of a quadratic form

  with addition of constant and linear terms

$$k(\mathbf{x}_n, \mathbf{x}_m) = \theta_0 \exp\left\{-\frac{\theta_1}{2} \| \mathbf{x}_n - \mathbf{x}_m \|^2\right\} + \theta_2 + \theta_3 \mathbf{x}_n^T \mathbf{x}_m$$

Corresponds to a parametric model that is a linear function of input variables

- Samples from this prior are plotted for various values of the parameters θ_0, θ_1, θ_2, and θ_3