Bayesian Linear Regression

Sargur Srihari
srihari@cedar.buffalo.edu

Topics in Bayesian Regression

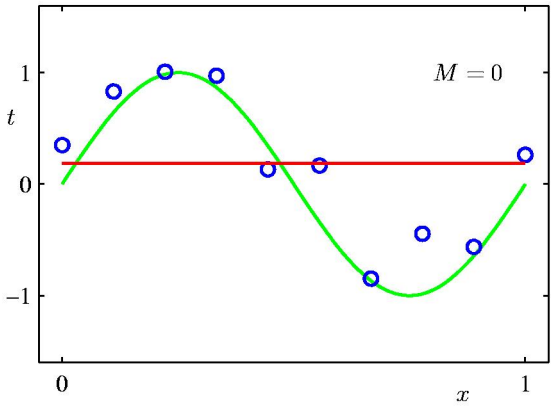
- Recall Max Likelihood Linear Regression
- Parameter Distribution
- Predictive Distribution
- Equivalent Kernel

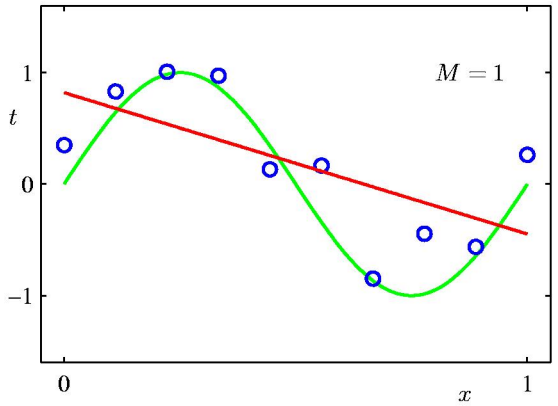
Linear Regression: model complexity M

- Polynomial regression

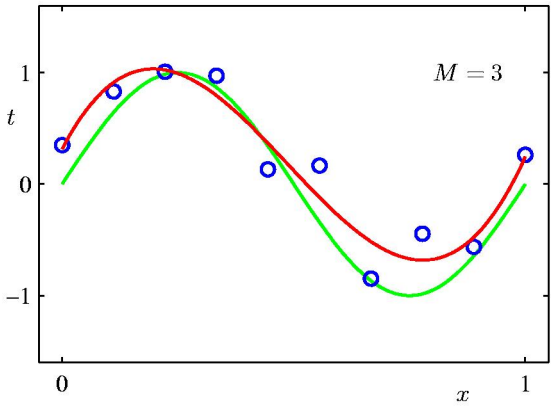
$$y(x, \mathrm{w}) = w_0 + w_1 x + w_2 x^2 + .. + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

  - Red lines are best fits with $M = 0,1,3,9$ and $N=10$

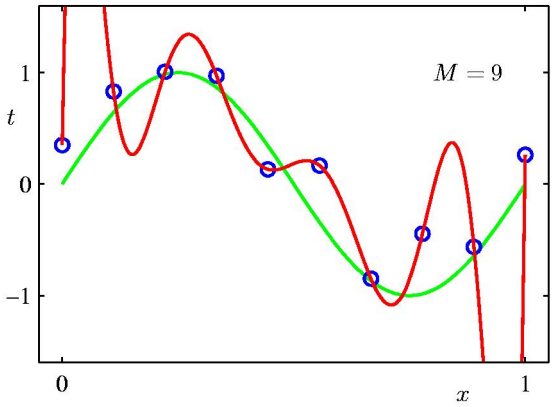




← Poor representations of $\sin(2\pi x)$



← Best Fit to $\sin(2\pi x)$



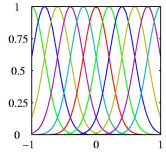
Over Fit Poor representation of $\sin(2\pi x)$

Max Likelihood Regression

- Input vector $\boldsymbol{x}$, basis functions $\{\phi_1(\boldsymbol{x}),..,\phi_M(\boldsymbol{x})\}$:

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

Radial basis fns: $$\phi_j(\boldsymbol{x}) = \exp\left(-\frac{1}{2}(\boldsymbol{x}-\mu_j)^t \Sigma^{-1}(\boldsymbol{x}-\mu_j)\right)$$



- Objective Function:

Max Likelihood objective with N examples $\{\boldsymbol{x}_1,..\boldsymbol{x}_N\}$:
(equivalent to Mean Squared Error Objective)

$$E(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\left\{ t_n - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n)\right\}^2$$

Regularized MSE with N examples:
(λ is the regularization coefficient)

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

- Closed-form ML solution is:

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

where Φ is the design matrix:
$(\Phi^T\Phi)^{-1}$ is Moore-Penrose inverse

$$\Phi = \begin{pmatrix} \phi_0(x_1) & \phi_1(x_1) & \ldots & \phi_{M-1}(x_1) \\ \phi_0(x_2) & & & \\ & & & \\ \phi_0(x_N) & & & \phi_{M-1}(x_N) \end{pmatrix}$$

Regularized solution is:

$$\boldsymbol{w}_{ML} = (\lambda I + \Phi^T\Phi)^{-1}\Phi^T\boldsymbol{t}$$

- Gradient Descent:

$$\boldsymbol{w}^{(\tau+1)} = \boldsymbol{w}^{(\tau)} - \eta\nabla E$$

$$\nabla E = -\sum_{n=1}^{N}\left\{ t_n - \boldsymbol{w}^{(\tau)T}\boldsymbol{\phi}(\boldsymbol{x}_n)\right\}\boldsymbol{\phi}(\boldsymbol{x}_n)$$

Regularized version:

$$\nabla E = \left[-\sum_{n=1}^{N}\left\{ t_n - \boldsymbol{w}^{(\tau)T}\boldsymbol{\phi}(x_n)\right\}\boldsymbol{\phi}(x_n)\right] - \lambda\boldsymbol{w}^{(\tau)}$$

Shortcomings of MLE

- M.L.E. of parameters $\boldsymbol{w}$ 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 effect)

$$E(\boldsymbol{w}) = E_D(\boldsymbol{w}) + \lambda E_W(\boldsymbol{w})$$ where $$E_D(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\left\{ t_n - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n) \right\}^2$$ $$E_W(\boldsymbol{w}) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}$$

  - 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 are better handled using Bayesian approach

Bayesian Linear Regression

- Using Bayes rule, posterior is proportional to Likelihood × Prior:

$$p(\boldsymbol{w} \mid \mathbf{t}) = \frac{p(\mathbf{t} \mid \boldsymbol{w}) p(\boldsymbol{w})}{p(\mathbf{t})}$$

  - where $p(\mathbf{t}|\boldsymbol{w})$ is the likelihood of observed data
  - $p(\boldsymbol{w})$ is prior distribution over the parameters
- We will look at:
  - A normal distribution for prior $p(\boldsymbol{w})$
  - Likelihood $p(\boldsymbol{t}|\boldsymbol{w})$ is a product of Gaussians based on the noise model
  - And conclude that posterior is also Gaussian

Gaussian Prior Parameters

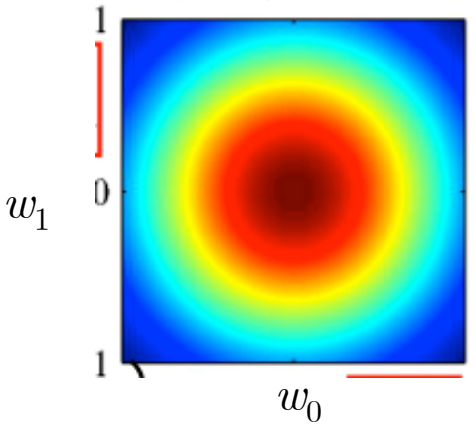
Assume multivariate Gaussian prior for $\boldsymbol{w}$ (which has components $w_0,..,w_{M-1}$)

$$p(\boldsymbol{w}) = N\left(\boldsymbol{w}|\boldsymbol{m}_0, \mathrm{S}_0\right)$$

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

If we choose $\mathrm{S}_0 = \alpha^{-1}\mathrm{I}$ it means that the variances of the weights are all equal to α^{-1} and covariances are zero

$p(\boldsymbol{w})$ with zero mean $(m_0=\mathbf{0})$
and isotropic over weights (*same variances*)



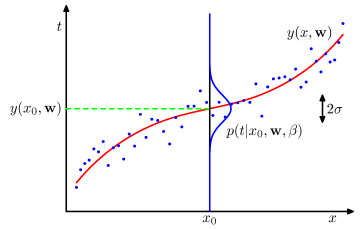
Likelihood of Data is Gaussian

- Assume noise precision parameter β

  $t = y(\boldsymbol{x},\boldsymbol{w})+\varepsilon$ where ε is defined probabilistically as Gaussian noise

  $p(t|\boldsymbol{x},\boldsymbol{w},\beta)=N\left(t|y(\boldsymbol{x},\boldsymbol{w}),\beta^{-1}\right)$

  Note that output t is a scalar



- Likelihood of $\boldsymbol{t} = \{t_1,..,t_N\}$ is then

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

  - This is the probability of target data $\boldsymbol{t}$ given the parameters $\boldsymbol{w}$ and input $X=\{\boldsymbol{x}_1,..,\boldsymbol{x}_N\}$
  - Due to Gaussian noise, likelihood $p(\boldsymbol{t}\,|\boldsymbol{w})$ is also a Gaussian

Posterior Distribution is also Gaussian

- Prior: $p(\boldsymbol{w}) \sim N\left(\boldsymbol{w}|\boldsymbol{m}_0\,,\,\mathrm{S}_0\right)$ i.e., it is Gaussian
- Likelihood comes from Gaussian noise

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

- It follows that posterior $p(\boldsymbol{w}|\boldsymbol{t})$ is also Gaussian
- *Proof:* use standard result from Gaussians:
  - If marginal $p(\boldsymbol{w})$ & conditional $p(\boldsymbol{t}|\boldsymbol{w})$ have Gaussian forms then the marginals $p(\boldsymbol{t})$ and $\boldsymbol{p}(\boldsymbol{w}|\boldsymbol{t})$ are also Gaussian:
    - Let $p(\boldsymbol{w}) = \mathrm{N}(\boldsymbol{w}|\mu,\Lambda^{-1})$ and $p(\boldsymbol{t}|\boldsymbol{w}) = \mathrm{N}(\mathrm{t}|\mathrm{A}\boldsymbol{w}+\boldsymbol{b},\mathrm{L}^{-1})$
    - Then marginal $p(\mathrm{t}) = \mathrm{N}(\mathrm{t}|\mathrm{A}\mu+\mathrm{b},\mathrm{L}^{-1}+\mathrm{A}\Lambda^{-1}\mathrm{A}^{\mathrm{T}})$ and

      conditional $p(\mathrm{w}|\mathrm{t}) = \mathrm{N}(\mathrm{w}|\Sigma\{\mathrm{A}^{\mathrm{t}}\mathrm{L}(\mathrm{t}\text{-}\mathrm{b})+\Lambda\mu\},\Sigma)$

      where $\Sigma=(\Lambda+\mathrm{A}^{\mathrm{T}}\mathrm{LA})^{-1}$

Exact form of Posterior Distribution

- We have $p(\boldsymbol{w}) = N(\boldsymbol{w}|\boldsymbol{m}_0, S_0)$ & $p(\boldsymbol{t}|X,\boldsymbol{w},\beta) = \prod_{n=1}^{N} N\left(t_n | \boldsymbol{w}^T\phi(\boldsymbol{x}_n), \beta^{-1}\right)$
- Posterior is also Gaussian, written directly as

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

  - where $\boldsymbol{m}_N$ is the mean of the posterior

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

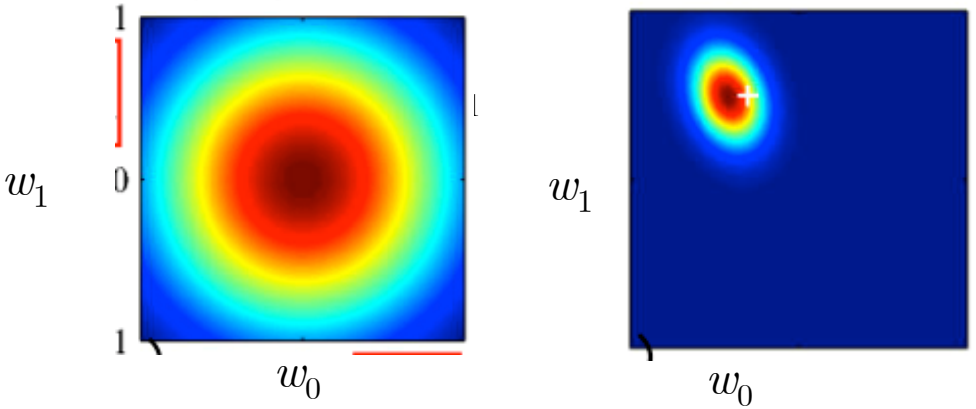
    Φ is the design matrix

$$\Phi = \begin{pmatrix} \phi_0(x_1) & \phi_1(x_1) & \cdots & \phi_{M-1}(x_1) \\ \phi_0(x_2) & & & \\ & & & \\ \phi_0(x_N) & & & \phi_{M-1}(x_N) \end{pmatrix}$$

  - and S_N is the covariance matrix of posterior

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

Prior $p(\boldsymbol{w}|\alpha) = N(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1}I)$
and Posterior in weight space
for scalar input x and
$y(x,\boldsymbol{w}) = w_0 + w_1 x$



Properties of Posterior

1. Since posterior $p(\boldsymbol{w}|\boldsymbol{t})=N(\boldsymbol{w}|\boldsymbol{m}_N,\mathrm{S}_N)$ is Gaussian its mode coincides with its mean
   - Thus maximum posterior weight is $\boldsymbol{w}_{\mathrm{MAP}}=\boldsymbol{m}_N$
2. Infinitely broad prior $\mathrm{S}_0=\alpha^{-1}\mathrm{I}$, i.e., precision $\alpha \to 0$
   - Then mean $\boldsymbol{m}_N$ reduces to the maximum likelihood value, i.e., mean is the solution vector

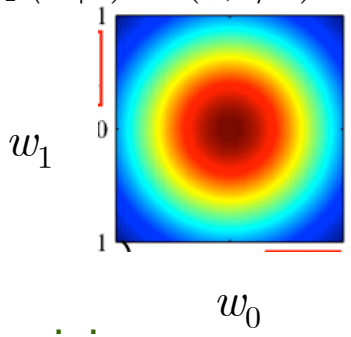
$$\boldsymbol{w}_{ML}=(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\boldsymbol{t}$$

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 $p(\boldsymbol{w})$

$y(x,\mathrm{w}) = w_0 + w_1 x$

$p(\mathrm{w}|\alpha) \sim N(0, 1/\alpha)$



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

$$p(\boldsymbol{w} \mid \alpha) \sim N(\boldsymbol{w} \mid \mathbf{0}, \alpha^{-1}\mathrm{I})$$

Single precision parameter α

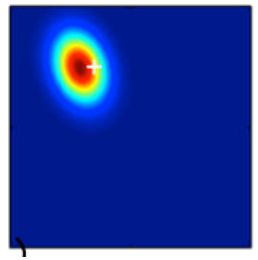
- Corresponding posterior distribution is

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

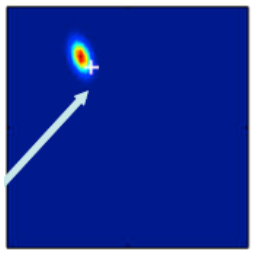
where

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

Note:
β is noise precision and
α is variance of parameter $\boldsymbol{w}$ in prior



Point Estimate with infinite samples



Equivalence to MLE with Regularization

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

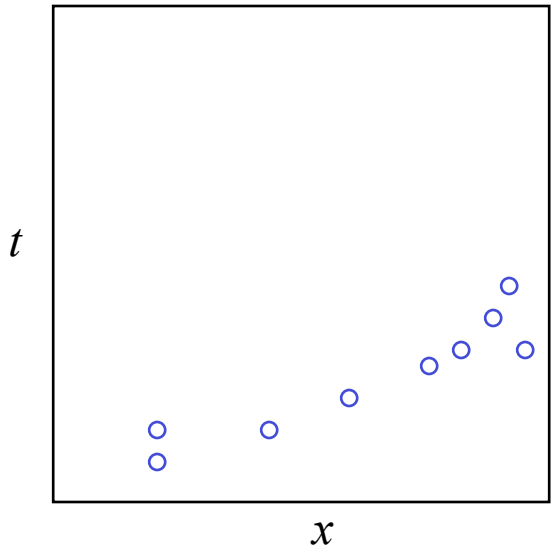
$$\ln p(\boldsymbol{w} \mid \boldsymbol{t}) = -\frac{\beta}{2} \sum_{n=1}^{N} \left\{ t_n - \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_n) \right\}^2 - \frac{\alpha}{2} \boldsymbol{w}^T \boldsymbol{w} + \text{const}$$

- Thus Maximization of posterior is equivalent to minimization of sum-of-squares error

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

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

Bayesian Linear Regression Example (Straight Line Fit)

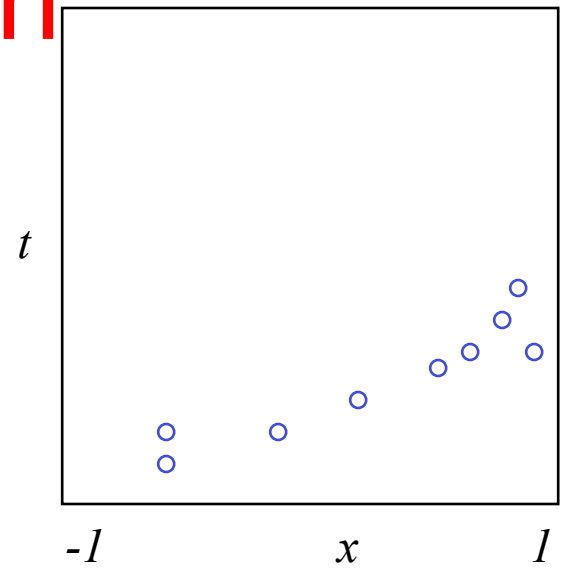


- Single input variable x
- Single target variable t
- Goal is to fit
  - Linear model $y(x, \boldsymbol{w}) = w_0 + w_1 x$
- Goal of Linear Regression is to recover

  $\boldsymbol{w} = [w_0, w_1]$ given the samples

Data Generation

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



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

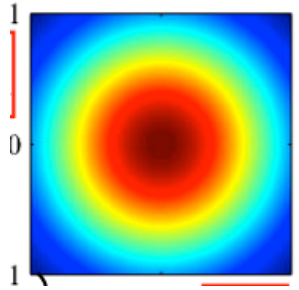
- For prior over $\boldsymbol{w}$ we choose $\alpha = 2$

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

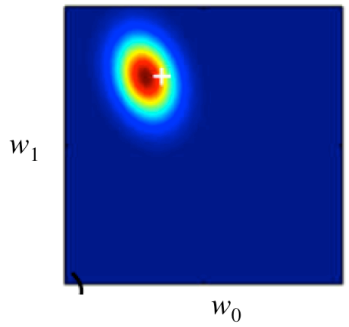
Sampling $p(\boldsymbol{w})$ and $p(\boldsymbol{w}|\boldsymbol{t})$

- Each sample represents a straight line in data space (modified by examples)

Distribution



$p(\boldsymbol{w})$

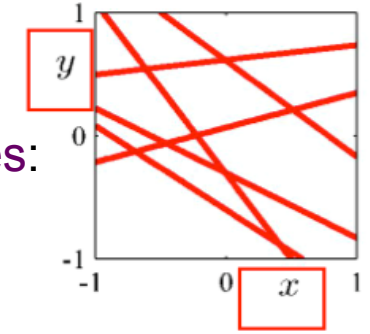


$p(\boldsymbol{w}|\boldsymbol{t})$

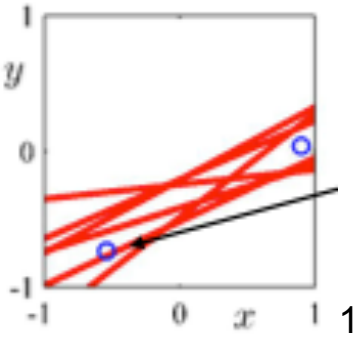
Six samples

$y(x,\boldsymbol{w}) = w_0 + w_1 x$

With no examples:



With two examples:



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

Sequential Bayesian Learning

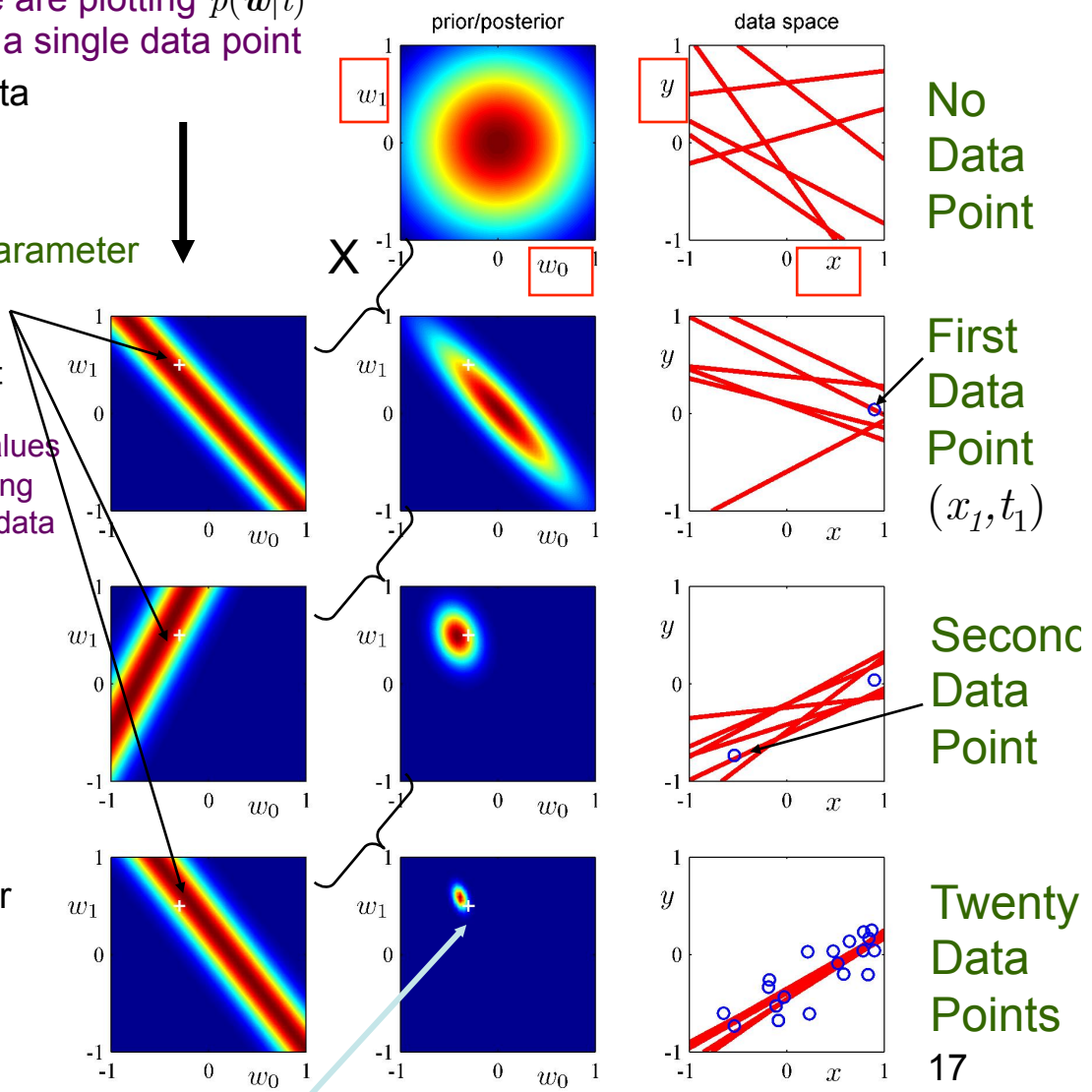
- Since there are only two parameters
  - We can plot prior and posterior distributions in parameter space
- We look at sequential update of posterior

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

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

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

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



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

Generalization of Gaussian prior

- The Gaussian prior over parameters is

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

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

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

- Other prior yields Lasso and variations:

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

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

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

Predictive Distribution

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

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

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

  - Leaving out conditioning variables X and $\boldsymbol{x}$ for convenience
- Marginalizing over parameter variable $\boldsymbol{w}$, is the standard Bayesian approach
  - Sum rule of probability $\boxed{p(t)=\int p(t,\boldsymbol{w})\,d\boldsymbol{w} = \int p(t|\boldsymbol{w})p(\boldsymbol{w})\,d\boldsymbol{w}}$
  - We can now write $\boxed{p(t \mid \boldsymbol{t})=\int p(t|\boldsymbol{w})p(\boldsymbol{w}|\boldsymbol{t})\,d\boldsymbol{w}}$

Predictive Distribution with $\alpha, \beta, \boldsymbol{x}, t$

- We can predict t for a new value of $\boldsymbol{x}$ using

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

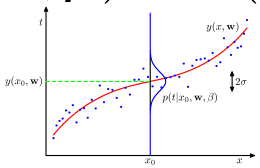
We have left out conditioning variables X and $\boldsymbol{x}$ for convenience.
Also we have applied sum rule of probability $p(t) = \Sigma_{\boldsymbol{w}} p(t \mid \boldsymbol{w}) p(\boldsymbol{w})$

  - With explicit dependence on prior parameter α, noise parameter β, & targets in training set $\boldsymbol{t}$

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

Conditional of target t given weight $\boldsymbol{w}$

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



posterior of weight $\boldsymbol{w}$

$$p(\boldsymbol{w} \mid \boldsymbol{t}) = N(\boldsymbol{w} \mid \boldsymbol{m}_N, \mathrm{S}_N)$$

where $\boldsymbol{m}_N = \beta\, \mathrm{S}_N \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{t}$

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

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

$$p(t \mid \boldsymbol{x}, \boldsymbol{t}, \alpha, \beta) = N(t \mid \boldsymbol{m}_N^T \boldsymbol{\phi}(\boldsymbol{x}), \sigma_N^2(\mathrm{x})) \quad \text{where} \quad \sigma_N^2(\boldsymbol{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\boldsymbol{x})^T S_N \boldsymbol{\phi}(\boldsymbol{x})$$

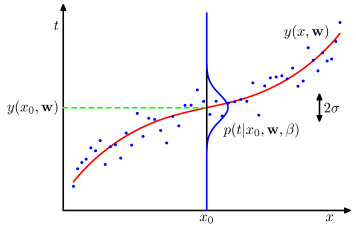
Variance of Predictive Distribution

- Predictive distribution is a Gaussian:

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

$$\text{where} \quad \sigma_N^2(\boldsymbol{x}) = \underbrace{\frac{1}{\beta}}_{\text{Noise in data}} + \underbrace{\boldsymbol{\phi}(\boldsymbol{x})^T S_N \boldsymbol{\phi}(\boldsymbol{x})}$$

Since noise process and distribution of w are independent Gaussians their variances are additive



Uncertainty associated with parameters $\boldsymbol{w}$:

where $S_N^{-1} = \alpha \mathrm{I} + \beta \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi}$ is the covariance of $p(\boldsymbol{w}|\alpha)$

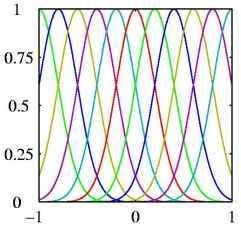
Since $\sigma_{N+1}^2(x) \leq \sigma_N^2(x)$ as no. of samples increases it becomes narrower

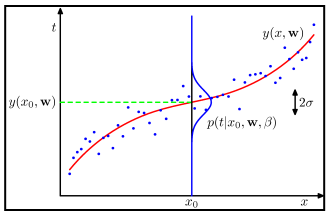
As $N \rightarrow \infty$, second term of variance goes to zero and variance of predictive distribution arises solely from the additive noise parameter β

Example of Predictive Distribution

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

$$y(x,\mathrm{w}) = \sum_{j=0}^{8} w_j \phi_j(x) = \mathrm{w}^T \phi(x) \qquad \phi_j(x) = \exp\left\{\frac{(x-\mu_j)^2}{2\sigma^2}\right\}$$

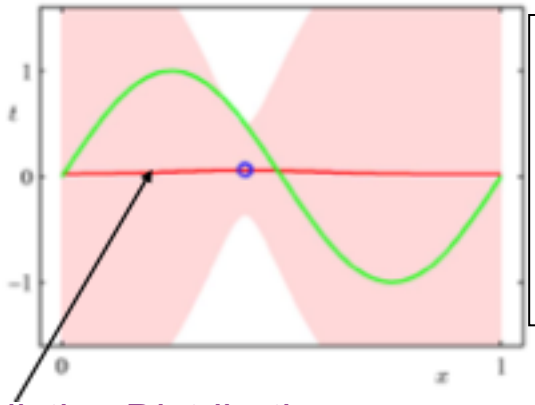




- Predictive distribution

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

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



Mean of Predictive Distribution

where $m_N = \beta\, S_N \Phi^T \mathrm{t}$, $S_N^{-1} = \alpha \mathrm{I} + \beta\, \Phi^T \Phi$
and α and β come from assumptions
$p(\mathrm{w}|\alpha) = N(\mathrm{w}|0, \alpha^{-1} I)$
$p(t \mid x, \boldsymbol{w}, \beta) = N(t \mid y(x, \boldsymbol{w}), \beta^{-1})$

Predictive Distribution Variance

Bayesian prediction:

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

where we have assumed Gaussian prior over parameters:

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

Noise model assumed Gaussian:

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

and use design matrix as:

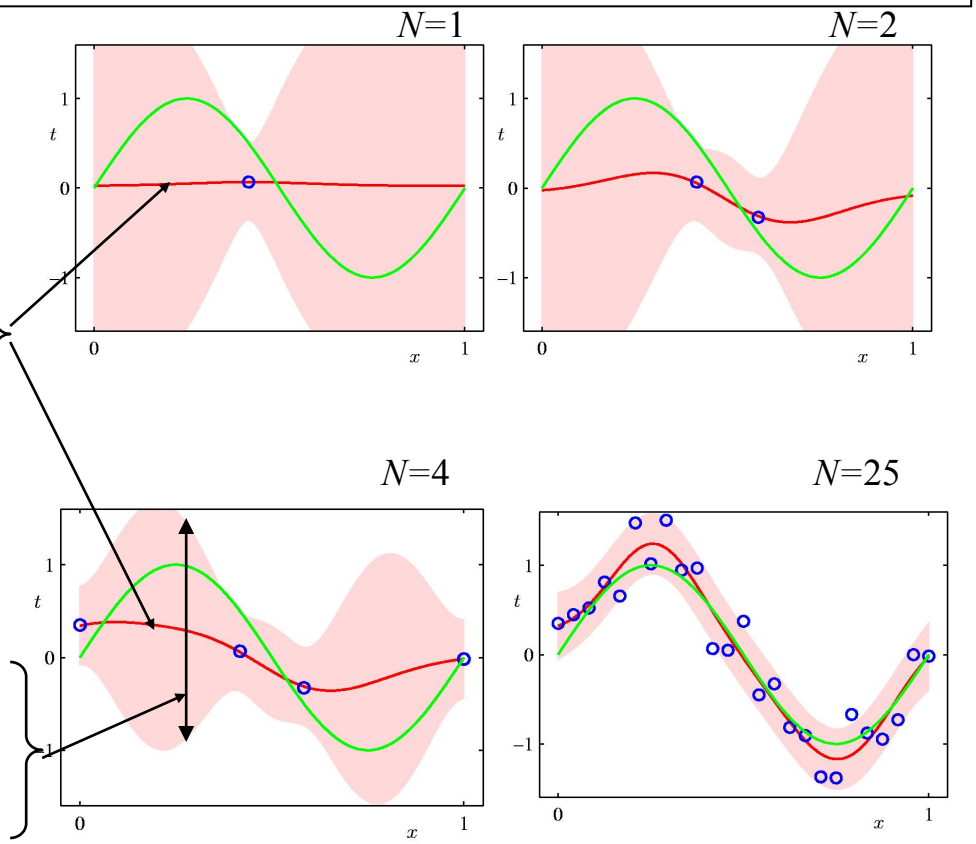
$$S_N^{-1} = \alpha \mathrm{I} + \beta \Phi^{\mathrm{T}} \Phi$$

$$\Phi = \begin{pmatrix} \phi_0(x_1) & \phi_1(x_1) & \dots & \phi_{M-1}(x_1) \\ \phi_0(x_2) & & & \\ & & & \\ \phi_0(x_N) & & & \phi_{M-1}(x_N) \end{pmatrix}$$

Using data from $\sin(2\pi x)$:

$\sigma_N^2(x)$, 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 $\boldsymbol{w}$ $p(\boldsymbol{w}|\boldsymbol{t})$ and plot corresponding functions $y(x, \boldsymbol{w})$

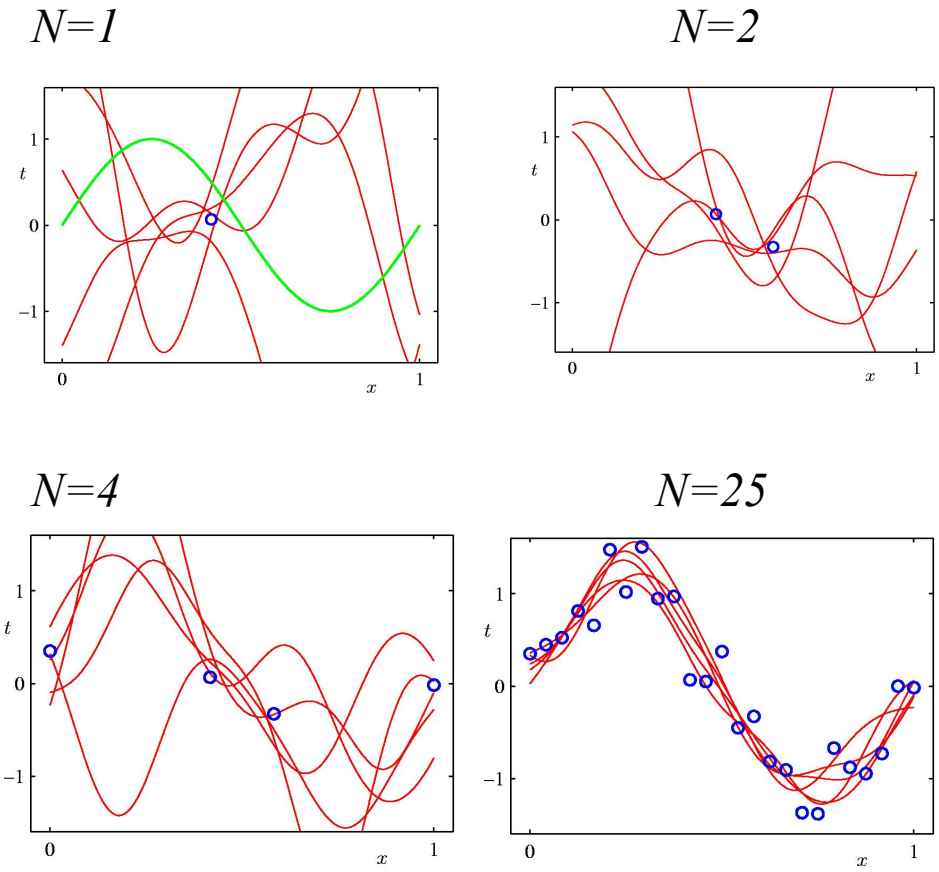
Plots of function $y(x,\boldsymbol{w})$

Draw samples $\mathbf{w}$ from from posterior distribution $p(\boldsymbol{w}|\boldsymbol{t})$

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

and plot samples from $y(x,\boldsymbol{w}) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(x)$

- Shows covariance between predictions at different values of x
  - For a given function, for a pair of x,x', the values of y,y' are determined by $k(x,x')$ which in turn is determined by the samples



Disadvantage of Local Basis

- Predictive distribution, assuming Gaussian prior
  - $p(\boldsymbol{w} \mid \alpha) = N(\boldsymbol{w} \mid \mathbf{0}, \alpha^{-1}\mathrm{I})$ and Gaussian noise $t = y(\boldsymbol{x},\boldsymbol{w})+\varepsilon$
  - where noise is defined probabilistically as $p(t|\boldsymbol{x},\boldsymbol{w},\beta)=N(t|y(\boldsymbol{x},\boldsymbol{w}),\beta^{-1})$

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

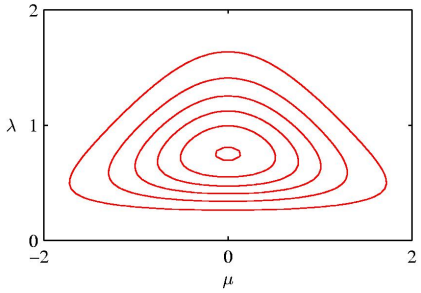
$S_N^{-1}=\alpha\mathrm{I}+\beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}$

- With localized basis functions, e.g., Gaussian
  - at regions away from basis function centers, contribution of second term of variance σ_n^2 in will go to zero leaving only noise contribution β^{-1}
    - Model becomes very confident outside of region occupied by basis functions
  - Problem avoided by alternative Bayesian approach of Gaussian Processes

Dealing with unknown β

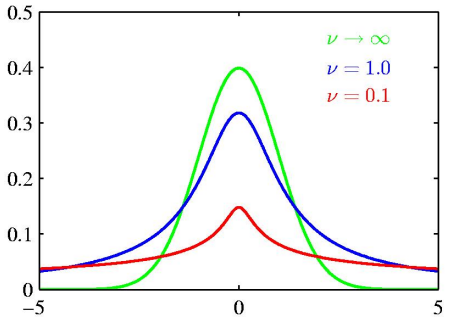
- If both $\boldsymbol{w}$ and β are treated as unknown then we can introduce a conjugate prior distribution $p(\boldsymbol{w},\beta)$ which is given by a *Gaussian-gamma* distribution

$$p(\mu,\lambda) = N\left(\mu \mid \mu_0,\left(\beta\lambda\right)^{-1}\right)Gam\left(\lambda \mid a,b\right)$$



  - In this case the predictive distribution is a *Student's t*-distribution

$$St(x \mid \mu,\lambda,\nu) = \frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu+2)}\left(\frac{\lambda}{\pi\nu}\right)^{1/2}\left[1+\frac{\lambda(x-\mu)^2}{\nu}\right]^{-\nu/2-1/2}$$



Mean of $p(\boldsymbol{w}|\boldsymbol{t})$ has Kernel Interpretation

- Regression function is:

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

- If we take a Bayesian approach with Gaussian prior $p(\boldsymbol{w}) = N(\boldsymbol{w}|\boldsymbol{m}_0\ ,\ \mathrm{S}_0)$ then we have:
  - Posterior $p(\boldsymbol{w}|\boldsymbol{t}) = N\ (\boldsymbol{w}|\boldsymbol{m}_N, S_N)$ where

$$\boldsymbol{m}_\mathrm{N} = \mathrm{S}_N\,(\mathrm{S}_0^{-1}\boldsymbol{m}_0 + \beta\Phi^\mathrm{T}\boldsymbol{t})$$

$$\mathrm{S}_N^{-1} = \mathrm{S}_0^{-1} + \beta\Phi^\mathrm{T}\Phi$$

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

$$\boldsymbol{m}_\mathrm{N} = \beta\ \mathrm{S}_N\Phi^\mathrm{T}\mathrm{t},$$

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

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

Equivalent Kernel

- Posterior mean of $\boldsymbol{w}$ is $\boldsymbol{m}_N = \beta \mathrm{S}_N \Phi^{\mathrm{T}} \mathrm{t}$
  - where $\mathrm{S}_N^{-1} = \mathrm{S}_0^{-1} + \beta \Phi^{\mathrm{T}} \Phi$,
    - S_0 is the covariance matrix of the prior $p(\boldsymbol{w})$, β is the noise parameter and Φ is the design matrix that depends on the samples
- Substitute mean value into Regression function

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

- Mean of predictive distribution at point $\boldsymbol{x}$ is

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

  - where $k(\boldsymbol{x}, \boldsymbol{x}') = \beta \phi(\boldsymbol{x})^T S_N \phi(\boldsymbol{x}')$ is the *equivalent kernel*
- Thus mean of predictive distribution is a linear combination of training set target variables t_n
  - Note: the equivalent kernel depends on input values $\boldsymbol{x}_n$ from the dataset because they appear in S_N

Kernel Function

- Regression functions such as

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

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

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

$$\Phi = \begin{pmatrix} \phi_0(x_1) & \phi_1(x_1) & \ldots & \phi_{M-1}(x_1) \\ \phi_0(x_2) & & & \\ & & & \\ \phi_0(x_N) & & & \phi_{M-1}(x_N) \end{pmatrix}$$

- That take a linear combination of the training set target values are known as *linear smoothers*
- They depend on the input values $\boldsymbol{x}_n$ from the data set since they appear in the definition of S_N

Example of kernel for Gaussian Basis

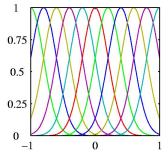
Equivalent Kernel

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

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

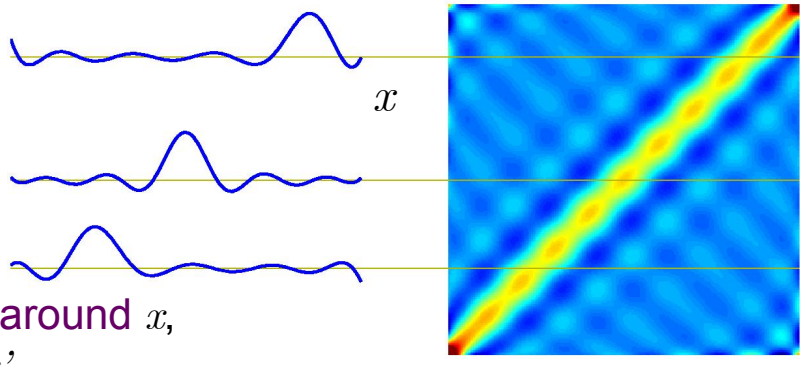
$$\Phi = \begin{pmatrix} \phi_0(x_1) & \phi_1(x_1) & \ldots & \phi_{M-1}(x_1) \\ \phi_0(x_2) & & & \\ & & & \\ \phi_0(x_N) & & & \phi_{M-1}(x_N) \end{pmatrix}$$

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



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

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



Kernels are localized around x, i.e., peaks when $x = x'$

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

Kernel used directly in regression.

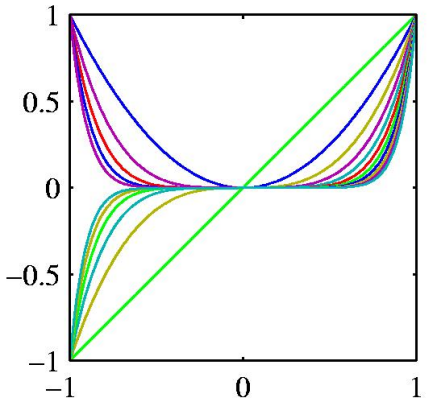
Mean of the predictive distribution is $y(\boldsymbol{x}, \boldsymbol{m}_N) = \sum_{n=1}^{N} k(\boldsymbol{x}, \boldsymbol{x}_n) t_n$

Obtained by forming a weighted combination of target values:

Data points close to x are given higher weight than points further removed from x

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

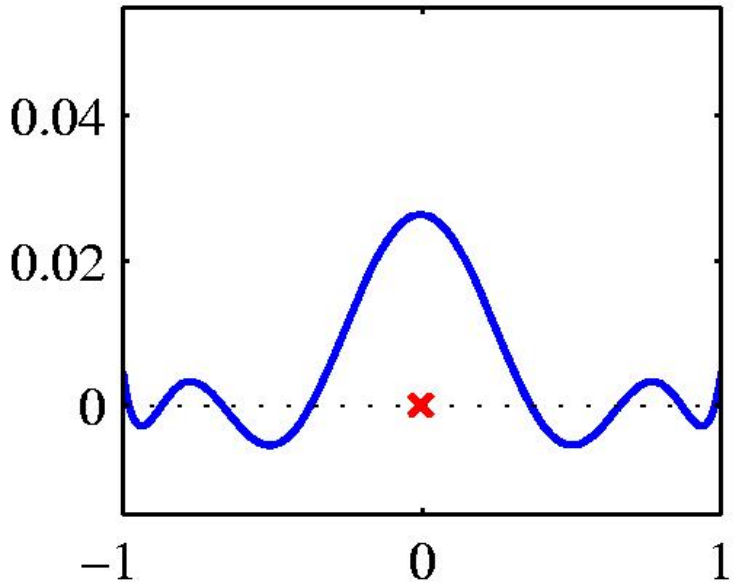
Equivalent Kernel for Polynomial Basis Function



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

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

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

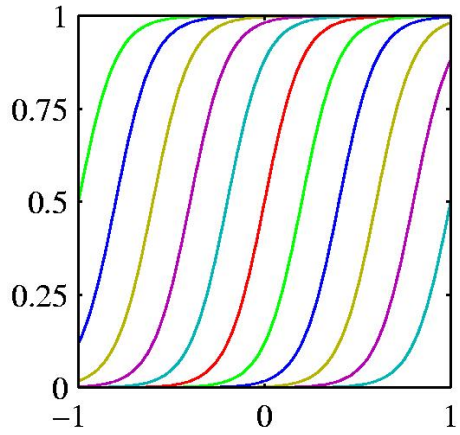


Plotted as a function of x' for $x=0$

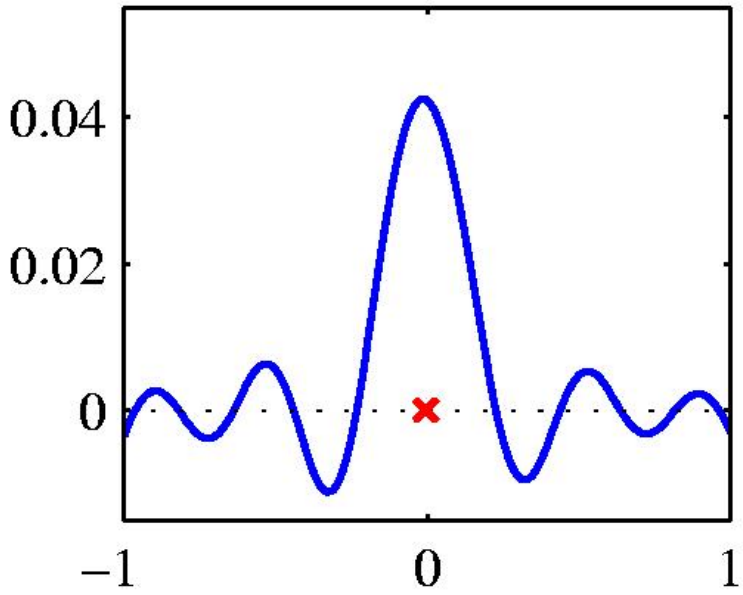
Data points close to x are given higher weight than points further removed from x

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

Equivalent Kernel for Sigmoidal Basis Function



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



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

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

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

An important insight:

The value of the kernel function between two points is directly related to the covariance between their target values

$$\begin{aligned}\text{cov}\,[y(\boldsymbol{x}),\, y(\boldsymbol{x}')] &= \text{cov}[\phi(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{w},\, \boldsymbol{w}^{\mathrm{T}}\phi\,(\boldsymbol{x}')] \\ &= \phi\,(\boldsymbol{x})^{\mathrm{T}} S_{\mathrm{N}} \phi\,(\boldsymbol{x}') \\ &= \beta^{-1} k(\boldsymbol{x},\, \boldsymbol{x}')\end{aligned}$$

where we have used:
$p(\boldsymbol{w}|\boldsymbol{t}) \sim N(\boldsymbol{w}|\boldsymbol{m}_N, S_N)$
$k(\boldsymbol{x},\, \boldsymbol{x}') = \beta\phi(\boldsymbol{x})^T S_N \phi\,(\boldsymbol{x})$

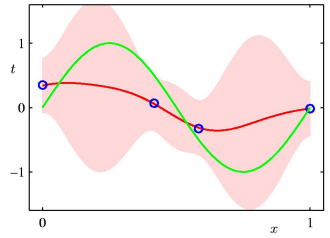
From the form of the equivalent kernel $k(\boldsymbol{x},\, \boldsymbol{x}')$
the predictive mean at nearby points $y(\boldsymbol{x})$, $y(\boldsymbol{x}')$ will be highly correlated
For more distant pairs correlation is smaller

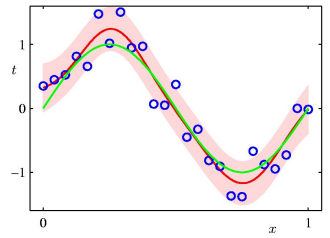
The kernel captures the covariance

Predictive plot vs. Posterior plots

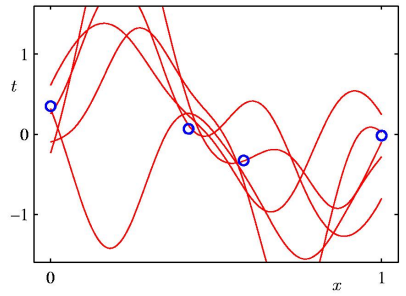
- Predictive distribution
  - allows us to visualize pointwise uncertainty in the predictions governed by

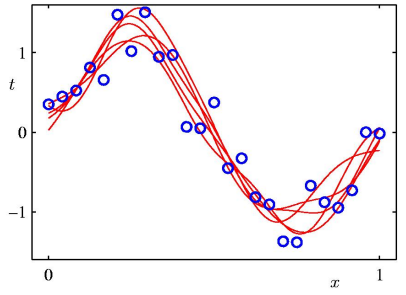
$$p(t \mid x, \mathbf{t}, \alpha, \beta) = N(t \mid m_N^T \boldsymbol{\phi}(x), \sigma_N^2(x)) \quad \text{where} \quad \sigma_N^2(x) = \frac{1}{\beta} + \boldsymbol{\phi}(x)^T S_N \boldsymbol{\phi}(x)$$





- Drawing samples from posterior $p(\boldsymbol{w}|\boldsymbol{t})$
  - Plotting corresponding functions $y(x, \boldsymbol{w})$ we visualize joint uncertainty in the posterior distribution between the y values at two or more x values as governed by the kernel





Directly Specifying Kernel Function

- Formulation of Linear Regression in terms of kernel function suggests an alternative approach to regression:
  - Instead of introducing a set of basis functions, which implicitly determines an equivalent kernel:
  - Directly define kernel functions and use to make predictions for new input $\boldsymbol{x}$, given the observation set
- This leads to a practical framework for regression (and classification) called *Gaussian Processes*

Summing Kernel Values Over samples

- Effective kernel defines weights by which
  - target values combined to make a prediction at $\boldsymbol{x}$
- It can be shown that weights sum to one, i.e.,

$$\sum_{n=1}^{N} k(\boldsymbol{x}, \boldsymbol{x}_n) = 1$$

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

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

  - For all values of $\boldsymbol{x}$
  - This result can be proven intuitively:
    - Since $y(\boldsymbol{x}, \mathrm{m}_N) = \sum_{n=1}^{N} k(\boldsymbol{x}, \boldsymbol{x}_n) t_n$ summation is equivalent to considering predictive mean $\hat{y}(\boldsymbol{x})$ for a set of integer data in which $t_n = 1$ for all n
    - Provided basis functions are linearly independent, that $N > M$, one of the basis functions is constant (corresponding to the bias parameter), then we can fit training data exactly, and hence $\hat{y}(\boldsymbol{x}) = 1$

Kernel Function Properties

- Equivalent kernel can be positive or negative

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

  - Although it satisfies a summation constraint, the corresponding predictions are not necessarily a convex combination of the training set target variables
- Equivalent kernel satisfies important property shared by kernel functions in general.
  - It can be expressed in the form of an inner product wrt a vector $\boldsymbol{\psi}(\boldsymbol{x})$ of nonlinear functions:

$$k(\boldsymbol{x}, \boldsymbol{z}) = \boldsymbol{\psi}(\boldsymbol{x})^T \boldsymbol{\psi}(\boldsymbol{z}) \quad \text{where} \quad \boldsymbol{\psi}(\boldsymbol{x}) = \beta^{1/2} S_N^{1/2} \boldsymbol{\phi}(\boldsymbol{x})$$