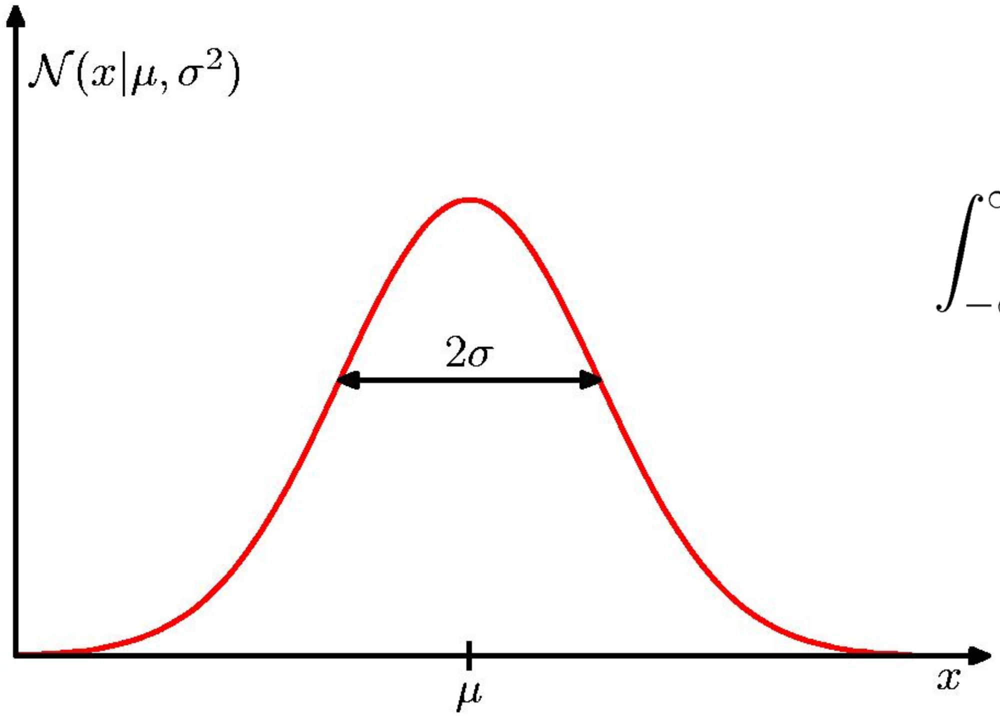
The Gaussian Distribution

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



$$\mathcal{N}(x|\mu,\sigma^2) > 0$$

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right)\,\mathrm{d}x = 1$$

Gaussian Mean and Variance

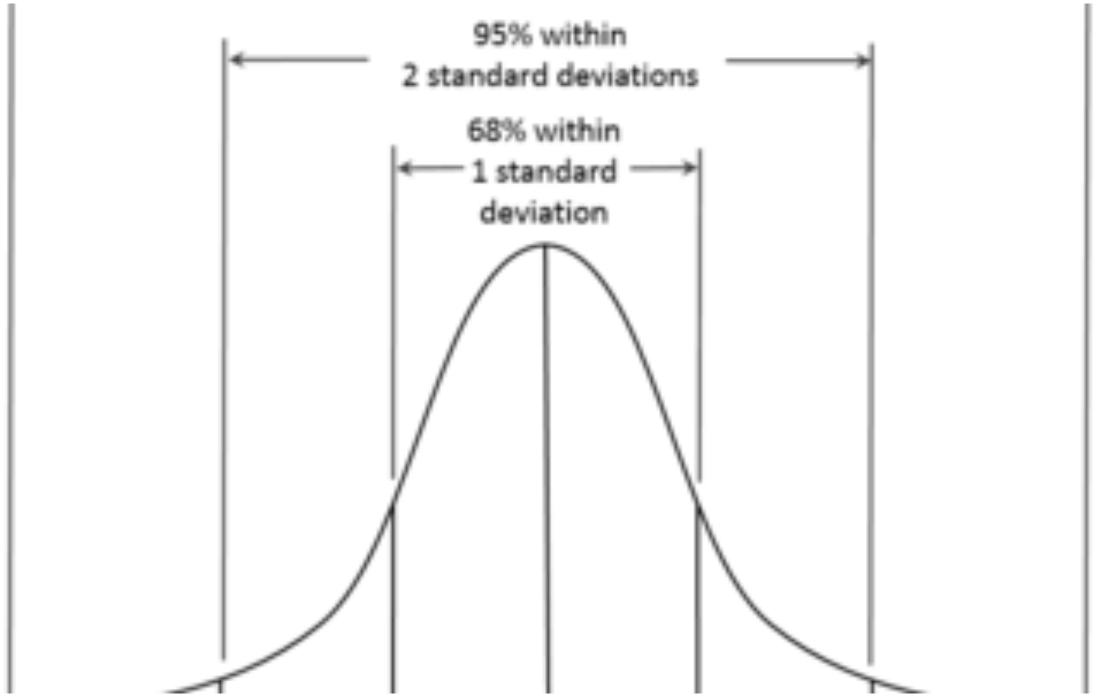
$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) x \, \mathrm{d}x = \mu$$ (mean)

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) x^2 \, \mathrm{d}x = \mu^2 + \sigma^2$$

$$\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2$$ (variance)

$\beta = 1/\sigma^2$ (precision – the bigger β is, the smaller σ is,
thus the more “precise” the distribution is.)

The Gaussian Distribution



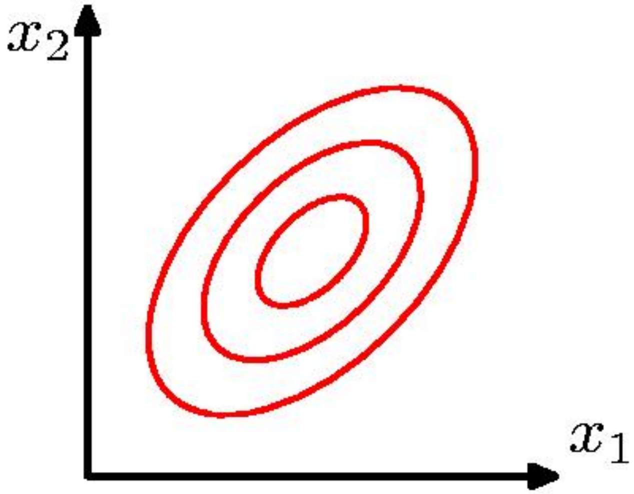
For the normal distribution, the values less than one standard deviation away from the mean account for 68.27% of the set; while two standard deviations from the mean account for 95.45%; and three standard deviations account for 99.73%.

The Multivariate Gaussian

Gaussian distribution defined over a D-dimensional vector x of continuous variables:

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

where the **D**-dimensional vector μ is called the mean, the **D x D** symmetric matrix Σ is called the covariance, and **|Σ|** denotes the determinant of Σ.



Bayes' Theorem Recall

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

$$p(X) = \sum_{Y} p(X|Y)p(Y)$$

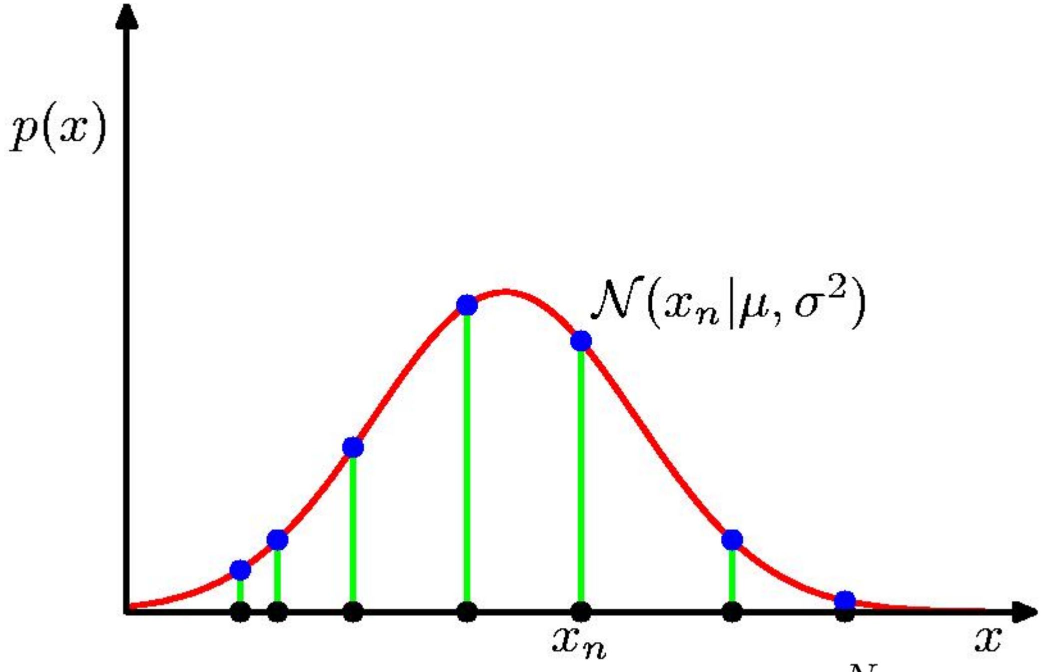
posterior ∝ likelihood × prior

In the example of boxes of fruits: (1) sample B, then (2) sample F

p(B=r | F=o) = p(F=o | B=r) p(B=r) / p(F=o)

prior: the probability available before we observe the identity of the fruit
posterior: the probability obtained after we have observed the identity of the fruit
likelihood: how probable the observed fruit is for different settings of the boxes

Warm Up - Gaussian Parameter Estimation



x = (x_1, x_2, … , x_N) are drawn independently from the Gaussian distribution $\mathcal{N}\left(x|\mu,\sigma^2\right)$.

$$p(\mathbf{x}|\mu,\sigma^2) = \prod_{n=1}^{N} \mathcal{N}\left(x_n|\mu,\sigma^2\right)$$

Likelihood function for the Gaussian distribution
i.e. the probability of the data given the parameters

Goal: Determine the parameters in the above probability distribution using an observed data set x.

Maximum (Log) Likelihood

One common criterion for determining the parameters in a probability distribution using an observed data set is to find the parameter values that maximize the likelihood function (equivalent to maximizing its log).

$$\ln p\left(\mathbf{x}|\mu,\sigma^2\right) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$$

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}x_n$$

sample mean, i.e., the mean of the observed values $\{x_n\}$

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2$$

sample variance, measured with respect to the sample mean μ_{ML}

Remark: It would seem more natural to maximize the probability of the parameters given the data, not the probability of the data given the parameters. In fact, these two criteria are related, as we shall discuss in the context of curve fitting.

Properties of μ_{ML} and σ^2_{ML}

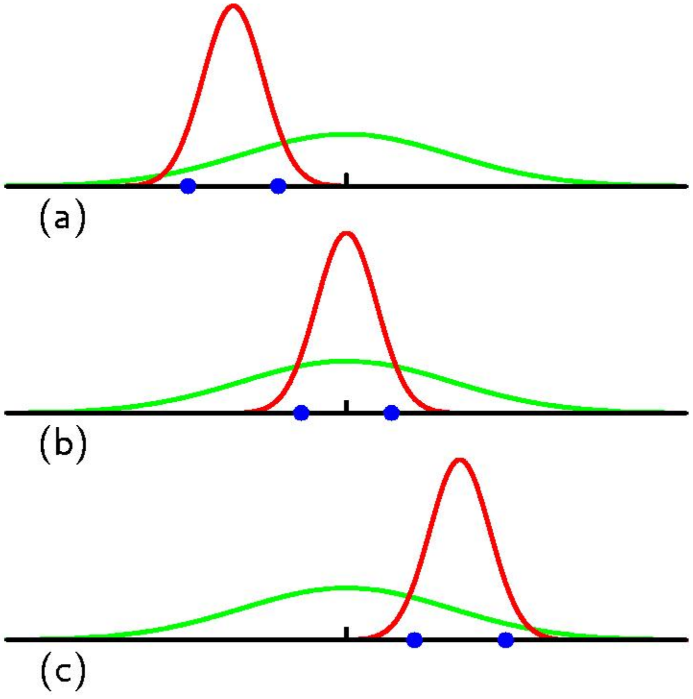
The maximum likelihood approach systematically underestimates the variance of the distribution – bias.

Consider the expectations of these quantities w.r.t. the joint distr. over x = (x_1, x_2, … , x_N), where each x_i comes from Gauss. distr. $\mathcal{N}\left(x|\mu,\sigma^2\right)$

$$\mathbb{E}[\mu_{\mathrm{ML}}] = \mu$$

$$\mathbb{E}[\sigma^2_{\mathrm{ML}}] = \left(\frac{N-1}{N}\right)\sigma^2$$

e.g.

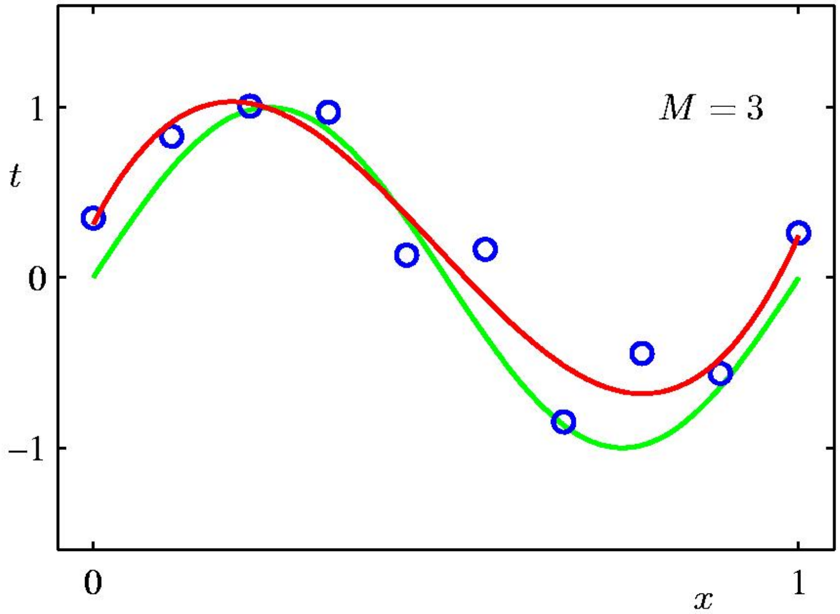


Unbiased estimate for the variance:

$$\widetilde{\sigma}^2 = \frac{N}{N-1}\sigma^2_{\mathrm{ML}} = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2$$

Averaged across the three data sets, the mean is correct, but the variance is systematically under-estimated because it is measured relative to the sample mean and not relative to the true mean.

Curve Fitting Re-visited

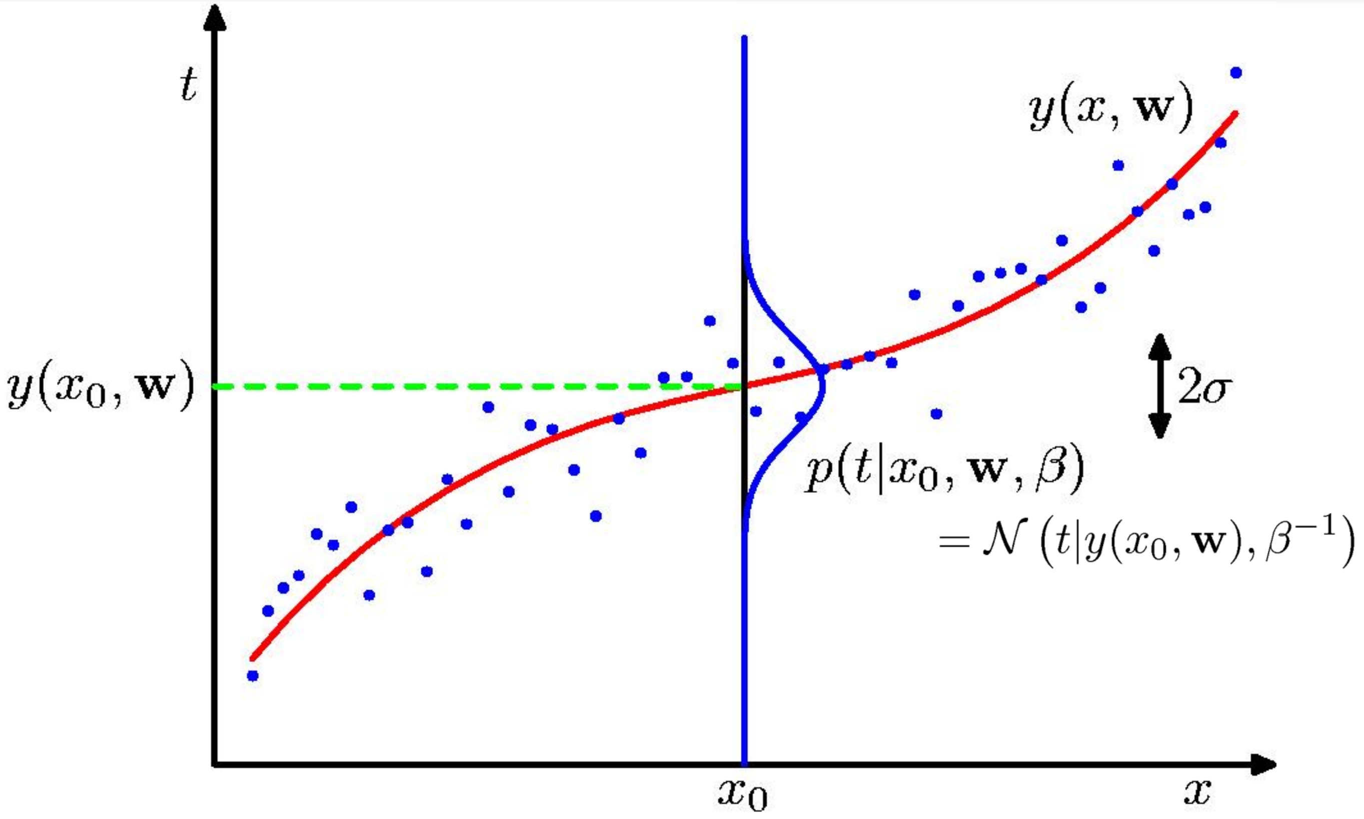


The goal in the curve fitting problem is to be able to make predictions for the target variable *t* given some new value of the input variable *x* on the basis of a set of training data comprising *N* input values *x = (x*$_1$*, . . . , x*$_N$*)* and their corresponding target values *t = (t*$_1$*, . . . , t*$_N$*)*.

Polynomial Curve Fitting: $y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$

Remark: The original (green) function may never be known from given condition.

Curve Fitting Re-visited



1) Obtain w according to a prior probability *p*(w)
2) For each input *x* from $x_1, x_2, \ldots, x_N$ distributed uniformly over an interval, generate target value *t* according to Gauss. distri. $p(t|x, \mathbf{w}, \beta) = \mathcal{N}\left(t|y(x, \mathbf{w}), \beta^{-1}\right)$

where $\beta = 1/\sigma^2$ (precision)

Frequentist - Maximum Likelihood

Frequentist: w is to be fixed.

Likelihood function: $p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}\left(t_n | y(x_n, \mathbf{w}), \beta^{-1}\right)$

$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \underbrace{-\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2}_{\beta E(\mathbf{w})} + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi)$$

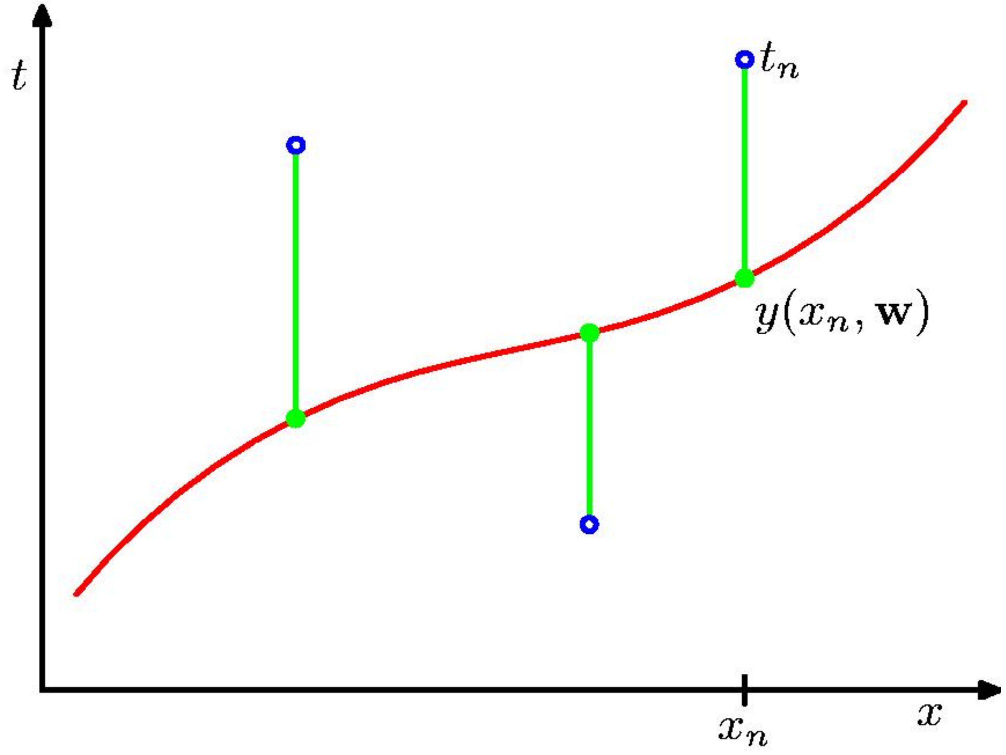
Determine $\mathbf{w}_{\mathrm{ML}}$ by minimizing sum-of-squares error, $E(\mathbf{w})$.

(Thus the sum-of-squares error function has arisen as a consequence of maximizing likelihood under the assumption of a Gaussian noise distribution.)

By maximizing the log max. likelihood w.r.t. β :

$$\frac{1}{\beta_{\mathrm{ML}}} = \frac{1}{N}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}_{\mathrm{ML}}) - t_n\}^2$$

Sum-of-Squares Error Function Recall

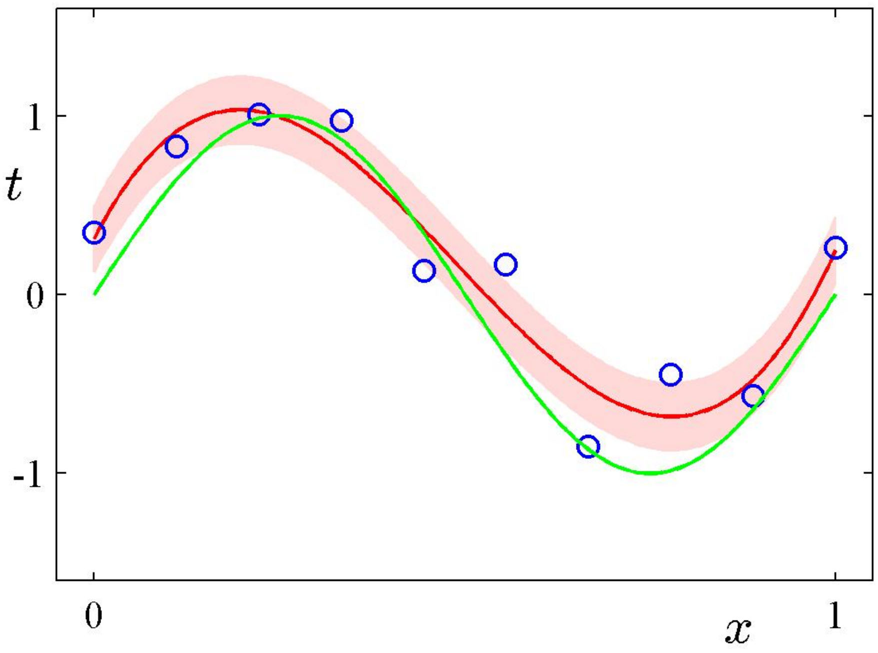


$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2$$

Predictive Distribution

Because we now have a probabilistic model, these are expressed in terms of the *predictive distribution* that gives the probability distribution over *t*, rather than simply a point estimate:

$$p(t|x, \mathbf{w}_{\mathrm{ML}}, \beta_{\mathrm{ML}}) = \mathcal{N}\left(t|y(x, \mathbf{w}_{\mathrm{ML}}), \beta_{\mathrm{ML}}^{-1}\right)$$



Towards Bayes – Maximal Posterior (MAP)

Bayesian: w is from a probability distribution.

For simplicity, $p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right\}$

where α is the precision of the distribution, and M+1 is the total number of elements in the vector w for an M^{th} order polynomial.

(α is to restrict the magnitude of polynomial coefficients w)

Recall: Frequentist maximizes the following likelihood function to determine w:

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}\left(t_n|y(x_n, \mathbf{w}), \beta^{-1}\right)$$

However, it would seem more natural to maximize the probability of the parameters given the data, not the probability of the data given the parameters.

By Bayes Theorem:

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha)$$

MAP: A Step towards more Bayes

Goal: Maximize $p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha)$

where $p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}\left(t_n|y(x_n, \mathbf{w}), \beta^{-1}\right)$

and $p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right\}$

By maximizing the log function of RHS, it comes to minimize the following:

$$\beta\widetilde{E}(\mathbf{w}) = \frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

Now, determine $\mathbf{w}_{\mathrm{MAP}}$ by minimizing regularized sum-of-squares error, $\widetilde{E}(\mathbf{w})$.
with a regularization parameter given by $\lambda = \alpha/\beta$.

Regularization Recall

Penalize large coefficient values

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

Bayesian Curve Fitting

Given: the training data set $\mathbf{x}$ and $\mathbf{t}$, along with a new test point x
Goal: predict the value of t,
i.e. evaluate the predictive distribution $p(t \mid x, \mathbf{x}, \mathbf{t})$

Assume that the parameters α and β are fixed and known in advance.
(In textbook, $\alpha = 5 \times 10^{-3}$, and $\beta = 11.1$)

By generalize Bayes Theorem, $p(W|T, X) = p(T|X, W)\, p(W|X) \,/\, p(T|X)$

thus, $p(T|X)$ = sum over W of $p(T|X, W)\, p(W|X)$

So, by setting X = $x, \mathbf{x}, \mathbf{t}$, we obtain

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

Bayesian Curve Fitting

Given: the training data set **x** and **t**, along with a new test point *x*
Goal: predict the value of *t*,
i.e. evaluate the predictive distribution *p*(*t* | *x*, **x**, **t**)

Assume that the parameters α and β are fixed and known in advance.
(In textbook, α = 5 x 10^{-3}, and β = 11.1)

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

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

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w}|\alpha)$$

LHS can be computed by normalizing the RHS.

Bayesian Curve Fitting

By some calculus,

$$p(t|x, \mathbf{x}, \mathbf{t}) = \int p(t|x, \mathbf{w}) p(\mathbf{w}|\mathbf{x}, \mathbf{t}) \, \mathrm{d}\mathbf{w} = \mathcal{N}\left(t | m(x), s^2(x)\right)$$

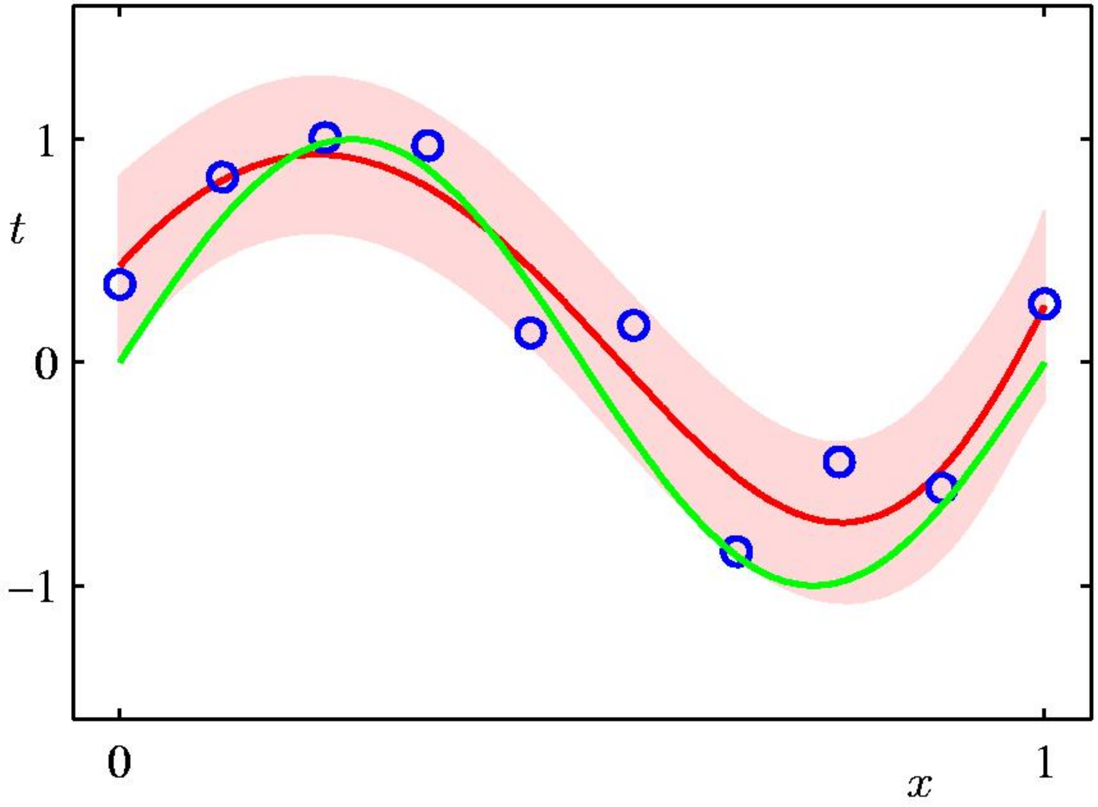
$$m(x) = \beta \boldsymbol{\phi}(x)^{\mathrm{T}} \mathbf{S} \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) t_n \qquad s^2(x) = \beta^{-1} + \boldsymbol{\phi}(x)^{\mathrm{T}} \mathbf{S} \boldsymbol{\phi}(x)$$

$$\mathbf{S}^{-1} = \underline{\alpha}\mathbf{I} + \underline{\beta} \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) \boldsymbol{\phi}(x_n)^{\mathrm{T}} \qquad \boldsymbol{\phi}(x_n) = \left(x_n^0, \ldots, x_n^M\right)^{\mathrm{T}}$$

devlaelen

Bayesian Predictive Distribution

$$p(t|x, \mathbf{x}, \mathbf{t}) = \mathcal{N}\left(t|m(x), s^2(x)\right)$$



The predictive distribution resulting from a Bayesian treatment of polynomial curve fitting using an M = 9 polynomial, with the fixed parameters α = 5x10^{-3} and β = 11.1 (corresponding to the known noise variance), in which the red curve denotes the mean of the predictive distribution and the red region corresponds to ± 1 standard deviation around the mean.