

Predictive distribution

Lecture 4

How to generate a new value of t ?
(predict)

$$P(t | \vec{t}, \alpha, \beta, \vec{X}) = \int d\vec{\omega} \underbrace{p(t | \vec{\omega}, \beta, \vec{x})}_{\text{likelihood}} \underbrace{p(\vec{\omega} | \vec{t}, \vec{X}, \alpha, \beta)}_{\text{posterior}}$$

training set \vec{t} , \vec{X}
 \vec{x} input corresponding to t

Here,

$$p(t | \vec{\omega}, \beta, \vec{x}) = \mathcal{N}(t | y(\vec{x}, \vec{\omega}), \beta^{-1})$$

$$p(\vec{\omega} | \vec{t}, \vec{X}, \alpha, \beta) = \mathcal{N}(\vec{\omega} | \vec{m}_N, S_N), \text{ where}$$

$$\begin{cases} \vec{m}_N = \beta S_N \Phi^T \vec{t}, \\ S_N = (\alpha \mathbf{I} + \beta \Phi^T \Phi)^{-1}. \end{cases}$$

In general, if

$$\rightarrow p(\vec{x}) = \mathcal{N}(\vec{x} | \vec{\mu}, \Lambda^{-1}),$$

$$\rightarrow p(\vec{y} | \vec{x}) = \mathcal{N}(\vec{y} | A\vec{x} + \vec{b}, L^{-1}).$$

$$p(\vec{y}) = \int d\vec{x} \underbrace{p(\vec{y} | \vec{x}) p(\vec{x})}_{p(\vec{y}, \vec{x})} = \mathcal{N}(\vec{y} | A\vec{\mu} + \vec{b}, L^{-1} + A\Lambda^{-1}A^T)$$

is also a gaussian

Here,

$$\textcircled{1D} \begin{cases} A\vec{x} + \vec{b} \Leftrightarrow y(\vec{x}, \vec{w}) = \underbrace{\vec{y}^T(\vec{x})}_{\text{"A}\vec{x}"} \vec{w} \\ L^{-1} \Leftrightarrow \beta^{-1} \end{cases}$$

linear in \vec{w} which is our f 's above (line \vec{x} in)

$\vec{y}^T(\vec{x}) \Leftrightarrow A$

$$\textcircled{\text{multi-D}} \begin{cases} \vec{\mu} \Leftrightarrow \vec{m}_N \\ S_N \Leftrightarrow \Lambda^{-1} \end{cases}$$

two indep. vars add up
"noise" due to prim. uncertainty
noise in data

We obtain

$$L^{-1} + A\Lambda^{-1}A^T \Rightarrow \beta^{-1} + \vec{y}^T(\vec{x}) S_N \vec{y}(\vec{x}) \equiv \sigma_N^2(\vec{x})$$

Further, the mean is given by

$$\vec{y}^T(\vec{x}) \vec{m}_N = \vec{m}_N^T \vec{y}(\vec{x}), \text{ s.t.}$$

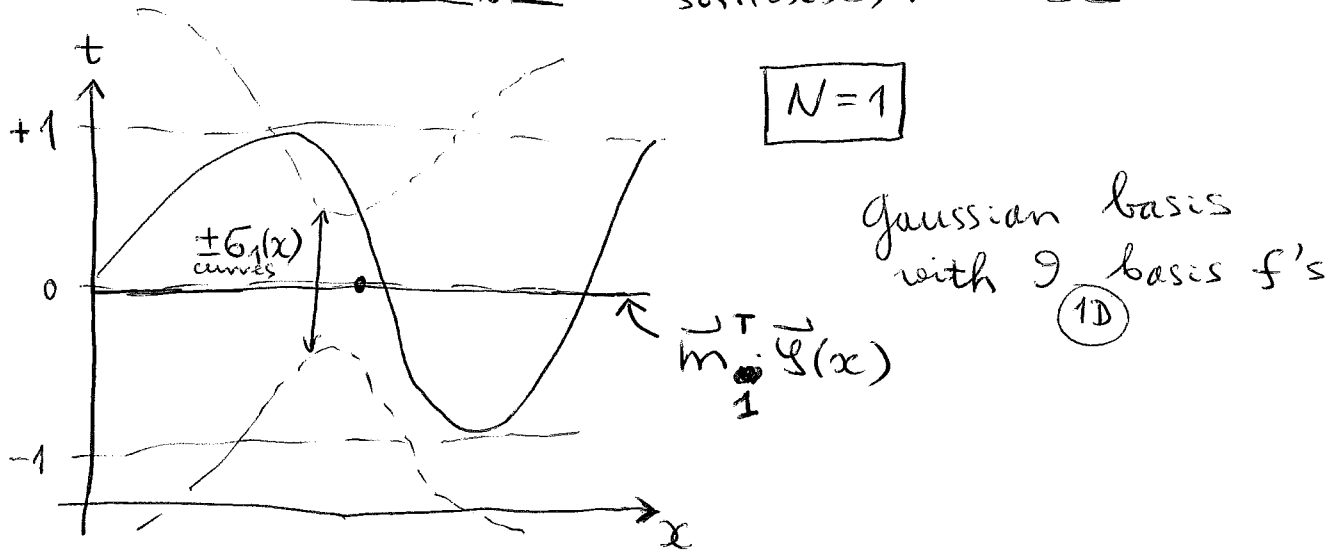
$$P(t|\vec{t}, \vec{X}, \vec{x}, \sigma, \beta) = \mathcal{N}(t | \underbrace{\vec{m}_N^T \vec{y}(\vec{x})}_{\text{current input}}, \underbrace{\sigma_N^2(\vec{x})}_{\text{training set}})$$

In the $N \rightarrow \infty$ limit, the posterior prob. for \vec{w} , $p(\vec{w} | \dots)$ should have vanishing variance: $S_N \rightarrow 0$ as $N \rightarrow \infty$.

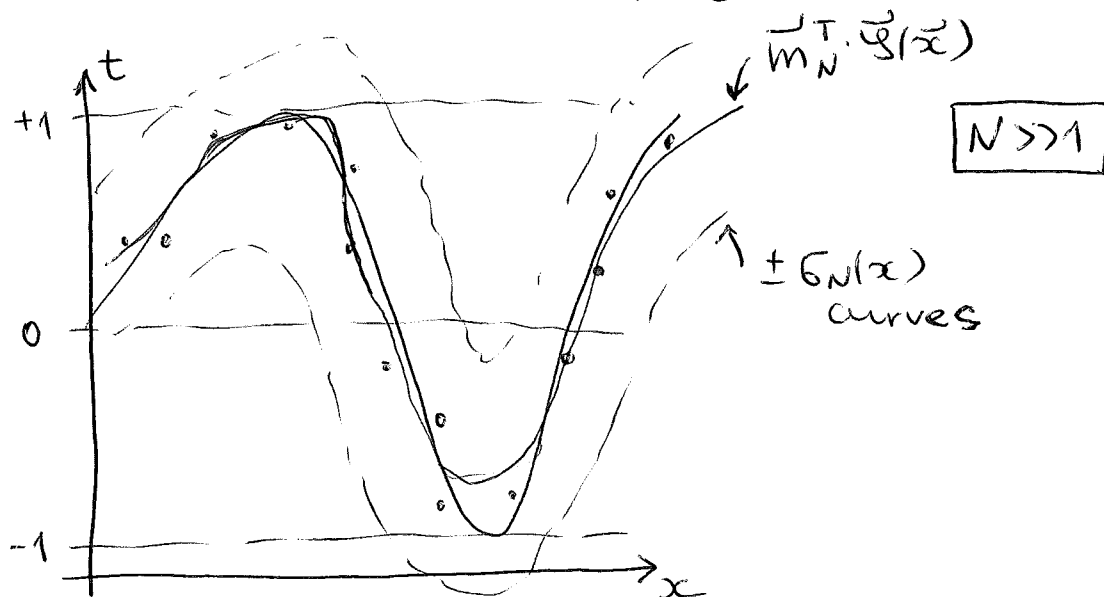
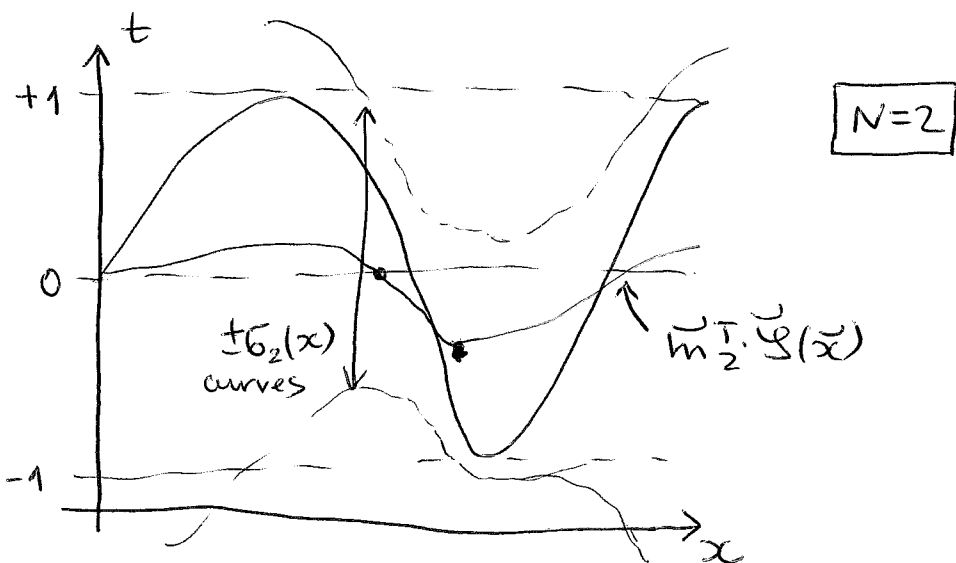
Then $\sigma_N^2(\vec{x}) \xrightarrow{N \rightarrow \infty} \beta^{-1}$, just noise in the data

$$\text{Indeed, } (S_N^{-1})_{ij} \sim \beta (\phi^T \phi)_{ij} = \beta \sum_{n=1}^N y_i(\vec{x}_n) y_j(\vec{x}_n) \uparrow \text{ as } N \uparrow$$

Consider our example, $\sin(2\pi x) + \text{noise}$ ^{gaussian}

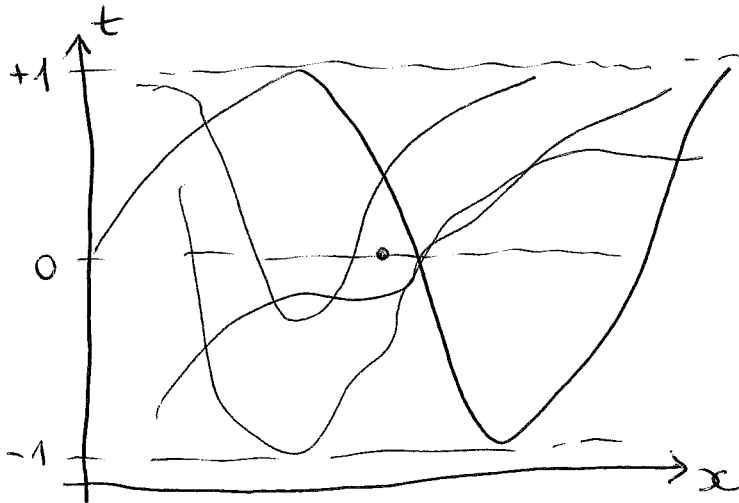


$\sigma_1(x)$ is smallest around the datapoint



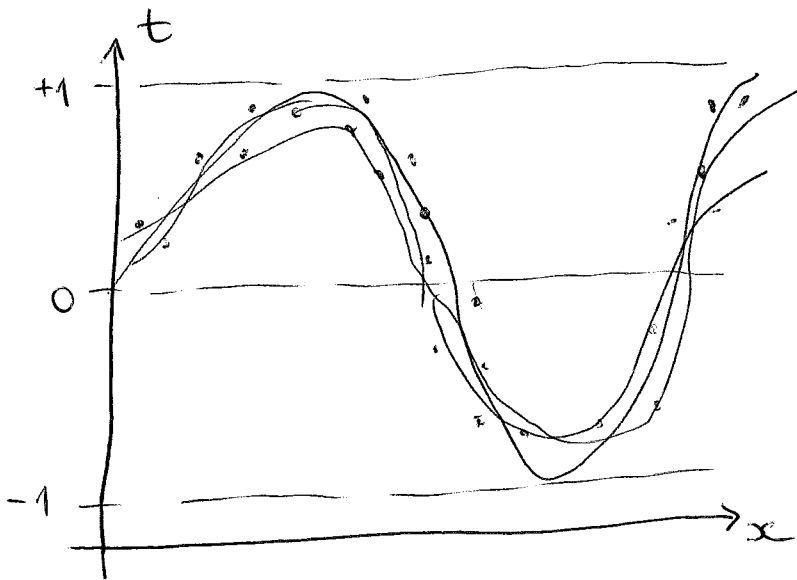
Note that far away from the localized basis f's ; $\sigma_N^2(\bar{x}) \rightarrow \beta^{-1}$ at any N , the model is too confident when extrapolating outside of the basis f's. Will be fixed later...

Now, consider sampling weights from the posterior distr'n $p(\vec{w} | \dots)$:



$N=1$

Functions are "attracted" to the datapoint but pretty random



$N \gg 1$

Functions "condensed" onto the datapoints

Equivalent kernel

Recall that the predictive mean

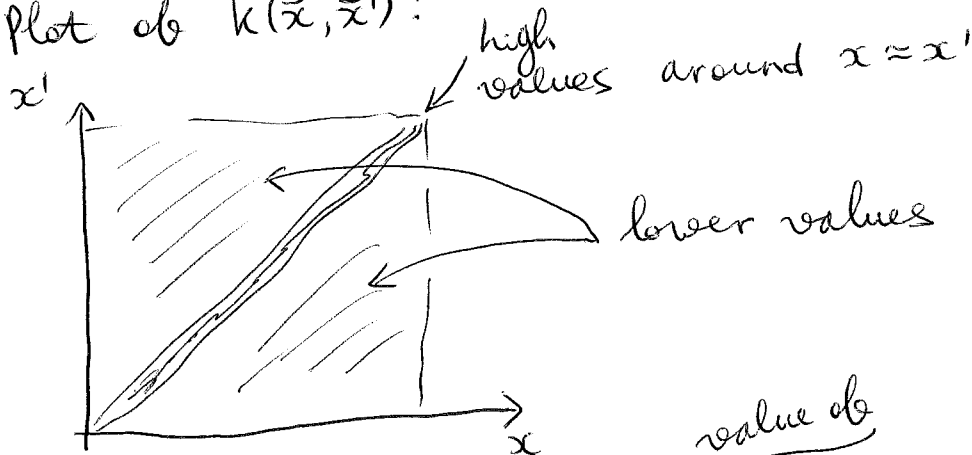
$$y(\vec{x}, \vec{m}_N) \equiv \underbrace{\vec{m}_N^T \vec{\Psi}(\vec{x})}_{\text{"}\vec{\Psi}^T \cdot \vec{m}_N\text{"}} = \beta \vec{\Psi}^T(\vec{x}) S_N \Phi^T \vec{t} =$$

$$= \beta \vec{\Psi}^T(\vec{x}) S_N \underbrace{\sum_{n=1}^N \vec{\Psi}(\vec{x}_n) t_n}_{\Phi^T \vec{t}} \equiv \sum_{n=1}^N k(\vec{x}, \vec{x}_n) t_n,$$

where $k(\vec{x}, \vec{x}') = \beta \vec{\Psi}^T(\vec{x}) S_N \vec{\Psi}(\vec{x}')$ is the equivalent kernel.

The predictive mean is therefore a linear combination of t_n 's.

Plot of $k(\vec{x}, \vec{x}')$:



So, $y(\vec{x}, \vec{m}_N)$ for a given \vec{x} will be dominated by terms in the $\sum_{n=1}^N$ sum for which $\vec{x}_n \approx \vec{x}$: "local evidence outweighs distant evidence".

Furthermore, consider

$$\text{cov}[y(\vec{x}, \vec{w}), y(\vec{x}', \vec{w})] = \text{cov}[\vec{\Psi}^T(\vec{x}) \vec{w}, \vec{w}^T \vec{\Psi}(\vec{x}')]$$

↑ ↑
same model
at \vec{x} & \vec{x}'

$$\textcircled{=} \vec{\Psi}^T(\vec{x}) \underbrace{\text{cov}[\vec{w}, \vec{w}^T]}_{\substack{\text{from posterior} \\ \text{distr'n for } \vec{w}, \text{ this} \\ \text{is } = S_N}} \vec{\Psi}(\vec{x}') = \vec{\Psi}^T(\vec{x}) S_N \vec{\Psi}(\vec{x}') = \beta^{-1} k(\vec{x}, \vec{x}')$$

So, $y(\vec{x}, \vec{w})$ [and in particular $\overline{y(\vec{x}, \vec{m}_N)}$ mean predictive model] will be highly correlated with $y(\vec{x}', \vec{w})$ if $\vec{x}' = \vec{x}$, and not otherwise.

Note that $k(\vec{x}, \vec{x}') = \Psi^T(\vec{x}) \Psi(\vec{x}')$, where $\Psi(\vec{x}) = \beta^{1/2} S_N^{1/2} \vec{\Psi}(\vec{x})$

Finally, one can argue that

$$\sum_{n=1}^N k(\vec{x}, \vec{x}_n) = 1, \quad \forall \vec{x} \quad (*)$$

Intuitively, consider

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

↑
predictive mean for $t_n=1, \forall n$

Clearly, $\bar{y}(\vec{x}) = 1$ with enough datapoints $\Rightarrow (*)$ follows.