Predictive distribution

{ Lecture 4 }

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

$$p(t|\vec{t}, \alpha, \beta, \vec{x}) = \int d\vec{w}\; p(t|\vec{w}, \beta, \vec{x})\; p(\vec{w}|\vec{t}, \vec{X}, \alpha, \beta)$$

- $\vec{t}$: training set
- $\vec{X}$: $\vec{x}$: input corresponding to t
- $p(t|\vec{w}, \beta, \vec{x})$: likelihood
- $p(\vec{w}|\vec{t}, \vec{X}, \alpha, \beta)$: posterior

HW#1: due 02/11
Bishop 3.4, 3.7, 3.12

Here,

$$\begin{cases} p(t|\vec{w}, \beta, \vec{x}) = \mathcal{N}(t|y(\vec{x}, \vec{w}), \beta^{-1}) \\ p(\vec{w}|\vec{t}, \vec{X}, \alpha, \beta) = \mathcal{N}(\vec{w}|\vec{m}_N, S_N), \text{ where} \end{cases}$$

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

In general, if (2.115)

$$\begin{cases} p(\vec{x}) = \mathcal{N}(\vec{x}|\vec{\mu}, \Lambda^{-1}), \\ p(\vec{y}|\vec{x}) = \mathcal{N}(\vec{y}|A\vec{x} + \vec{b}, L^{-1}). \end{cases}$$

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

is also a Gaussian

Here,

linear in $\vec{w}$ which is our f'las above (like $\vec{x}$ in)

(1D) $\begin{cases} A\vec{x}+\vec{b} \iff y(\vec{x},\vec{w}) = \vec{\varphi}^T(\vec{x})\,\vec{w} \\ L^{-1} \iff \beta^{-1} \end{cases}$

"$A\vec{x}$" ; "$\vec{b}$"$=0$ $\quad \vec{\varphi}^T(\vec{x}) \Leftrightarrow A$

(multi.D) $\begin{cases} \vec{\mu} \iff \vec{m}_N, \\ S_N \iff \Lambda^{-1} \end{cases}$

two indep. vars add up

noise in data ; "noise" due to prm. uncertainty

we obtain

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

Further, the mean is given by

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

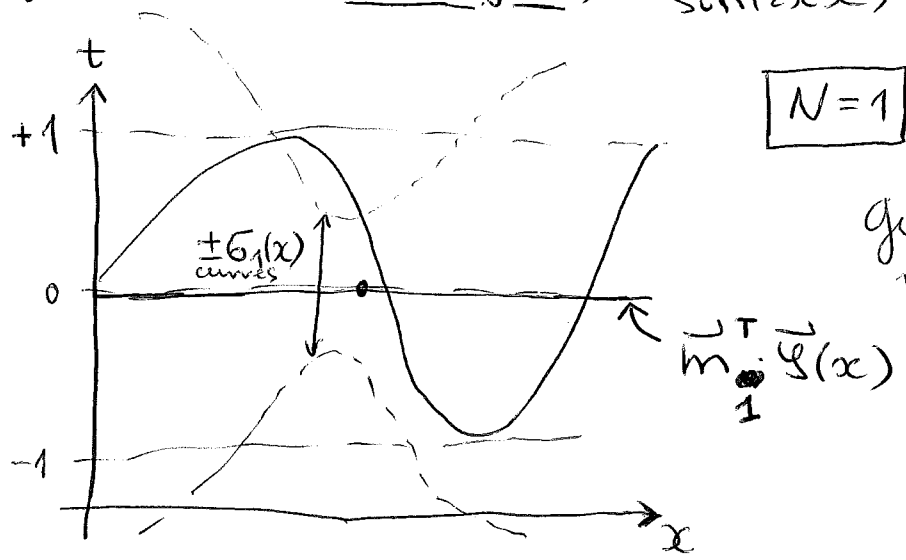
$$p(t|\underbrace{\vec{t},\vec{X}}_{\text{training set}},\underbrace{\vec{x}}_{\text{current input}},\alpha,\beta) = \mathcal{N}(t|\vec{m}_N^T\vec{\varphi}(\vec{x}),\sigma_N^2(\vec{x}))$$

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

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

Indeed, $\left(S_N^{-1}\right)_{ij} \sim \beta(\Phi^T\Phi)_{ij} = \beta\sum_{n=1}^{N}\varphi_i(\vec{x}_n)\varphi_j(\vec{x}_n) \nearrow$ as $N\nearrow$.

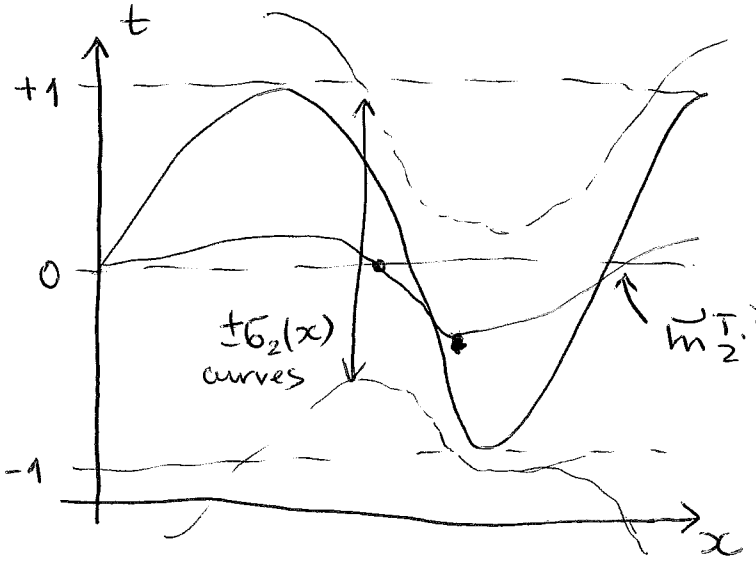
Consider our example, $\sin(2\pi x)$ + gaussian noise



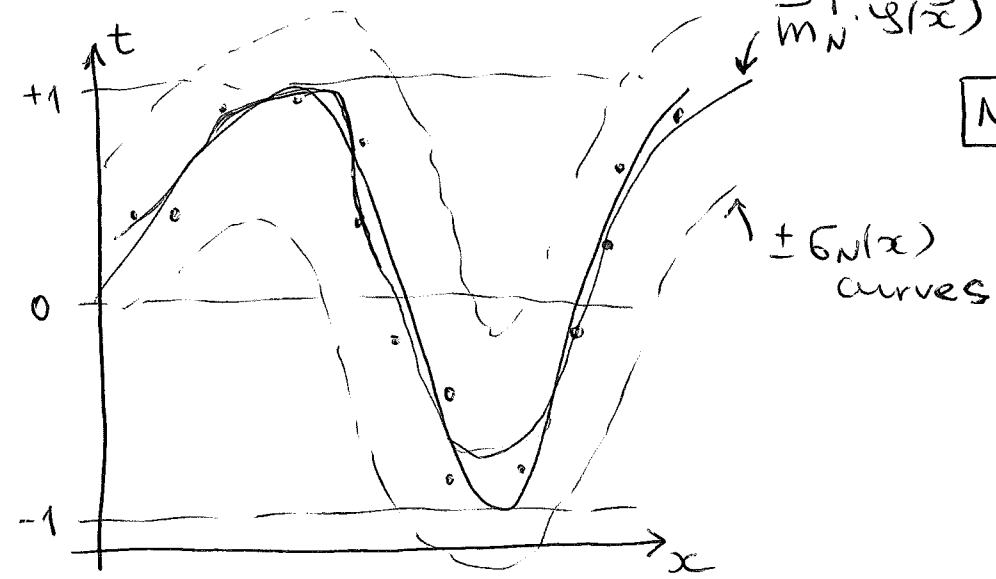
$N = 1$

Gaussian basis with 9 basis f's (1D)

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



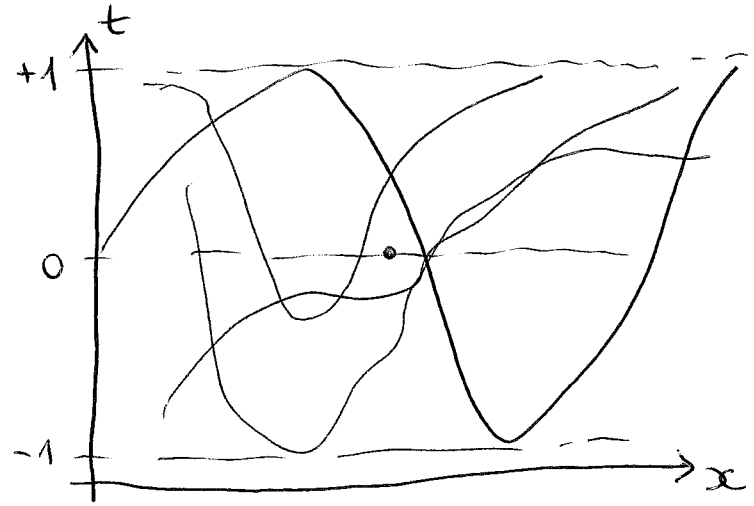
$N = 2$



$N >> 1$

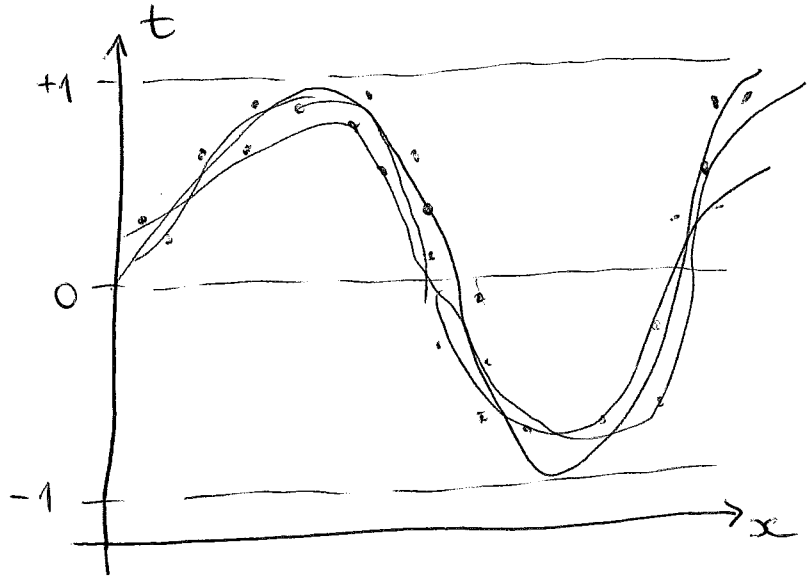
Note that far away from the localized basis f's; $\sigma_N^2(\vec{x}) \to \beta^{-1}$ at any N, the model is too confident when extrapolating outside of the basis f's. Will be fixed later...

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

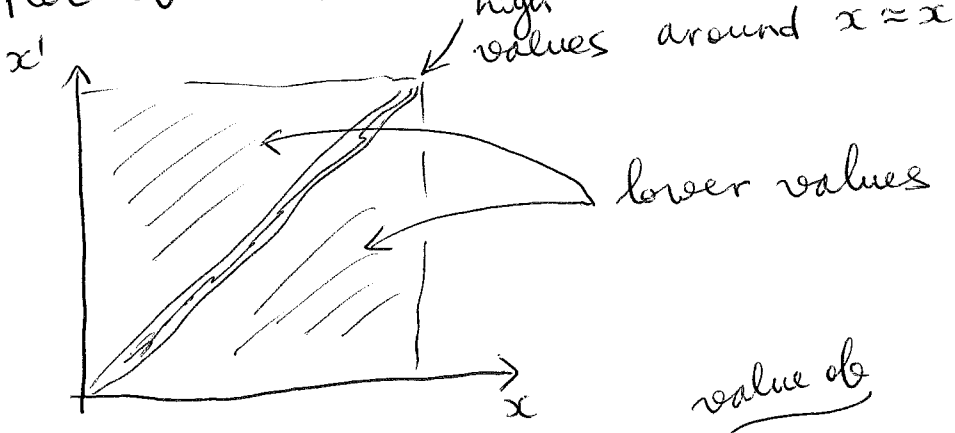
$$y(\vec{x}, \vec{m}_N) \equiv \underbrace{\vec{m}_N^T \cdot \vec{\varphi}(\vec{x})}_{\text{"}\vec{\varphi}^T \cdot \vec{m}_N} = \beta \vec{\varphi}^T(\vec{x}) S_N \Phi^T \vec{t} =$$

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

where $k(\vec{x}, \vec{x}') = \beta \vec{\varphi}^T(\vec{x}) S_N \vec{\varphi}(\vec{x}')$ 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 value of $\vec{x}$ 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

$$\mathrm{cov}[\,y(\vec{x},\vec{w}),\, y(\vec{x}',\vec{w})\,] = \mathrm{cov}[\,\vec{\varphi}^T(\vec{x})\,\vec{w},\, \vec{w}^T\vec{\varphi}(\vec{x}')\,] \ominus$$

($y(\vec{x},\vec{w})$: same model at $\vec{x}$ & $\vec{x}'$)

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

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 $\vec{x}' \simeq \vec{x}$, and not otherwise.

Note that $k(\vec{x},\vec{x}') = \psi^T(\vec{x})\,\psi(\vec{x}')$, where

$$\psi(\vec{x}) = \beta^{1/2} S_N^{1/2}\, \vec{\varphi}(\vec{x})$$

Finally, one can argue that

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

Intuitively, consider

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

($\bar{y}(\vec{x})$: predictive mean for $t_n = 1$, $\forall n$)

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