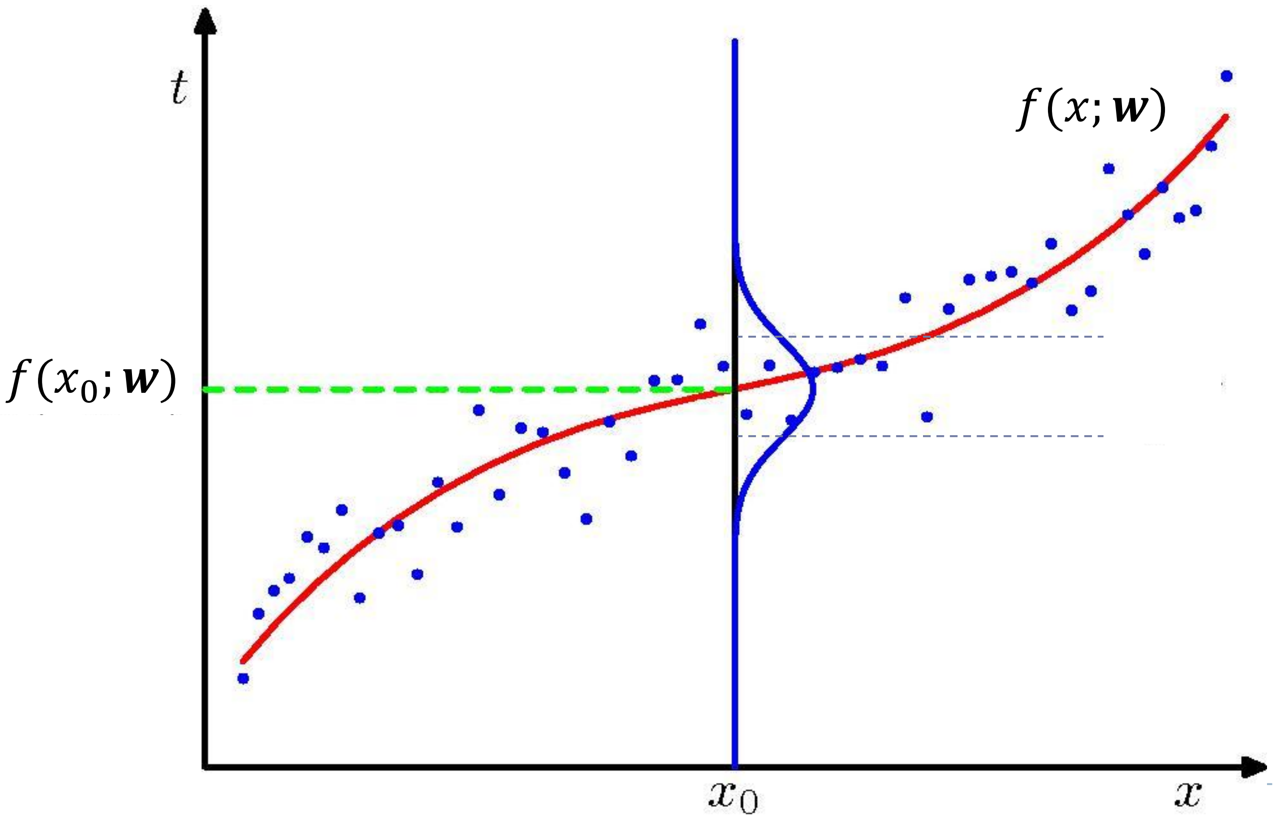
Regression and generalization

CE-717: Machine Learning
Sharif University of Technology

M. Soleymani
Fall 2016

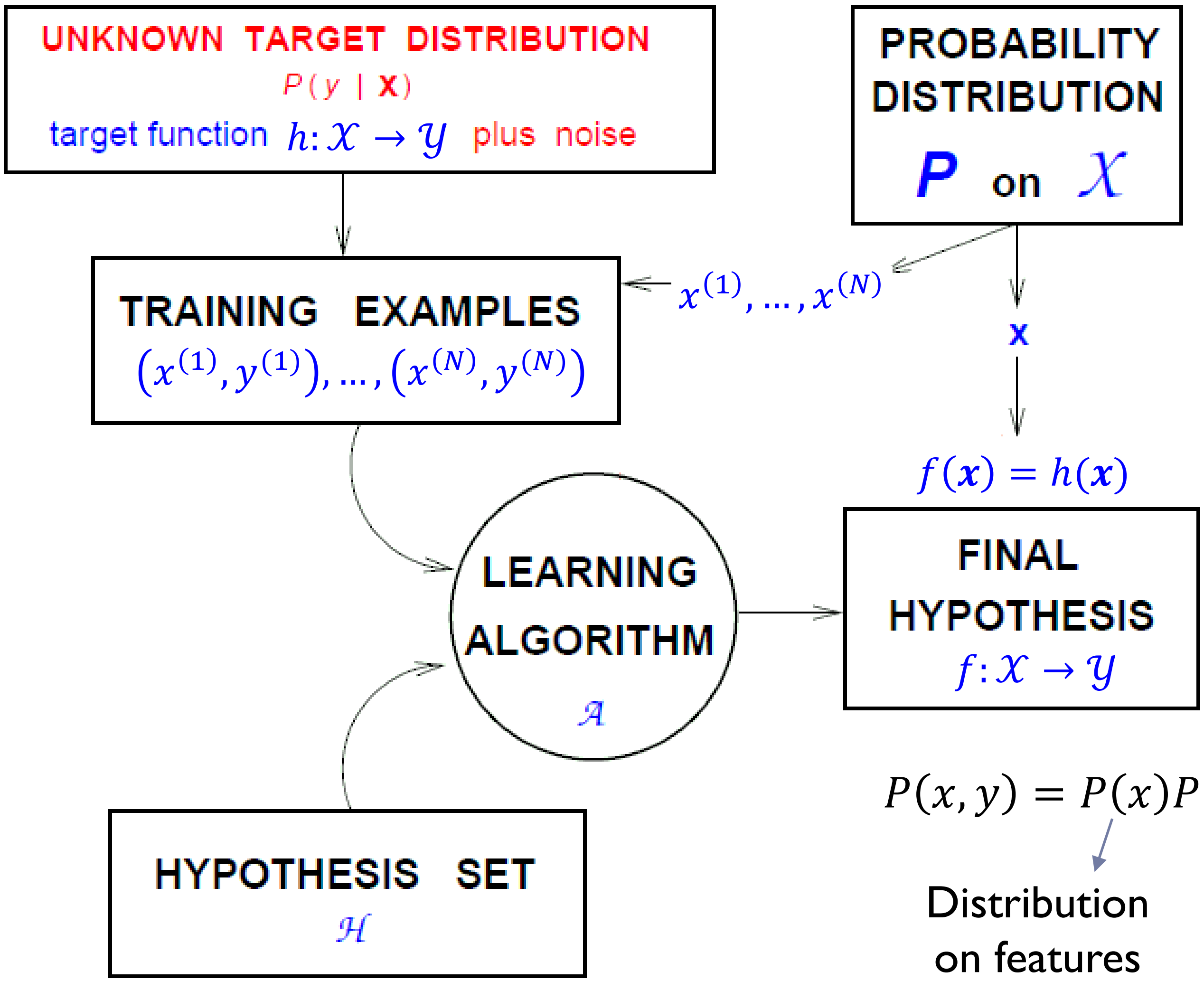
Curve fitting: probabilistic perspective

- Describing uncertainty over value of target variable as a probability distribution
- Example:



The learning diagram including noisy target

- Type



$$P(x, y) = P(x)P(y|x)$$

Distribution on features

Target distribution

[Y.S. Abou Mostafa, et. al]

Curve fitting: probabilistic perspective (Example)

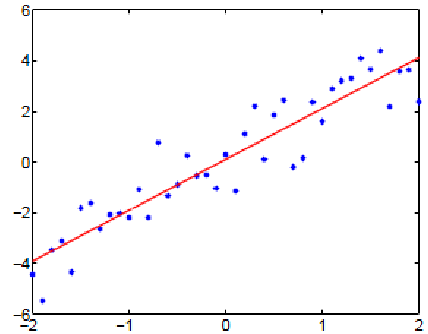
- Special case:

Observed output = function + noise

$$y = f(\boldsymbol{x}; \boldsymbol{w}) + \epsilon$$

e.g., $\epsilon \sim N(0, \sigma^2)$

- Noise: Whatever we cannot capture with our chosen family of functions



Curve fitting: probabilistic perspective (Example)

- Best regression

$$\mathbb{E}[y|\boldsymbol{x}] = E[f(\boldsymbol{x};\boldsymbol{w}) + \epsilon] = f(\boldsymbol{x};\boldsymbol{w})$$

$$\epsilon \sim N(0,\sigma^2)$$

- $f(\boldsymbol{x};\boldsymbol{w})$ is trying to capture the <u>mean of the observations y given the input $\boldsymbol{x}$</u>:

- $\mathbb{E}[y|\boldsymbol{x}]$: conditional expectation of y given $\boldsymbol{x}$
  - evaluated according to the model (not according to the underlying distribution P)

Curve fitting using probabilistic estimation

- Maximum Likelihood (ML) estimation
- Maximum A Posteriori (MAP) estimation
- Bayesian approach

Maximum likelihood estimation

- Given observations $\mathcal{D} = \left\{\left(\boldsymbol{x}^{(i)}, y^{(i)}\right)\right\}_{i=1}^{n}$
- Find the parameters that maximize the (conditional) likelihood of the outputs:

$$L(\mathcal{D}; \boldsymbol{\theta}) = p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}) = \prod_{i=1}^{n} p(y^{(i)}|\boldsymbol{x}^{(i)}, \boldsymbol{\theta})$$

$$\boldsymbol{y} = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix} \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_1^{(1)} & \cdots & x_d^{(1)} \\ 1 & x_1^{(2)} & \cdots & x_d^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(n)} & \cdots & x_d^{(n)} \end{bmatrix}$$

Maximum likelihood estimation (Cont'd)

$$y = f(\boldsymbol{x};\boldsymbol{w}) + \epsilon\ , \quad \epsilon \sim N(0,\sigma^2)$$

- y given $\boldsymbol{x}$ is normally distributed with mean $f(\boldsymbol{x};\boldsymbol{w})$ and variance σ^2:
  - we model the uncertainty in the predictions, not just the mean

$$p(y|\boldsymbol{x},\boldsymbol{w},\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{1}{2\sigma^2}\left(y - f(\boldsymbol{x};\boldsymbol{w})\right)^2\}$$

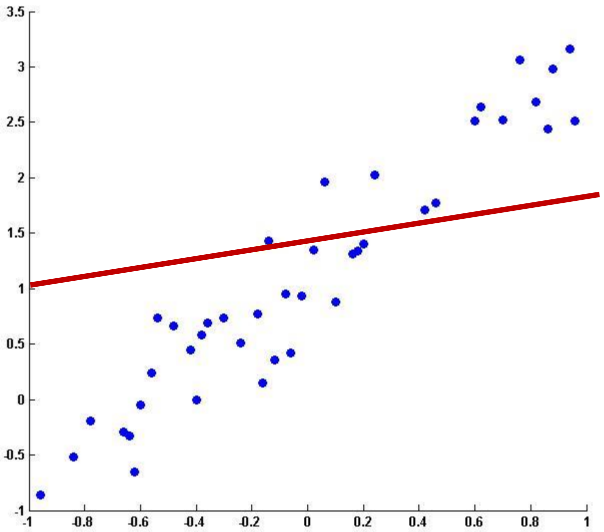
Maximum likelihood estimation (Cont'd)

- Example: univariate linear function

$$p(y|\boldsymbol{x}, \boldsymbol{w}, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2\sigma^2}(y - w_0 - w_1 x)^2\}$$

Why is this line a bad fit according to the likelihood criterion?

$p(y|\boldsymbol{x},\boldsymbol{w},\sigma^2)$ for most of the points will be near zero (as they are far from this line)



Maximum likelihood estimation (Cont'd)

- Maximize the likelihood of the outputs (i.i.d):

$$L(\mathcal{D}; \boldsymbol{w}, \sigma^2) = \prod_{i=1}^{n} p(y^{(i)} | \boldsymbol{x}^{(i)}, \boldsymbol{w}, \sigma^2)$$

$$\widehat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmax}}\, L(\mathcal{D}; \boldsymbol{w}, \sigma^2)$$

$$= \underset{\boldsymbol{w}}{\operatorname{argmax}} \prod_{i=1}^{n} p(y^{(i)} | \boldsymbol{x}^{(i)}, \boldsymbol{w}, \sigma^2)$$

Maximum likelihood estimation (Cont'd)

- It is often easier (but equivalent) to try to maximize the log-likelihood:

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmax}} \ln p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}, \sigma^2)$$

$$\ln \prod_{i=1}^{n} p(y^{(i)}|\boldsymbol{x}^{(i)}, \boldsymbol{w}, \sigma^2) = \sum_{i=1}^{n} \ln \mathcal{N}(y^{(i)}|\boldsymbol{x}^{(i)}, \boldsymbol{w}, \sigma^2)$$

$$= -n \ln \sigma - \frac{n}{2} \ln 2\pi - \frac{1}{2\sigma^2} \underbrace{\sum_{i=1}^{n} \left(y^{(i)} - f(\boldsymbol{x}^{(i)}; \boldsymbol{w})\right)^2}_{\text{sum of squares error}}$$

Maximum likelihood estimation (Cont'd)

- Maximizing log-likelihood (when we assume $y = f(\boldsymbol{x};\boldsymbol{w}) + \epsilon$, $\epsilon \sim N(0, \sigma^2)$) is equivalent to minimizing SSE

- Let $\widehat{\boldsymbol{w}}$ be the maximum likelihood (here least squares) setting of the parameters.

- What is the maximum likelihood estimate of σ^2?

$$\frac{\partial \log L(\mathcal{D};\boldsymbol{w},\sigma^2)}{\partial \sigma^2} = 0$$

$$\Rightarrow \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)} - f(\boldsymbol{x}^{(i)};\widehat{\boldsymbol{w}})\right)^2$$

Mean squared prediction error

Maximum likelihood estimation (Cont'd)

- Generally, maximizing log-likelihood is equivalent to minimizing empirical loss when the loss is defined according to:

$$Loss\left(y^{(i)}, f\left(\boldsymbol{x}^{(i)}, \boldsymbol{w}\right)\right) = -\ln\ p(y^{(i)}|\boldsymbol{x}^{(i)}, \boldsymbol{w}, \boldsymbol{\theta})$$

- Loss: negative log-probability
  - More general distributions for $p(y|\boldsymbol{x})$ can be considered

Maximum A Posterior (MAP) estimation

- MAP:
  - Given observations $\mathcal{D}$
  - Find the parameters that maximize the probabilities of the parameters after observing the data (posterior probabilities):

$$\boldsymbol{\theta}_{MAP} = \max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathcal{D}))$$

Since $p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$

$$\boldsymbol{\theta}_{MAP} = \max_{\boldsymbol{\theta}} \ p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

Maximum A Posterior (MAP) estimation

- Given observations $\mathcal{D} = \left\{\left(\boldsymbol{x}^{(i)}, y^{(i)}\right)\right\}_{i=1}^{n}$

$$\max_{\boldsymbol{w}} p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})p(\boldsymbol{w})$$

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{0}, \alpha^2 \boldsymbol{I}) = \left(\frac{1}{\sqrt{2\pi}\alpha}\right)^{d+1} exp\left\{-\frac{1}{2\alpha^2}\boldsymbol{w}^T\boldsymbol{w}\right\}$$

Maximum A Posterior (MAP) estimation

- Given observations $\mathcal{D} = \left\{\left(\boldsymbol{x}^{(i)}, y^{(i)}\right)\right\}_{i=1}^{n}$

$$\max_{\boldsymbol{w}} \ln p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}, \sigma^2)p(\boldsymbol{w})$$

$$\min_{\boldsymbol{w}} \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(y^{(i)} - f(\boldsymbol{x}^{(i)}; \boldsymbol{w})\right)^2 + \frac{1}{\alpha^2}\boldsymbol{w}^T\boldsymbol{w}$$

- Equivalent to regularized SSE with $\lambda = \frac{\sigma^2}{\alpha^2}$

Bayesian approach

- Given observations $\mathcal{D} = \left\{\left(\boldsymbol{x}^{(i)}, y^{(i)}\right)\right\}_{i=1}^{n}$
- Find the parameters that maximize the probabilities of observations

$$p(y|\boldsymbol{x}, \mathcal{D}) = \int p(y|\boldsymbol{w}, \boldsymbol{x})p(\boldsymbol{w}|\mathcal{D})d\boldsymbol{w}$$

- Example of prior distribution

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{0}, \alpha^2 \boldsymbol{I})$$

Bayesian approach

- Given observations $\mathcal{D} = \left\{\left(\boldsymbol{x}^{(i)}, y^{(i)}\right)\right\}_{i=1}^{N}$
- Find the parameters that maximize the probabilities of observations

$$p(\mathcal{D}|\boldsymbol{w}) = L(\mathcal{D}; \boldsymbol{w}, \boldsymbol{\theta}) = \prod_{i=1}^{N} p\left(y^{(i)} \middle| \boldsymbol{w}^T \boldsymbol{x}^{(i)}, \boldsymbol{\theta}\right)$$

$$p\left(y^{(i)} \middle| f\left(\boldsymbol{x}^{(i)}, \boldsymbol{w}\right), \boldsymbol{\theta}\right) = \mathcal{N}(y^{(i)}|\boldsymbol{w}^T \boldsymbol{x}^{(i)}, \sigma^2)$$

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{0}, \alpha^2 \boldsymbol{I})$$

$$p(\boldsymbol{w}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{w}) p(\boldsymbol{w})$$

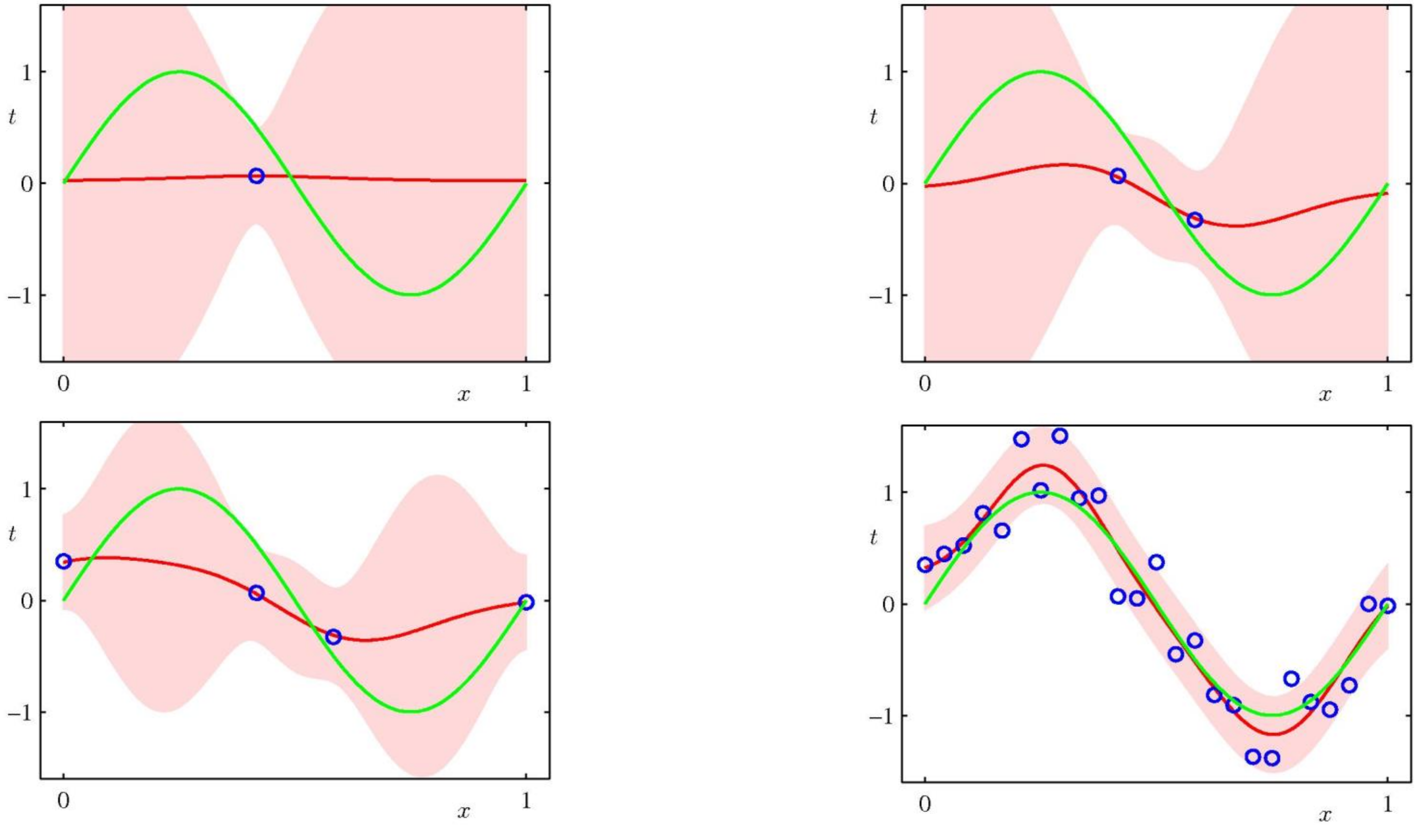
Predictive distribution

$$p(y|\boldsymbol{x}, \mathcal{D}) = \int p(y|\boldsymbol{w}, \boldsymbol{x}) p(\boldsymbol{w}|\mathcal{D}) d\boldsymbol{w}$$

$$p(y|\boldsymbol{x}, \mathcal{D}) = N(\boldsymbol{m}_N^T \boldsymbol{x}, \sigma_N^2(\boldsymbol{x}))$$

Predictive distribution: example

- Example: Sinusoidal data, 9 Gaussian basis functions

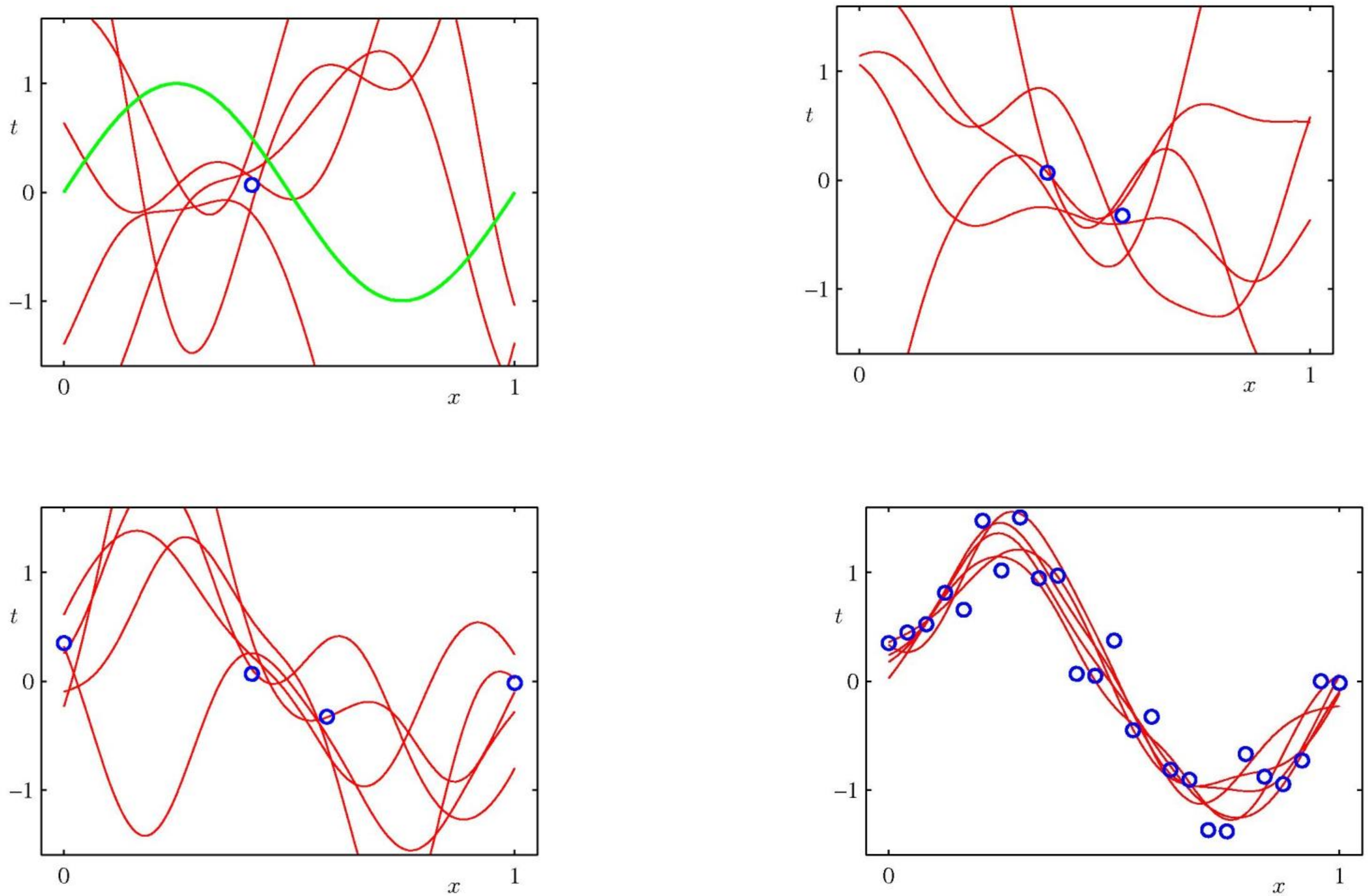


Red curve shows the mean of the predictive distribution
Pink region spans one standard deviation either side of the mean

[Bishop]

Predictive distribution: example

- Functions whose parameters are sampled from $p(\boldsymbol{w}|\mathcal{D})$



[Bishop]