# Bayesian Machine Learning

Evgeny Burnaev

Skoltech, Moscow, Russia





Burnaev, ML Skoltech

#### Outline

- Main Context
- 2 Reminder: Gaussian Distribution
- Bayesian Probability
- 4 Curve fitting re-visited
- 5 Linear Basis Function Models
- 6 Bayesian Linear Regression

Skoltech Stationa Institute all Science and Technology

- Main Context
- Reminder: Gaussian Distribution
- Bayesian Probability
- 4 Curve fitting re-visited
- 6 Linear Basis Function Models
- 6 Bayesian Linear Regression

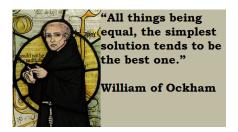
Burnaev, ML Skoltech

### Main Principles



Thomas Bayes (c. 1701 – 7 April 1761) was an English statistician, philosopher and Presbyterian minister

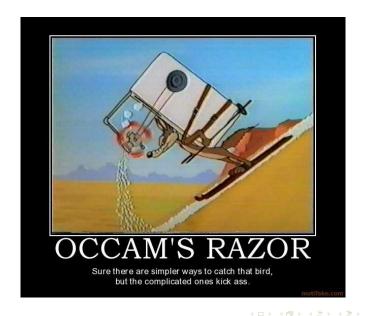
$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)}$$



William of Ockham (c. 1287 – 1347) was an English Franciscan friar and scholastic philosopher and theologian

4/43

Burnaev, ML Skottech



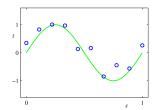


Figure - Plot of a training data

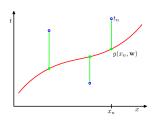


Figure – Residuals

• 
$$\mathcal{D}_m = \{\mathbf{X}_m, \mathbf{Y}_m\} = \{(x_i, y_i)\}_{i=1}^m$$
, where  $y_i = \sin(2\pi x_i) + \varepsilon_i$ ,  $\varepsilon_i$  is a Gaussian white noise

Surface britished of Science and Technology 6/43

## Example: Polynomial Curve Fitting

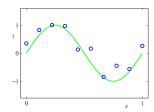


Figure – Plot of a training data

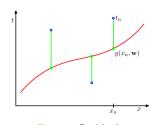


Figure – Residuals

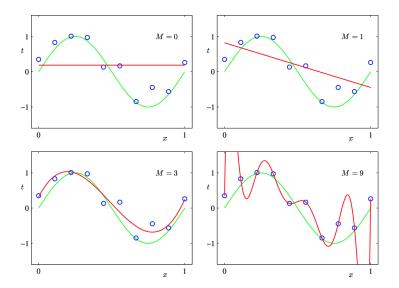
We fit a model

$$f(x, \mathbf{w}) = \sum_{j=0}^{M} w_j x^j,$$

by minimizing the error

$$E(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{m} \{f(x_i, \mathbf{w}) - y_i\}^2$$

# Plots of polynomials having various orders ${\cal M}$



koltech

8/43

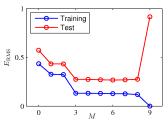


Figure –  $E_{RMS} = \sqrt{2E(\mathbf{w}^*)/n}$ 

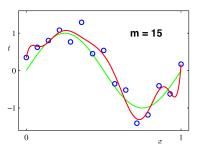
	M = 0	M = 1	M = 6	M = 9
$w_0^{\star}$	0.19	0.82	0.31	0.35
$w_1^{\star}$		-1.27	7.99	232.37
$w_2^{\star}$			-25.43	-5321.83
$w_3^{\star}$			17.37	48568.31
$w_4^{\star}$				-231639.30
$w_5^{\star}$				640042.26
$w_6^{\star}$				-1061800.52
$w_7^{\star}$				1042400.18
$w_8^{\star}$				-557682.99
$w_9^{\star}$				125201.43

Figure – Coefficients w\*

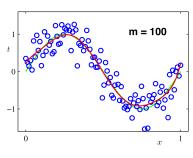
Skoltech

9/43

## Overfitting vs. Sample size



 $\mathsf{Figure} - M = 9, m = 15$ 



 ${\color{red}\mathsf{Figure}} - M = 9, m = 100$ 

Skoltech Statione trattate of Science and Technology 10/43

- $\bullet$  Limit the number of parameters M w.r.t. the size of the available training set?
- Instead choose the complexity of the model (effective model parameters)
   according to the complexity of the problem!

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{m} \{ f(x_i, \mathbf{w}) - y_i \}^2 + \frac{\lambda}{2} ||\mathbf{w}||^2$$

- ullet Limit the number of parameters M w.r.t. the size of the available training set?
- Instead choose the complexity of the model (effective model parameters) according to the complexity of the problem!

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{m} \{ f(x_i, \mathbf{w}) - y_i \}^2 + \frac{\lambda}{2} ||\mathbf{w}||^2$$

- $\bullet$  Limit the number of parameters M w.r.t. the size of the available training set?
- Instead choose the complexity of the model (effective model parameters) according to the complexity of the problem!

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{m} \{ f(x_i, \mathbf{w}) - y_i \}^2 + \frac{\lambda}{2} ||\mathbf{w}||^2$$

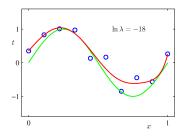


Figure –  $\lambda = e^{-18} \approx 0$ 

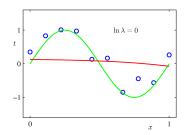


Figure –  $\lambda = 1$ 

	$\ln \lambda = -\infty$	$\ln \lambda = -18$	$\ln \lambda = 0$
$w_0^{\star}$	0.35	0.35	0.13
$w_1^{\star}$	232.37	4.74	-0.05
$w_2^{\star}$	-5321.83	-0.77	-0.06
$w_3^{\star}$	48568.31	-31.97	-0.05
$w_4^{\star}$	-231639.30	-3.89	-0.03
$w_5^{\star}$	640042.26	55.28	-0.02
$w_6^{\star}$	-1061800.52	41.32	-0.01
$w_7^{\star}$	1042400.18	-45.95	-0.00
$w_8^{\star}$	-557682.99	-91.53	0.00
$\widetilde{w_9^\star}$	125201.43	72.68	0.01

Figure – Dependence of  $\mathbf{w}^*$  on  $\lambda$ 

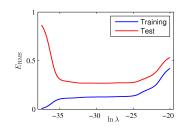


Figure – Dependence of  $E_{RMS}$  on  $\lambda$ 

- We would have to find a way to determine a suitable value for the model complexity!
- Hold-out set to select a model complexity (either M or  $\lambda$ )? Too wasteful  $\Rightarrow$  Bayesian Learning!

	$\ln \lambda = -\infty$	$\ln \lambda = -18$	$\ln \lambda = 0$
$w_0^{\star}$	0.35	0.35	0.13
$w_1^{\star}$	232.37	4.74	-0.05
$w_2^{\star}$	-5321.83	-0.77	-0.06
$w_3^{\star}$	48568.31	-31.97	-0.05
$w_4^{\star}$	-231639.30	-3.89	-0.03
$w_5^{\star}$	640042.26	55.28	-0.02
$w_6^{\star}$	-1061800.52	41.32	-0.01
$w_7^{\star}$	1042400.18	-45.95	-0.00
$w_8^{\star}$	-557682.99	-91.53	0.00
$\widetilde{w_9^\star}$	125201.43	72.68	0.01

Figure – Dependence of  $\mathbf{w}^*$  on  $\lambda$ 

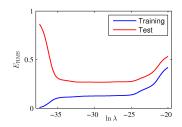


Figure – Dependence of  $E_{RMS}$  on  $\lambda$ 

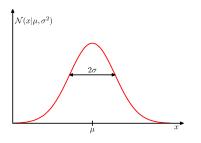
- We would have to find a way to determine a suitable value for the model complexity!
- Hold-out set to select a model complexity (either M or  $\lambda$ )? Too wasteful  $\Rightarrow$  Bayesian Learning!

- Main Context
- 2 Reminder: Gaussian Distribution
- Bayesian Probability
- 4 Curve fitting re-visited
- 6 Linear Basis Function Models
- 6 Bayesian Linear Regression

13/43

Burnaev, ML

### 1d Gaussian distribution



• Gaussian distribution of  $x \in \mathbb{R}^1$  with  $\mathbb{E}[x] = \mu$ ,  $\text{var}[x] = \sigma^2$ 

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

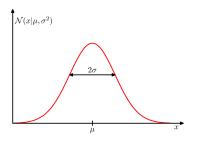
• Multivariate Gaussian distribution of  $\mathbf{x} \in \mathbb{R}^d$  with  $\mathbb{E}[\mathbf{x}] = \mu$   $\mathrm{cov}[\mathbf{x}] = \boldsymbol{\varSigma}$ 

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

←□ → ←□ → ← 글 → ← 글

Skoltech Institute of Science and Technology 14/43

### 1d Gaussian distribution



• Gaussian distribution of  $x \in \mathbb{R}^1$  with  $\mathbb{E}[x] = \mu$ ,  $\mathrm{var}[x] = \sigma^2$ 

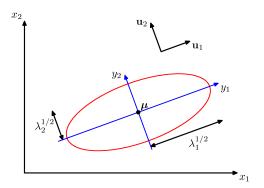
$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

• Multivariate Gaussian distribution of  $\mathbf{x} \in \mathbb{R}^d$  with  $\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$ ,  $\mathrm{cov}[\mathbf{x}] = \boldsymbol{\varSigma}$ 

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

14/43

Burnaev, ML Suban notice of Science and Technology

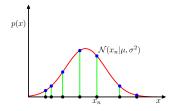


- The red curve shows the elliptical surface of constant probability density for  $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}),\ d=2$
- $\bullet$  Curve corresponds to the density  $\exp(-1/2)$  of its value at  $\mathbf{x}=\boldsymbol{\mu}$
- The major axes of the ellipse are defined by the eigenvectors  $\mathbf{u}_i$  of the covariance matrix  $\Sigma$ , with eigenvalues  $\lambda_i$

15/43

Burnaev, ML Substitute of Second Individual

#### Gaussian MLE



ullet Likelihood of an i.i.d. Gaussian sample  ${f X}_m=\{x_1,\ldots,x_m\}$ 

$$p(\mathbf{X}_m|\mu,\sigma^2) = \prod_{i=1}^m \mathcal{N}(x_i|\mu,\sigma^2)$$

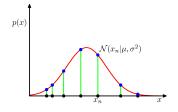
Log-likelihood is equal to

$$\log p(\mathbf{X}_m | \mu, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{i=1}^m (x_i - \mu)^2 - \frac{m}{2} \log \sigma^2 - \frac{m}{2} \log(2\pi) \to \max_{\mu, \sigma^2}$$

MLE is equal to

$$\mu_{ML} = \frac{1}{m} \sum_{i=1}^{m} x_i, \ \sigma_{ML}^2 = \frac{1}{m} \sum_{i=1}^{m} (x_i - \mu_{ML})$$

#### Gaussian MLE



• Likelihood of an i.i.d. Gaussian sample  $\mathbf{X}_m = \{x_1, \dots, x_m\}$ 

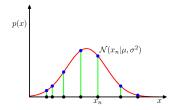
$$p(\mathbf{X}_m|\mu,\sigma^2) = \prod_{i=1}^m \mathcal{N}(x_i|\mu,\sigma^2)$$

Log-likelihood is equal to

$$\log p(\mathbf{X}_m | \mu, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{i=1}^m (x_i - \mu)^2 - \frac{m}{2} \log \sigma^2 - \frac{m}{2} \log(2\pi) \to \max_{\mu, \sigma^2}$$

$$\mu_{ML} = \frac{1}{m} \sum_{i=1}^{m} x_i, \ \sigma_{ML}^2 = \frac{1}{m} \sum_{i=1}^{m} (x_i - \mu_{ML})^2$$

#### Gaussian MLE



• Likelihood of an i.i.d. Gaussian sample  $\mathbf{X}_m = \{x_1, \dots, x_m\}$ 

$$p(\mathbf{X}_m|\mu,\sigma^2) = \prod_{i=1} \mathcal{N}(x_i|\mu,\sigma^2)$$

Log-likelihood is equal to

$$\log p(\mathbf{X}_m | \mu, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{i=1}^m (x_i - \mu)^2 - \frac{m}{2} \log \sigma^2 - \frac{m}{2} \log(2\pi) \to \max_{\mu, \sigma^2}$$

MLE is equal to

$$\mu_{ML} = \frac{1}{m} \sum_{i=1}^{m} x_i, \ \sigma_{ML}^2 = \frac{1}{m} \sum_{i=1}^{m} (x_i - \mu_{ML})^2$$

Burnaev, ML

- Main Context
- 2 Reminder: Gaussian Distribution
- Bayesian Probability
- 4 Curve fitting re-visited
- 5 Linear Basis Function Models
- 6 Bayesian Linear Regression

Burnaev, ML Skoltech

- Repeatable events ⇒ classical (frequentist) interpretation of probability
- Bayesian view: probabilities provide a quantification of uncertainty
- Consider an uncertain (non-repeatable) event
  - "whether the Arctic ice cap will have disappeared by the end of the century?"
  - we can generally have some idea how quickly we think the polar ice is melting
  - we obtain fresh data: e.g. from an Earth observation satellite we may revise our opinion on the rate of ice loss
  - we need to quantify our expression of uncertainty and make precise revisions of uncertainty in the light of new data

Skoltech
Skalkow Profit and all Science and Technology 18/43

### Uncertainty

- Repeatable events ⇒ classical (frequentist) interpretation of probability
- Bayesian view: probabilities provide a quantification of uncertainty
- Consider an uncertain (non-repeatable) event:
  - "whether the Arctic ice cap will have disappeared by the end of the century?"
  - we can generally have some idea how quickly we think the polar ice is melting
  - we obtain fresh data: e.g. from an Earth observation satellite we may revise our opinion on the rate of ice loss
    - we need to quantify our expression of uncertainty and make



18/43

Burnaev. ML Skoltech

- Repeatable events ⇒ classical (frequentist) interpretation of probability
- Bayesian view: probabilities provide a quantification of uncertainty
- Consider an uncertain (non-repeatable) event:
  - "whether the Arctic ice cap will have disappeared by the end of the century?"
  - we can generally have some idea how quickly we think the polar ice is melting
  - we obtain fresh data: e.g. from an Earth observation satellite we may revise our opinion on the rate of ice loss
  - we need to quantify our expression of uncertainty and make precise revisions of uncertainty in the light of new data

Skoltech
Station retrieve at Science and Technology 18/43

- Repeatable events ⇒ classical (frequentist) interpretation of probability
- Bayesian view: probabilities provide a quantification of uncertainty
- Consider an uncertain (non-repeatable) event:
  - "whether the Arctic ice cap will have disappeared by the end of the century?"
  - we can generally have some idea how quickly we think the polar ice is melting
  - we obtain fresh data: e.g. from an Earth observation satellite we may revise our opinion on the rate of ice loss
  - we need to quantify our expression of uncertainty and make precise revisions of uncertainty in the light of new data

- ullet Repeatable events  $\Rightarrow$  classical (frequentist) interpretation of probability
- Bayesian view: probabilities provide a quantification of uncertainty
- Consider an uncertain (non-repeatable) event:
  - "whether the Arctic ice cap will have disappeared by the end of the century?"
  - we can generally have some idea how quickly we think the polar ice is melting
  - we obtain fresh data: e.g. from an Earth observation satellite we may revise our opinion on the rate of ice loss
  - we need to quantify our expression of uncertainty and make precise revisions of uncertainty in the light of new data

- Data model:  $y = f(\mathbf{x}, \mathbf{w}) + \varepsilon$ ,  $\varepsilon$  is a noise
- Quantify uncertainty about model parameters w?
- Prior  $p(\mathbf{w})$  captures our assumptions about  $\mathbf{w}$  before observing the data!

- It is almost impossible to predict random rare events ⇒ their description is very long ⇒ complex
- w defines "complexity" of the model
- $-p(\mathbf{w})$  quantifies this complexity, as "small probability"  $\equiv$  "complex"



Figure – Kolmogorov A.N. (1903-1987)

- Data model:  $y = f(\mathbf{x}, \mathbf{w}) + \varepsilon$ ,  $\varepsilon$  is a noise
- Quantify uncertainty about model parameters w?
- Prior  $p(\mathbf{w})$  captures our assumptions about  $\mathbf{w}$  before observing the data!

- It is almost impossible to predict random rare events ⇒ their description is very long ⇒ complex
- w defines "complexity" of the model
- $-p(\mathbf{w})$  quantifies this complexity, as "small probability"  $\equiv$  "complex"



Figure – Kolmogorov A.N. (1903-1987)

- Data model:  $y = f(\mathbf{x}, \mathbf{w}) + \varepsilon$ ,  $\varepsilon$  is a noise
- Quantify uncertainty about model parameters w?
- $\bullet$  Prior  $p(\mathbf{w})$  captures our assumptions about  $\mathbf{w}$  before observing the data!

- It is almost impossible to predict random rare events ⇒ their description is very long ⇒ complex
- w defines "complexity" of the model
- $-p(\mathbf{w})$  quantifies this complexity, as "small probability"  $\equiv$  "complex"



Figure – Kolmogorov A.N. (1903-1987)

- Data model:  $y = f(\mathbf{x}, \mathbf{w}) + \varepsilon$ ,  $\varepsilon$  is a noise
- Quantify uncertainty about model parameters w?
- Prior  $p(\mathbf{w})$  captures our assumptions about  $\mathbf{w}$  before observing the data!

- It is almost impossible to predict random rare events ⇒ their description is very long ⇒ complex
- w defines "complexity" of the model
- $-p(\mathbf{w})$  quantifies this complexity, as "small probability"  $\equiv$  "complex"



Figure – Kolmogorov A.N. (1903-1987)

- Data model:  $y = f(\mathbf{x}, \mathbf{w}) + \varepsilon$ ,  $\varepsilon$  is a noise
- Quantify uncertainty about model parameters w?
- Prior  $p(\mathbf{w})$  captures our assumptions about  $\mathbf{w}$  before observing the data!

- It is almost impossible to predict random rare events ⇒ their description is very long ⇒ complex
- w defines "complexity" of the model
- $-p(\mathbf{w})$  quantifies this complexity, as "small probability"  $\equiv$  "complex"



Figure – Kolmogorov A.N. (1903-1987)

• Observed data  $\mathcal{D}_m = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$  influences the conditional probability  $p(\mathbf{w}|\mathcal{D}_m)$ :

$$p(\mathbf{w}|\mathcal{D}_m) = \frac{p(\mathcal{D}_m|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D}_m)}$$

- $p(\mathcal{D}_m|\mathbf{w})$  is a likelihood function (how probable the observed data set is for different settings of the parameter vector  $\mathbf{w}$ )
- Normalization constant (evidence)

$$p(\mathcal{D}_m) = \int p(\mathcal{D}_m | \mathbf{w}) p(\mathbf{w}) d\mathbf{w}$$

General form

posterior  $\sim$  likelihood  $\times$  prior

 $\log$  posterior  $~\sim~\log$  likelihood  $+\log$  prior

• Observed data  $\mathcal{D}_m = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$  influences the conditional probability  $p(\mathbf{w}|\mathcal{D}_m)$ :

$$p(\mathbf{w}|\mathcal{D}_m) = \frac{p(\mathcal{D}_m|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D}_m)}$$

- $p(\mathcal{D}_m|\mathbf{w})$  is a likelihood function (how probable the observed data set is for different settings of the parameter vector  $\mathbf{w}$ )
- Normalization constant (evidence)

$$p(\mathcal{D}_m) = \int p(\mathcal{D}_m | \mathbf{w}) p(\mathbf{w}) d\mathbf{w}$$

General form:

posterior  $\sim$  likelihood  $\times$  prior

 $\log$  posterior  $\,\sim\,\log$  likelihood  $\,+\,\log$  prior

• Observed data  $\mathcal{D}_m = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$  influences the conditional probability  $p(\mathbf{w}|\mathcal{D}_m)$ :

$$p(\mathbf{w}|\mathcal{D}_m) = \frac{p(\mathcal{D}_m|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D}_m)}$$

- $p(\mathcal{D}_m|\mathbf{w})$  is a likelihood function (how probable the observed data set is for different settings of the parameter vector  $\mathbf{w}$ )
- Normalization constant (evidence)

$$p(\mathcal{D}_m) = \int p(\mathcal{D}_m | \mathbf{w}) p(\mathbf{w}) d\mathbf{w}$$

• General form:

posterior  $\sim$  likelihood  $\times$  prior

 $\log$  posterior  $\,\sim\,\log$  likelihood  $+\log$  prior

### Example: curve fitting problem

• Observed data  $\mathcal{D}_m = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$  influences the conditional probability  $p(\mathbf{w}|\mathcal{D}_m)$ :

$$p(\mathbf{w}|\mathcal{D}_m) = \frac{p(\mathcal{D}_m|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D}_m)}$$

- $p(\mathcal{D}_m|\mathbf{w})$  is a likelihood function (how probable the observed data set is for different settings of the parameter vector  $\mathbf{w}$ )
- Normalization constant (evidence)

$$p(\mathcal{D}_m) = \int p(\mathcal{D}_m | \mathbf{w}) p(\mathbf{w}) d\mathbf{w}$$

• General form:

posterior ~ likelihood × prior

 $\log$  posterior  $\,\sim\,\log$  likelihood  $+\log$  prior

### Example: curve fitting problem

• Observed data  $\mathcal{D}_m = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$  influences the conditional probability  $p(\mathbf{w}|\mathcal{D}_m)$ :

$$p(\mathbf{w}|\mathcal{D}_m) = \frac{p(\mathcal{D}_m|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D}_m)}$$

- $p(\mathcal{D}_m|\mathbf{w})$  is a likelihood function (how probable the observed data set is for different settings of the parameter vector  $\mathbf{w}$ )
- Normalization constant (evidence)

$$p(\mathcal{D}_m) = \int p(\mathcal{D}_m | \mathbf{w}) p(\mathbf{w}) d\mathbf{w}$$

• General form:

posterior  $\sim$  likelihood  $\times$  prior

 $\log$  posterior  $\sim \log$  likelihood  $+\log$  prior

Statistics Institute at Science and Technology 20/43

### • Frequentist setting:

- $-\mathbf{w}$  is a fixed parameter,
- error bars on its estimates obtained by considering the distribution of possible data sets  $\mathcal{D}_m$
- Bayesian setting:
  - the uncertainty in the parameters is expressed through a probability distribution over w,
  - we reduce uncertainty about w by observing more and more data
- The inclusion of prior knowledge arises naturally

#### • Frequentist setting:

- $-\mathbf{w}$  is a fixed parameter,
- error bars on its estimates obtained by considering the distribution of possible data sets  $\mathcal{D}_m$

#### Bayesian setting:

- the uncertainty in the parameters is expressed through a probability distribution over w,
- we reduce uncertainty about  ${\bf w}$  by observing more and more data
- The inclusion of prior knowledge arises naturally

#### • Frequentist setting:

- $-\mathbf{w}$  is a fixed parameter,
- error bars on its estimates obtained by considering the distribution of possible data sets  $\mathcal{D}_m$

#### Bayesian setting:

- the uncertainty in the parameters is expressed through a probability distribution over  $\mathbf{w}$ ,
- we reduce uncertainty about  ${f w}$  by observing more and more data
- The inclusion of prior knowledge arises naturally

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \log p(\mathcal{D}_m | \mathbf{w})$$

- MAP (Maximum posterior) estimate
  - Posterior

$$p(\mathbf{w}|\mathcal{D}_m) = \frac{p(\mathcal{D}_m|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D}_m)}$$

— Maf

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} p(\mathbf{w}|\mathcal{D}_m)$$

- $\mathbf{w}^* = \arg\max_{\mathbf{z}} \log p(\mathbf{w}|\mathcal{D}_m)$
- MAP ≡ regularized MLE

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} [\log p(\mathcal{D}_m | \mathbf{w}) + \log p(\mathbf{w})]$$

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \log p(\mathcal{D}_m | \mathbf{w})$$

- MAP (Maximum posterior) estimate
  - Posterior

$$p(\mathbf{w}|\mathcal{D}_m) = \frac{p(\mathcal{D}_m|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D}_m)}$$

- MAP

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} p(\mathbf{w}|\mathcal{D}_m)$$

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \log p(\mathbf{w}|\mathcal{D}_m)$$

— MAP ≡ regularized MLE:

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} [\log p(\mathcal{D}_m | \mathbf{w}) + \log p(\mathbf{w})]$$

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \log p(\mathcal{D}_m | \mathbf{w})$$

- MAP (Maximum posterior) estimate
  - Posterior

$$p(\mathbf{w}|\mathcal{D}_m) = \frac{p(\mathcal{D}_m|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D}_m)}$$

- MAP

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} p(\mathbf{w}|\mathcal{D}_m)$$

$$\mathbf{w}^{\uparrow} = \arg\max_{\mathbf{w}} \log p(\mathbf{w}|\mathcal{D}_m)$$

- MAP  $\equiv$  regularized MLE:

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} [\log p(\mathcal{D}_m | \mathbf{w}) + \log p(\mathbf{w})]$$

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \log p(\mathcal{D}_m | \mathbf{w})$$

- MAP (Maximum posterior) estimate
  - Posterior

$$p(\mathbf{w}|\mathcal{D}_m) = \frac{p(\mathcal{D}_m|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D}_m)}$$

- MAP

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} p(\mathbf{w}|\mathcal{D}_m)$$

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \log p(\mathbf{w}|\mathcal{D}_m)$$

— MAP ≡ regularized MLE:

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} [\log p(\mathcal{D}_m | \mathbf{w}) + \log p(\mathbf{w})]$$

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \log p(\mathcal{D}_m | \mathbf{w})$$

- MAP (Maximum posterior) estimate
  - Posterior

$$p(\mathbf{w}|\mathcal{D}_m) = \frac{p(\mathcal{D}_m|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D}_m)}$$

— MAP

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} p(\mathbf{w}|\mathcal{D}_m)$$

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \log p(\mathbf{w}|\mathcal{D}_m)$$

— MAP ≡ regularized MLE:

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} [\log p(\mathcal{D}_m | \mathbf{w}) + \log p(\mathbf{w})]$$

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \log p(\mathcal{D}_m | \mathbf{w})$$

- MAP (Maximum posterior) estimate
  - Posterior

$$p(\mathbf{w}|\mathcal{D}_m) = \frac{p(\mathcal{D}_m|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D}_m)}$$

— MAP

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} p(\mathbf{w}|\mathcal{D}_m)$$

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \log p(\mathbf{w}|\mathcal{D}_m)$$

- MAP  $\equiv$  regularized MLE:

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} [\log p(\mathcal{D}_m | \mathbf{w}) + \log p(\mathbf{w})]$$

Skoltech

22/43

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \log p(\mathcal{D}_m | \mathbf{w})$$

- MAP (Maximum posterior) estimate
  - Posterior

$$p(\mathbf{w}|\mathcal{D}_m) = \frac{p(\mathcal{D}_m|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D}_m)}$$

MAP

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} p(\mathbf{w}|\mathcal{D}_m)$$

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \log p(\mathbf{w}|\mathcal{D}_m)$$

— MAP  $\equiv$  regularized MLE:

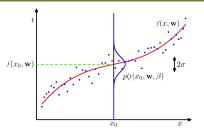
$$\mathbf{w}^* = \arg\max_{\mathbf{w}} [\log p(\mathcal{D}_m | \mathbf{w}) + \log p(\mathbf{w})]$$

Skoltech
Skilone treitzer et Science and Technology 22/43

- Main Context
- 2 Reminder: Gaussian Distribution
- Bayesian Probability
- 4 Curve fitting re-visited
- 6 Linear Basis Function Models
- 6 Bayesian Linear Regression

23/43

Burnaev, ML Skoltech



• Sample 
$$\mathcal{D}_m = \{\mathbf{X}_m, \mathbf{Y}_m\} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$$
, 
$$y_i = f(\mathbf{x}_i, \mathbf{w}) + \varepsilon_i, \text{ with i.i.d. } \varepsilon_i \sim \mathcal{N}(0, \beta^{-1})$$

Probabilistic model

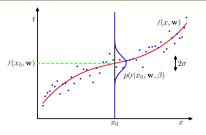
$$p(y|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(y|f(\mathbf{x}, \mathbf{w}), \beta^{-1})$$

where

- the mean is given by a polynomial  $f(\mathbf{x}, \mathbf{w})$
- the noise precision is given by the parameter  $\beta^{-1} = \sigma^2$
- Likelihood

$$p(\mathbf{Y}_m|\mathbf{X}_m, \mathbf{w}, \beta) = \prod_{i=1}^m \mathcal{N}(y_i|f(\mathbf{x}_i, \mathbf{w}), \beta^{-1})$$

Burnaev, ML Skoltec



• Sample 
$$\mathcal{D}_m = \{\mathbf{X}_m, \mathbf{Y}_m\} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$$
,  $y_i = f(\mathbf{x}_i, \mathbf{w}) + \varepsilon_i$ , with i.i.d.  $\varepsilon_i \sim \mathcal{N}(0, \beta^{-1})$ 

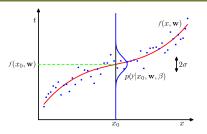
Probabilistic model

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

where

- the mean is given by a polynomial  $f(\mathbf{x}, \mathbf{w})$
- the noise precision is given by the parameter  $\beta^{-1} = \sigma^2$
- Likelihood

$$p(\mathbf{Y}_m|\mathbf{X}_m, \mathbf{w}, \beta) = \prod_{i=1}^m \mathcal{N}(y_i|f(\mathbf{x}_i, \mathbf{w}), \beta^{-1})$$



• Sample 
$$\mathcal{D}_m = \{\mathbf{X}_m, \mathbf{Y}_m\} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$$
,  $y_i = f(\mathbf{x}_i, \mathbf{w}) + \varepsilon_i$ , with i.i.d.  $\varepsilon_i \sim \mathcal{N}(0, \beta^{-1})$ 

Probabilistic model

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

where

- the mean is given by a polynomial  $f(\mathbf{x}, \mathbf{w})$
- the noise precision is given by the parameter  $\beta^{-1} = \sigma^2$
- Likelihood

$$p(\mathbf{Y}_m|\mathbf{X}_m, \mathbf{w}, \beta) = \prod_{i=1}^m \mathcal{N}(y_i|f(\mathbf{x}_i, \mathbf{w}), \beta^{-1})$$

Burnaev, ML Skoltec

Log-likelihood

$$\log p(\mathbf{Y}_m | \mathbf{X}_m, \mathbf{w}, \beta) = -\frac{\beta}{2} \sum_{i=1}^m (f(\mathbf{x}_i, \mathbf{w}) - y_i)^2 + \frac{m}{2} \log \beta - \frac{m}{2} (2\pi)$$

 $\bullet$  MLE of  $\beta$ 

$$\frac{1}{\beta_{ML}} = \frac{1}{m} \sum_{i=1}^{m} (f(\mathbf{x}_i, \mathbf{w}_{ML}) - y_i)^2$$

Predictive distribution

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

SKUILECII

Authors institute al Science and Technology 25/43

Log-likelihood

$$\log p(\mathbf{Y}_m | \mathbf{X}_m, \mathbf{w}, \beta) = -\frac{\beta}{2} \sum_{i=1}^m (f(\mathbf{x}_i, \mathbf{w}) - y_i)^2 + \frac{m}{2} \log \beta - \frac{m}{2} (2\pi)$$

 $\bullet$  MLE of  $\beta$ 

$$\frac{1}{\beta_{ML}} = \frac{1}{m} \sum_{i=1}^{m} (f(\mathbf{x}_i, \mathbf{w}_{ML}) - y_i)^2$$

Predictive distribution

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

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

Posterior

$$p(\mathbf{w}|\mathbf{X}_m, \mathbf{Y}_m, \alpha, \beta) \sim p(\mathbf{Y}_m|\mathbf{X}_m, \mathbf{w}, \beta) \cdot p(\mathbf{w}|\alpha)$$

Maximum posterior

 $\mathbf{w}^* = \arg\max_{\mathbf{w}} p(\mathbf{Y}_m | \mathbf{X}_m, \mathbf{w}, \beta) \cdot p(\mathbf{w} | \alpha)$ 

 $\mathbf{w}^* = \arg\max_{\mathbf{w}} [\log p(\mathbf{Y}_m | \mathbf{X}_m, \mathbf{w}, \beta) + \log p(\mathbf{w} | \alpha)]$ 

SKOITECII
stone tratase al Science and Technology 26

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

Posterior

$$p(\mathbf{w}|\mathbf{X}_m, \mathbf{Y}_m, \alpha, \beta) \sim p(\mathbf{Y}_m|\mathbf{X}_m, \mathbf{w}, \beta) \cdot p(\mathbf{w}|\alpha)$$

Maximum posterior

$$\mathbf{w}^* = \arg \max_{\mathbf{w}} p(\mathbf{Y}_m | \mathbf{X}_m, \mathbf{w}, \beta) \cdot p(\mathbf{w} | \alpha)$$
$$\mathbf{w}^* = \arg \max_{\mathbf{w}} \left[ \log p(\mathbf{Y}_m | \mathbf{X}_m, \mathbf{w}, \beta) + \log p(\mathbf{w} | \alpha) \right]$$

NOITECH Profitate el Science and Technology 26/43

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

Posterior

$$p(\mathbf{w}|\mathbf{X}_m, \mathbf{Y}_m, \alpha, \beta) \sim p(\mathbf{Y}_m|\mathbf{X}_m, \mathbf{w}, \beta) \cdot p(\mathbf{w}|\alpha)$$

Maximum posterior

$$\mathbf{w}^* = \arg \max_{\mathbf{w}} p(\mathbf{Y}_m | \mathbf{X}_m, \mathbf{w}, \beta) \cdot p(\mathbf{w} | \alpha)$$
$$\mathbf{w}^* = \arg \max_{\mathbf{w}} \left[ \log p(\mathbf{Y}_m | \mathbf{X}_m, \mathbf{w}, \beta) + \log p(\mathbf{w} | \alpha) \right]$$

tase of Science and Technology 26/43

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

Posterior

$$p(\mathbf{w}|\mathbf{X}_m, \mathbf{Y}_m, \alpha, \beta) \sim p(\mathbf{Y}_m|\mathbf{X}_m, \mathbf{w}, \beta) \cdot p(\mathbf{w}|\alpha)$$

Maximum posterior

$$\mathbf{w}^* = \arg \max_{\mathbf{w}} p(\mathbf{Y}_m | \mathbf{X}_m, \mathbf{w}, \beta) \cdot p(\mathbf{w} | \alpha)$$
$$\mathbf{w}^* = \arg \max_{\mathbf{w}} [\log p(\mathbf{Y}_m | \mathbf{X}_m, \mathbf{w}, \beta) + \log p(\mathbf{w} | \alpha)]$$

e at Science and Technology 26/43

Maximum posterior

$$\mathbf{w}^* = \arg \max_{\mathbf{w}} \left[ -\frac{\beta}{2} \sum_{i=1}^{m} (f(\mathbf{x}_i, \mathbf{w}) - y_i)^2 + \frac{m}{2} \log \frac{\beta}{2\pi} + \frac{\alpha}{2} \mathbf{w} \cdot \mathbf{w}^\top + \frac{(M+1)}{2} \log \frac{\alpha}{2\pi} \right]$$

Thus we get that

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \left[ \frac{\beta}{2} \sum_{i=1}^{n} (f(\mathbf{x}_i, \mathbf{w}) - y_i)^2 + \frac{\alpha}{2} \mathbf{w} \cdot \mathbf{w}^\top \right]$$

• MAP  $\equiv L_2$ -penalized regressions with  $\lambda = rac{a}{eta}$ 

NOITECH no treatage of Science and Technology 27/43 Maximum posterior

$$\mathbf{w}^* = \arg \max_{\mathbf{w}} \left[ -\frac{\beta}{2} \sum_{i=1}^{m} (f(\mathbf{x}_i, \mathbf{w}) - y_i)^2 + \frac{m}{2} \log \frac{\beta}{2\pi} + \frac{\alpha}{2} \mathbf{w} \cdot \mathbf{w}^\top + \frac{(M+1)}{2} \log \frac{\alpha}{2\pi} \right]$$

Thus we get that

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \left[ \frac{\beta}{2} \sum_{i=1}^{n} (f(\mathbf{x}_i, \mathbf{w}) - y_i)^2 + \frac{\alpha}{2} \mathbf{w} \cdot \mathbf{w}^\top \right]$$

ullet MAP  $\equiv L_2$ -penalized regressions with  $\lambda=rac{lpha}{eta}$ 

hose freezings and Technology 27.

Maximum posterior

$$\mathbf{w}^* = \arg \max_{\mathbf{w}} \left[ -\frac{\beta}{2} \sum_{i=1}^{m} (f(\mathbf{x}_i, \mathbf{w}) - y_i)^2 + \frac{m}{2} \log \frac{\beta}{2\pi} + \frac{\alpha}{2} \mathbf{w} \cdot \mathbf{w}^\top + \frac{(M+1)}{2} \log \frac{\alpha}{2\pi} \right]$$

Thus we get that

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \left[ \frac{\beta}{2} \sum_{i=1}^{n} (f(\mathbf{x}_i, \mathbf{w}) - y_i)^2 + \frac{\alpha}{2} \mathbf{w} \cdot \mathbf{w}^\top \right]$$

ullet MAP  $\equiv L_2$ -penalized regressions with  $\lambda=rac{lpha}{eta}$ 

e at Science and Technology 27/43

- Given the training data  $\mathbf{X}_m$  and  $\mathbf{Y}_m$ , and a new test point  $\mathbf{x}$ , our goal is to predict the value of y
- We would like to evaluate the predictive distribution  $p(y|\mathbf{x},\mathbf{X}_m,\mathbf{Y}_m)$
- The predictive distribution

$$p(y|\mathbf{x}, \mathbf{X}_m, \mathbf{Y}_m) = \int p(y|\mathbf{x}, \mathbf{w}) p(\mathbf{w}|\mathbf{X}_m, \mathbf{Y}_m) d\mathbf{w}$$

- Given the training data  $X_m$  and  $Y_m$ , and a new test point x, our goal is to predict the value of y
- ullet We would like to evaluate the predictive distribution  $p(y|\mathbf{x},\mathbf{X}_m,\mathbf{Y}_m)$
- The predictive distribution

$$p(y|\mathbf{x}, \mathbf{X}_m, \mathbf{Y}_m) = \int p(y|\mathbf{x}, \mathbf{w}) p(\mathbf{w}|\mathbf{X}_m, \mathbf{Y}_m) d\mathbf{w}$$

- Given the training data  $\mathbf{X}_m$  and  $\mathbf{Y}_m$ , and a new test point  $\mathbf{x}$ , our goal is to predict the value of y
- ullet We would like to evaluate the predictive distribution  $p(y|\mathbf{x},\mathbf{X}_m,\mathbf{Y}_m)$
- The predictive distribution

$$p(y|\mathbf{x}, \mathbf{X}_m, \mathbf{Y}_m) = \int p(y|\mathbf{x}, \mathbf{w}) p(\mathbf{w}|\mathbf{X}_m, \mathbf{Y}_m) d\mathbf{w}$$

Skoltech

28/43

- Given the training data  $X_m$  and  $Y_m$ , and a new test point x, our goal is to predict the value of y
- ullet We would like to evaluate the predictive distribution  $p(y|\mathbf{x},\mathbf{X}_m,\mathbf{Y}_m)$
- The predictive distribution

$$p(y|\mathbf{x}, \mathbf{X}_m, \mathbf{Y}_m) = \int p(y|\mathbf{x}, \mathbf{w}) p(\mathbf{w}|\mathbf{X}_m, \mathbf{Y}_m) d\mathbf{w}$$

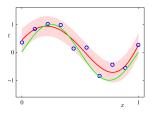


Figure – The predictive distribution for a polynomial with M=9, parameters  $\alpha=5\times 10^{-3}$  and  $\beta=11.1$  (known noise variance) are fixed

- Main Context
- 2 Reminder: Gaussian Distribution
- Bayesian Probability
- Curve fitting re-visited
- 5 Linear Basis Function Models
- Bayesian Linear Regression

Burnaev, ML Skoltech

Linear Basis Function Models

$$f(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w} \cdot \boldsymbol{\phi}(\mathbf{x})^{\top}$$

where  $\phi_j(\mathbf{x})$  are known basis functions

Typical basis functions

$$\phi_j(\mathbf{x}) = x_{j_1}^{j_0}, \ \phi_j(\mathbf{x}) = \exp\left\{-\frac{\|\mathbf{x} - \boldsymbol{\mu}_j\|^2}{2s^2}\right\},$$
$$\phi(\mathbf{x}) = \sigma\left(\boldsymbol{\mu}_{j,1} \cdot \mathbf{x}^\top + \boldsymbol{\mu}_{j,0}\right), \ \sigma(a) = \frac{1}{1 + \frac{1}{2s^2}}$$

 We assume that parameters of basis functions are fixed to some known values

Linear Basis Function Models

$$f(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w} \cdot \boldsymbol{\phi}(\mathbf{x})^{\top}$$

where  $\phi_i(\mathbf{x})$  are known basis functions

Typical basis functions

$$\phi_j(\mathbf{x}) = x_{j_1}^{j_0}, \ \phi_j(\mathbf{x}) = \exp\left\{-\frac{\|\mathbf{x} - \boldsymbol{\mu}_j\|^2}{2s^2}\right\},$$
$$\phi(\mathbf{x}) = \sigma\left(\boldsymbol{\mu}_{j,1} \cdot \mathbf{x}^\top + \mu_{j,0}\right), \ \sigma(a) = \frac{1}{1 + e^{-a}}$$

 We assume that parameters of basis functions are fixed to some known values

Skoltech
Stations for Example at Technology 30/43

Optimizing log-likelihood:

$$\mathbf{w}_{ML} = (\boldsymbol{\Phi}^{\top} \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^{\top} \mathbf{Y}_{m}, \quad \boldsymbol{\Phi} = \{(\boldsymbol{\phi}_{i}(\mathbf{x}_{j}))_{j=0}^{M-1}\}_{i=1}^{m}$$
$$\frac{1}{\beta_{ML}} = \frac{1}{m} \sum_{i=1}^{m} \{y_{i} - \mathbf{w}_{ML} \cdot \boldsymbol{\phi}(\mathbf{x}_{i})^{\top}\}^{2}$$

Regularized Least Squares

$$E_D(\mathbf{w}) + \lambda E_W(\mathbf{w}) \to \min_{\mathbf{w}}$$

$$\frac{1}{2} \sum_{i=1}^m \{ y_i - \mathbf{w} \cdot \boldsymbol{\phi}(\mathbf{x}_i)^\top \}^2 + \frac{\lambda}{2} \mathbf{w} \cdot \mathbf{w}^\top \to \min_{\mathbf{w}}$$

$$\mathbf{w}_{LS} = (\lambda \mathbf{I} + \boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^\top \mathbf{Y}_m$$

ロト (個) (注) (注) 注 の(の

Optimizing log-likelihood:

$$\mathbf{w}_{ML} = (\boldsymbol{\Phi}^{\top} \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^{\top} \mathbf{Y}_{m}, \quad \boldsymbol{\Phi} = \{(\boldsymbol{\phi}_{i}(\mathbf{x}_{j}))_{j=0}^{M-1}\}_{i=1}^{m}$$
$$\frac{1}{\beta_{ML}} = \frac{1}{m} \sum_{i=1}^{m} \{y_{i} - \mathbf{w}_{ML} \cdot \boldsymbol{\phi}(\mathbf{x}_{i})^{\top}\}^{2}$$

Regularized Least Squares

$$E_D(\mathbf{w}) + \lambda E_W(\mathbf{w}) \to \min_{\mathbf{w}}$$

$$\frac{1}{2} \sum_{i=1}^m \{ y_i - \mathbf{w} \cdot \boldsymbol{\phi}(\mathbf{x}_i)^\top \}^2 + \frac{\lambda}{2} \mathbf{w} \cdot \mathbf{w}^\top \to \min_{\mathbf{w}}$$

$$\mathbf{w}_{LS} = (\lambda \mathbf{I} + \boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^\top \mathbf{Y}_m$$

Skoltech
Stations Frotzer of Science and Technology 31/43

- Main Context
- 2 Reminder: Gaussian Distribution
- Bayesian Probability
- 4 Curve fitting re-visited
- 6 Linear Basis Function Models
- 6 Bayesian Linear Regression

Burnaev, ML Skoltech

Likelihood

$$p(\mathcal{D}_m|\mathbf{w}) = \prod_{i=1}^m \mathcal{N}(y_i|\mathbf{w} \cdot \phi(\mathbf{x}_i)^\top, \beta^{-1})$$

Thus the likelihood is Gaussian

$$p(\mathcal{D}_m|\mathbf{w}) = \mathcal{N}(\mathbf{Y}_m|\boldsymbol{\Phi}\cdot\mathbf{w}^\top, \beta^{-1}\mathbf{I})$$

The typical prior is Gaussian as well

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

Burnaev, ML 33/43

#### Parameter distribution

Likelihood

$$p(\mathcal{D}_m|\mathbf{w}) = \prod_{i=1}^m \mathcal{N}(y_i|\mathbf{w} \cdot \phi(\mathbf{x}_i)^\top, \beta^{-1})$$

• Thus the likelihood is Gaussian

$$p(\mathcal{D}_m|\mathbf{w}) = \mathcal{N}(\mathbf{Y}_m|\boldsymbol{\Phi}\cdot\mathbf{w}^\top, \beta^{-1}\mathbf{I})$$

The typical prior is Gaussian as well

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

soltane all Science and Technology 33/43

Likelihood

$$p(\mathcal{D}_m|\mathbf{w}) = \prod_{i=1}^m \mathcal{N}(y_i|\mathbf{w} \cdot \phi(\mathbf{x}_i)^\top, \beta^{-1})$$

• Thus the likelihood is Gaussian

$$p(\mathcal{D}_m|\mathbf{w}) = \mathcal{N}(\mathbf{Y}_m|\boldsymbol{\Phi}\cdot\mathbf{w}^\top, \beta^{-1}\mathbf{I})$$

The typical prior is Gaussian as well

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

state of Science and Technology 33/43

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$$
$$p(\mathbf{y}|\mathbf{z}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{z}, \mathbf{L}^{-1}),$$

we get that

$$p(\mathbf{z}|\mathbf{y}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\Sigma}\{\mathbf{A}^{\top}\mathbf{L}\mathbf{y} + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma})$$

where

$$\Sigma = (\Lambda + \mathbf{A}^{\mathsf{T}} \mathbf{L} \mathbf{A})^{-1}$$

Thus the posterior is defined by

$$p(\mathbf{w}|\mathcal{D}_m) = \mathcal{N}(\mathbf{w}|\boldsymbol{\omega}_m, \mathbf{S}_m)$$
$$\mathbf{S}_m = (\alpha^{-1}\mathbf{I} + \beta\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi})^{-}$$
$$\boldsymbol{\omega}_m = \beta\mathbf{S}_m\boldsymbol{\Phi}^{\top}\mathbf{Y}_m$$

Skoltech

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$$
$$(\mathbf{y}|\mathbf{z}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{z}, \mathbf{L}^{-1}),$$

we get that

$$p(\mathbf{z}|\mathbf{y}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\Sigma}\{\mathbf{A}^{\top}\mathbf{L}\mathbf{y} + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma})$$

where

$$\Sigma = (\Lambda + \mathbf{A}^{\mathsf{T}} \mathbf{L} \mathbf{A})^{-1}$$

Thus the posterior is defined by

$$p(\mathbf{w}|\mathcal{D}_m) = \mathcal{N}(\mathbf{w}|\boldsymbol{\omega}_m, \mathbf{S}_m)$$
$$\mathbf{S}_m = \left(\alpha^{-1}\mathbf{I} + \beta\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi}\right)^{-1}$$
$$\boldsymbol{\omega}_m = \beta\mathbf{S}_m\boldsymbol{\Phi}^{\top}\mathbf{Y}_m$$

Skoltech

34/43

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$$
$$p(\mathbf{y}|\mathbf{z}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{z}, \mathbf{L}^{-1}),$$

we get that

$$p(\mathbf{z}|\mathbf{y}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\Sigma}\{\mathbf{A}^{\top}\mathbf{L}\mathbf{y} + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma})$$

where

$$\boldsymbol{\varSigma} = (\boldsymbol{\varLambda} + \mathbf{A}^{\top} \mathbf{L} \mathbf{A})^{-1}$$

Thus the posterior is defined by

$$p(\mathbf{w}|\mathcal{D}_m) = \mathcal{N}(\mathbf{w}|\boldsymbol{\omega}_m, \mathbf{S}_m)$$
$$\mathbf{S}_m = \left(\alpha^{-1}\mathbf{I} + \beta\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi}\right)^{-1}$$
$$\boldsymbol{\omega}_m = \beta\mathbf{S}_m\boldsymbol{\Phi}^{\top}\mathbf{Y}_m$$

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$$
$$p(\mathbf{y}|\mathbf{z}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{z}, \mathbf{L}^{-1}),$$

we get that

$$p(\mathbf{z}|\mathbf{y}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\Sigma}\{\mathbf{A}^{\top}\mathbf{L}\mathbf{y} + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma}),$$

where

$$\Sigma = (\mathbf{\Lambda} + \mathbf{A}^{\mathsf{T}} \mathbf{L} \mathbf{A})^{-1}$$

Thus the posterior is defined by

$$p(\mathbf{w}|\mathcal{D}_m) = \mathcal{N}(\mathbf{w}|\boldsymbol{\omega}_m, \mathbf{S}_m)$$
$$\mathbf{S}_m = \left(\alpha^{-1}\mathbf{I} + \beta\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi}\right)^{-1}$$
$$\boldsymbol{\omega}_m = \beta\mathbf{S}_m\boldsymbol{\Phi}^{\top}\mathbf{Y}_m$$

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$$
$$p(\mathbf{y}|\mathbf{z}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{z}, \mathbf{L}^{-1}),$$

we get that

$$p(\mathbf{z}|\mathbf{y}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\Sigma}\{\mathbf{A}^{\top}\mathbf{L}\mathbf{y} + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma}),$$

where

$$\Sigma = (\mathbf{\Lambda} + \mathbf{A}^{\mathsf{T}} \mathbf{L} \mathbf{A})^{-1}$$

Thus the posterior is defined by

$$p(\mathbf{w}|\mathcal{D}_m) = \mathcal{N}(\mathbf{w}|\boldsymbol{\omega}_m, \mathbf{S}_m)$$
$$\mathbf{S}_m = (\alpha^{-1}\mathbf{I} + \beta\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi})^{-1}$$
$$\boldsymbol{\omega}_m = \beta\mathbf{S}_m\boldsymbol{\Phi}^{\top}\mathbf{Y}_m$$

SKOITECH Bulloon trettare of Science and Technology 34/43

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$$
$$p(\mathbf{y}|\mathbf{z}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{z}, \mathbf{L}^{-1}),$$

we get that

$$p(\mathbf{z}|\mathbf{y}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\Sigma}\{\mathbf{A}^{\top}\mathbf{L}\mathbf{y} + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma}),$$

where

$$\Sigma = (\mathbf{\Lambda} + \mathbf{A}^{\top} \mathbf{L} \mathbf{A})^{-1}$$

• Thus the posterior is defined by

$$p(\mathbf{w}|\mathcal{D}_m) = \mathcal{N}(\mathbf{w}|\boldsymbol{\omega}_m, \mathbf{S}_m)$$
$$\mathbf{S}_m = (\alpha^{-1}\mathbf{I} + \beta\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi})^{-1}$$
$$\boldsymbol{\omega}_m = \beta\mathbf{S}_m\boldsymbol{\Phi}^{\top}\mathbf{Y}_m$$

Skoltech

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$$
$$p(\mathbf{y}|\mathbf{z}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{z}, \mathbf{L}^{-1}),$$

we get that

$$p(\mathbf{z}|\mathbf{y}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\Sigma}\{\mathbf{A}^{\top}\mathbf{L}\mathbf{y} + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma}),$$

where

$$\Sigma = (\mathbf{\Lambda} + \mathbf{A}^{\mathsf{T}} \mathbf{L} \mathbf{A})^{-1}$$

• Thus the posterior is defined by

$$p(\mathbf{w}|\mathcal{D}_m) = \mathcal{N}(\mathbf{w}|\boldsymbol{\omega}_m, \mathbf{S}_m)$$
$$\mathbf{S}_m = (\alpha^{-1}\mathbf{I} + \beta\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi})^{-1}$$
$$\boldsymbol{\omega}_m = \beta\mathbf{S}_m\boldsymbol{\Phi}^{\top}\mathbf{Y}_m$$

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$$
$$p(\mathbf{y}|\mathbf{z}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{z}, \mathbf{L}^{-1}),$$

we get that

$$p(\mathbf{z}|\mathbf{y}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\Sigma}\{\mathbf{A}^{\top}\mathbf{L}\mathbf{y} + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma}),$$

where

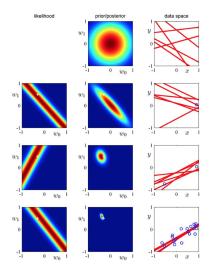
$$\Sigma = (\mathbf{\Lambda} + \mathbf{A}^{\mathsf{T}} \mathbf{L} \mathbf{A})^{-1}$$

Thus the posterior is defined by

$$p(\mathbf{w}|\mathcal{D}_m) = \mathcal{N}(\mathbf{w}|\boldsymbol{\omega}_m, \mathbf{S}_m)$$
$$\mathbf{S}_m = (\alpha^{-1}\mathbf{I} + \beta\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi})^{-1}$$
$$\boldsymbol{\omega}_m = \beta\mathbf{S}_m\boldsymbol{\Phi}^{\top}\mathbf{Y}_m$$

on trottage of Science and Technology 34/43

# Sequential Bayesian Learning



The Model  $f(x, \mathbf{w}) = w_0 + w_1 x$ 



Skoltech

• Make prediction of y for new value of x:

$$p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta) = \int p(y|\mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w}|\mathcal{D}_m, \alpha, \beta) d\mathbf{w}$$

• Since  $p(y|\mathbf{x}, \mathbf{w}, \beta)$  is Gaussian and the posterior  $p(\mathbf{w}|\mathcal{D}_m, \alpha, \beta)$  is Gaussian, then

$$p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta) = \mathcal{N}(y|\boldsymbol{\omega}_m \cdot \boldsymbol{\phi}(\mathbf{x})^\top, \sigma_m^2(\mathbf{x})),$$
$$\sigma_m^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^\top \mathbf{S}_m \boldsymbol{\phi}(\mathbf{x}),$$
$$\mathbf{S}_m = \left(\alpha^{-1} \mathbf{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi}\right)^{-1}$$

•  $p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta)$  depends on  $\alpha$  and  $\beta$ ! How to define them?  $\Rightarrow$  Full Bayesian approach!

Skoltech

• Make prediction of y for new value of x:

$$p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta) = \int p(y|\mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w}|\mathcal{D}_m, \alpha, \beta) d\mathbf{w}$$

• Since  $p(y|\mathbf{x}, \mathbf{w}, \beta)$  is Gaussian and the posterior  $p(\mathbf{w}|\mathcal{D}_m, \alpha, \beta)$  is Gaussian, then

$$p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta) = \mathcal{N}(y|\boldsymbol{\omega}_m \cdot \boldsymbol{\phi}(\mathbf{x})^\top, \sigma_m^2(\mathbf{x})),$$
$$\sigma_m^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^\top \mathbf{S}_m \boldsymbol{\phi}(\mathbf{x}),$$
$$\mathbf{S}_m = (\alpha^{-1}\mathbf{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1}$$

•  $p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta)$  depends on  $\alpha$  and  $\beta$ ! How to define them?  $\Rightarrow$  Full Bayesian approach!

Skoltech 36/43 Make prediction of y for new value of x:

$$p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta) = \int p(y|\mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w}|\mathcal{D}_m, \alpha, \beta) d\mathbf{w}$$

• Since  $p(y|\mathbf{x}, \mathbf{w}, \beta)$  is Gaussian and the posterior  $p(\mathbf{w}|\mathcal{D}_m, \alpha, \beta)$  is Gaussian, then

$$p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta) = \mathcal{N}(y|\boldsymbol{\omega}_m \cdot \boldsymbol{\phi}(\mathbf{x})^\top, \sigma_m^2(\mathbf{x})),$$
$$\sigma_m^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^\top \mathbf{S}_m \boldsymbol{\phi}(\mathbf{x}),$$
$$\mathbf{S}_m = (\alpha^{-1} \mathbf{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1}$$

•  $p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta)$  depends on  $\alpha$  and  $\beta$ ! How to define them?  $\Rightarrow$  Full Bayesian approach!

□ > 《□ > 《토 > 《토 > · 토 · · · 오○

Burnaev, ML Subsentitive distriction and facilities and facilities

• Make prediction of y for new value of x:

$$p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta) = \int p(y|\mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w}|\mathcal{D}_m, \alpha, \beta) d\mathbf{w}$$

• Since  $p(y|\mathbf{x}, \mathbf{w}, \beta)$  is Gaussian and the posterior  $p(\mathbf{w}|\mathcal{D}_m, \alpha, \beta)$  is Gaussian, then

$$p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta) = \mathcal{N}(y|\boldsymbol{\omega}_m \cdot \boldsymbol{\phi}(\mathbf{x})^\top, \sigma_m^2(\mathbf{x})),$$
$$\sigma_m^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^\top \mathbf{S}_m \boldsymbol{\phi}(\mathbf{x}),$$
$$\mathbf{S}_m = (\alpha^{-1} \mathbb{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1}$$

•  $p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta)$  depends on  $\alpha$  and  $\beta$ ! How to define them?  $\Rightarrow$  Full Bayesian approach!

ロ > 《御 > 《 差 > 《 差 > ~ 差 · の へ ②

Skoltech

36/43

• Make prediction of y for new value of x:

$$p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta) = \int p(y|\mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w}|\mathcal{D}_m, \alpha, \beta) d\mathbf{w}$$

• Since  $p(y|\mathbf{x}, \mathbf{w}, \beta)$  is Gaussian and the posterior  $p(\mathbf{w}|\mathcal{D}_m, \alpha, \beta)$  is Gaussian, then

$$p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta) = \mathcal{N}(y|\boldsymbol{\omega}_m \cdot \boldsymbol{\phi}(\mathbf{x})^\top, \sigma_m^2(\mathbf{x})),$$
$$\sigma_m^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^\top \mathbf{S}_m \boldsymbol{\phi}(\mathbf{x}),$$
$$\mathbf{S}_m = (\alpha^{-1}\mathbf{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1}$$

•  $p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta)$  depends on  $\alpha$  and  $\beta$ ! How to define them?  $\Rightarrow$  Full Bayesian approach!

ロ ト (日) (注) (注) 注 り()

Koltech

Make prediction of y for new value of x:

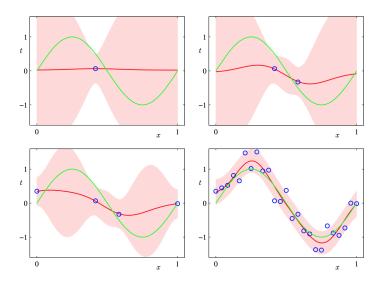
$$p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta) = \int p(y|\mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w}|\mathcal{D}_m, \alpha, \beta) d\mathbf{w}$$

• Since  $p(y|\mathbf{x}, \mathbf{w}, \beta)$  is Gaussian and the posterior  $p(\mathbf{w}|\mathcal{D}_m, \alpha, \beta)$  is Gaussian, then

$$p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta) = \mathcal{N}(y|\boldsymbol{\omega}_m \cdot \boldsymbol{\phi}(\mathbf{x})^\top, \sigma_m^2(\mathbf{x})),$$
$$\sigma_m^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^\top \mathbf{S}_m \boldsymbol{\phi}(\mathbf{x}),$$
$$\mathbf{S}_m = (\alpha^{-1}\mathbf{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1}$$

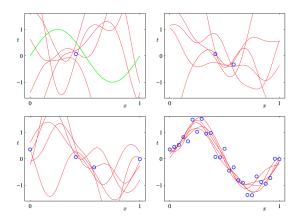
•  $p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta)$  depends on  $\alpha$  and  $\beta$ ! How to define them?  $\Rightarrow$  Full Bayesian approach!

Burnaev, ML Subar vitas of fices until street and subar vitas of fices



M=9 Gaussian functions





Plots of  $f(\mathbf{x}, \mathbf{w})$  using samples from the posterior distributions over  $\mathbf{w} \sim p(\mathbf{w}|\mathcal{D}_m, \alpha, \beta)$  for some  $\alpha$  and  $\beta$ 

Skoltech

$$p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta) = \int p(y|\mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w}|\mathcal{D}_m, \alpha, \beta) d\mathbf{w}$$
$$p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta) = \mathcal{N}(y|\boldsymbol{\omega}_m \cdot \boldsymbol{\phi}(\mathbf{x})^\top, \sigma_m^2(\mathbf{x})),$$
$$\sigma_m^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^\top \mathbf{S}_m \boldsymbol{\phi}(\mathbf{x}),$$
$$\mathbf{S}_m = \left(\alpha^{-1}\mathbf{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi}\right)^{-1}$$

•  $p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta)$  depends on  $\alpha$  and  $\beta$ ! We introduce hyperpriors over  $\alpha$  and  $\beta$ !

ㅁ▶ ◀畵▶ ◀불▶ ◀불▶ 불 쒸٩♡

Skoltech
Sations from and Technology 39/43

$$p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta) = \int p(y|\mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w}|\mathcal{D}_m, \alpha, \beta) d\mathbf{w}$$

$$p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta) = \mathcal{N}(y|\boldsymbol{\omega}_m \cdot \boldsymbol{\phi}(\mathbf{x})^\top, \sigma_m^2(\mathbf{x})),$$

$$\sigma_m^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^\top \mathbf{S}_m \boldsymbol{\phi}(\mathbf{x}),$$

$$\mathbf{S}_m = (\alpha^{-1}\mathbf{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1}$$

•  $p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta)$  depends on  $\alpha$  and  $\beta$ ! We introduce hyperpriors over  $\alpha$  and  $\beta$ !

ロト (個) (注) (注) 注 の(の

Skoltech
Sudhous brokkers at Science and Technology 39/43

$$p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta) = \int p(y|\mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w}|\mathcal{D}_m, \alpha, \beta) d\mathbf{w}$$

$$p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta) = \mathcal{N}(y|\boldsymbol{\omega}_m \cdot \boldsymbol{\phi}(\mathbf{x})^\top, \sigma_m^2(\mathbf{x})),$$

$$\sigma_m^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^\top \mathbf{S}_m \boldsymbol{\phi}(\mathbf{x}),$$

$$\mathbf{S}_m = (\alpha^{-1} \mathbf{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1}$$

•  $p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta)$  depends on  $\alpha$  and  $\beta$ ! We introduce hyperpriors over  $\alpha$  and  $\beta$ !

ㅁㅏ ◀畵ㅏ ◀불ㅏ 4 불ㅏ / 월 / / / / / /

Skoltech

$$p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta) = \int p(y|\mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w}|\mathcal{D}_m, \alpha, \beta) d\mathbf{w}$$

$$p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta) = \mathcal{N}(y|\boldsymbol{\omega}_m \cdot \boldsymbol{\phi}(\mathbf{x})^\top, \sigma_m^2(\mathbf{x})),$$

$$\sigma_m^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^\top \mathbf{S}_m \boldsymbol{\phi}(\mathbf{x}),$$

$$\mathbf{S}_m = (\alpha^{-1} \mathbf{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1}$$

•  $p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta)$  depends on  $\alpha$  and  $\beta$ ! We introduce hyperpriors over  $\alpha$  and  $\beta$ !

ロト (個) (注) (注) 注 の(の

Skoltech
Stations institute at Science and Technology 39/43

$$p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta) = \int p(y|\mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w}|\mathcal{D}_m, \alpha, \beta) d\mathbf{w}$$

$$p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta) = \mathcal{N}(y|\boldsymbol{\omega}_m \cdot \boldsymbol{\phi}(\mathbf{x})^\top, \sigma_m^2(\mathbf{x})),$$

$$\sigma_m^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^\top \mathbf{S}_m \boldsymbol{\phi}(\mathbf{x}),$$

$$\mathbf{S}_m = (\alpha^{-1}\mathbf{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1}$$

$$p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta) = \int p(y|\mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w}|\mathcal{D}_m, \alpha, \beta) d\mathbf{w}$$
$$p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta) = \mathcal{N}(y|\boldsymbol{\omega}_m \cdot \boldsymbol{\phi}(\mathbf{x})^\top, \sigma_m^2(\mathbf{x})),$$
$$\sigma_m^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^\top \mathbf{S}_m \boldsymbol{\phi}(\mathbf{x}),$$
$$\mathbf{S}_m = (\alpha^{-1}\mathbf{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1}$$

•  $p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta)$  depends on  $\alpha$  and  $\beta$ ! We introduce hyperpriors over  $\alpha$  and  $\beta$ !

Skoltech
Suddoos Institute at Science and Technology 39/43

ullet We introduce hyperpriors over lpha and eta

$$p(y|\mathbf{x}, \mathcal{D}_m) = \int \int \int p(y|\mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w}|\mathcal{D}_m, \alpha, \beta) p(\alpha, \beta|\mathcal{D}_m) d\mathbf{w} d\alpha d\beta$$

- We assume that the posterior distribution  $p(\alpha, \beta | \mathcal{D}_m)$  is sharply peaked around values  $\widehat{\alpha}$  and  $\widehat{\beta}$
- Then we simply marginalize over  $\mathbf{w}$ , where  $\alpha$  and  $\beta$  are fixed to the values  $\widehat{\alpha}$  and  $\widehat{\beta}$ , so that

$$p(y|\mathbf{x}, \mathcal{D}_m) \approx p(y|\mathbf{x}, \mathcal{D}_m, \widehat{\alpha}, \widehat{\beta}) = \int p(y|\mathbf{x}, \mathbf{w}, \widehat{\beta}) p(\mathbf{w}|\mathcal{D}_m, \widehat{\alpha}, \widehat{\beta}) d\mathbf{w}$$

ㅁ▶ ◀畵▶ ◀불▶ ◀불▶ 불 쒸٩♡

Burnaev, ML Skoltech

ullet We introduce hyperpriors over lpha and eta

$$p(y|\mathbf{x}, \mathcal{D}_m) = \int \int \int p(y|\mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w}|\mathcal{D}_m, \alpha, \beta) p(\alpha, \beta|\mathcal{D}_m) d\mathbf{w} d\alpha d\beta$$

- We assume that the posterior distribution  $p(\alpha,\beta|\mathcal{D}_m)$  is sharply peaked around values  $\widehat{\alpha}$  and  $\widehat{\beta}$
- Then we simply marginalize over  $\mathbf{w}$ , where  $\alpha$  and  $\beta$  are fixed to the values  $\widehat{\alpha}$  and  $\widehat{\beta}$ , so that

$$p(y|\mathbf{x}, \mathcal{D}_m) \approx p(y|\mathbf{x}, \mathcal{D}_m, \widehat{\alpha}, \widehat{\beta}) = \int p(y|\mathbf{x}, \mathbf{w}, \widehat{\beta}) p(\mathbf{w}|\mathcal{D}_m, \widehat{\alpha}, \widehat{\beta}) d\mathbf{w}$$

Burnaev, ML Subsection of Control

40/43

ullet We introduce hyperpriors over lpha and eta

$$p(y|\mathbf{x}, \mathcal{D}_m) = \int \int \int p(y|\mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w}|\mathcal{D}_m, \alpha, \beta) p(\alpha, \beta|\mathcal{D}_m) d\mathbf{w} d\alpha d\beta$$

- We assume that the posterior distribution  $p(\alpha, \beta | \mathcal{D}_m)$  is sharply peaked around values  $\widehat{\alpha}$  and  $\widehat{\beta}$
- Then we simply marginalize over  $\mathbf{w}$ , where  $\alpha$  and  $\beta$  are fixed to the values  $\widehat{\alpha}$  and  $\widehat{\beta}$ , so that

$$p(y|\mathbf{x}, \mathcal{D}_m) \approx p(y|\mathbf{x}, \mathcal{D}_m, \widehat{\alpha}, \widehat{\beta}) = \int p(y|\mathbf{x}, \mathbf{w}, \widehat{\beta}) p(\mathbf{w}|\mathcal{D}_m, \widehat{\alpha}, \widehat{\beta}) d\mathbf{w}$$

□ > 《□ > 《토 > 《토 > · 토 · · · 오○

Burnaev, ML

40/43

ullet The posterior for lpha and eta is given by

$$p(\alpha, \beta | \mathcal{D}_m) \sim p(\mathcal{D}_m | \alpha, \beta) \cdot p(\alpha, \beta)$$

• If the prior  $p(\alpha, \beta)$  is relatively flat, then in the evidence framework

$$(\widehat{\alpha}, \widehat{\beta}) = \arg \max_{\alpha, \beta} p(\mathcal{D}_m | \alpha, \beta)$$

• To obtain  $(\widehat{\alpha}, \widehat{\beta})$  iterative optimization is used

(ロ) (部) (注) (注) 注 り(C)

Burnaev, ML SKOITECT 41/43

• The posterior for  $\alpha$  and  $\beta$  is given by

$$p(\alpha, \beta | \mathcal{D}_m) \sim p(\mathcal{D}_m | \alpha, \beta) \cdot p(\alpha, \beta)$$

ullet If the prior  $p(\alpha, \beta)$  is relatively flat, then in the evidence framework

$$(\widehat{\alpha}, \widehat{\beta}) = \arg \max_{\alpha, \beta} p(\mathcal{D}_m | \alpha, \beta)$$

• To obtain  $(\widehat{\alpha}, \widehat{\beta})$  iterative optimization is used

ロ ト (回) (注) (注) 注 り()

ullet The posterior for lpha and eta is given by

$$p(\alpha, \beta | \mathcal{D}_m) \sim p(\mathcal{D}_m | \alpha, \beta) \cdot p(\alpha, \beta)$$

• If the prior  $p(\alpha, \beta)$  is relatively flat, then in the evidence framework

$$(\widehat{\alpha}, \widehat{\beta}) = \arg \max_{\alpha, \beta} p(\mathcal{D}_m | \alpha, \beta)$$

 $\bullet$  To obtain  $(\widehat{\alpha},\widehat{\beta})$  iterative optimization is used

Skoltech

• Let us calculate the evidence for  $(\alpha, \beta)$ 

$$p(\mathcal{D}_m|\alpha,\beta) = \int p(\mathcal{D}_m|\mathbf{w},\beta)p(\mathbf{w}|\alpha)d\mathbf{w}$$

• Let us denote by  $E(\mathbf{w})$  the sum of the fit and the regularization on coefficients  $\mathbf{w}$ 

$$E(\mathbf{w}) = \beta E_D(\beta) + \alpha E_W(\mathbf{w}) = \frac{\beta}{2} \|\mathbf{Y}_m - \boldsymbol{\Phi} \cdot \mathbf{w}^\top\|^2 + \frac{\alpha}{2} \mathbf{w} \cdot \mathbf{w}^\top$$

• Since  $p(\mathcal{D}_m|\mathbf{w},\beta)$  and  $p(\mathbf{w}|\alpha)$  are Gaussians with quadratic forms  $E_D(\beta)$  and  $E_W(\mathbf{w})$ , we get that

$$p(\mathcal{D}_m | \alpha, \beta) = \left(\frac{\beta}{2\pi}\right)^{m/2} \left(\frac{\alpha}{2\pi}\right)^{M/2} \int \exp\{-E(\mathbf{w})\} d\mathbf{w}$$

Burnaev, ML 42/43

• Let us calculate the evidence for  $(\alpha, \beta)$ 

$$p(\mathcal{D}_m|\alpha,\beta) = \int p(\mathcal{D}_m|\mathbf{w},\beta)p(\mathbf{w}|\alpha)d\mathbf{w}$$

 $\bullet$  Let us denote by  $E(\mathbf{w})$  the sum of the fit and the regularization on coefficients  $\mathbf{w}$ 

$$E(\mathbf{w}) = \beta E_D(\beta) + \alpha E_W(\mathbf{w}) = \frac{\beta}{2} \|\mathbf{Y}_m - \boldsymbol{\Phi} \cdot \mathbf{w}^\top\|^2 + \frac{\alpha}{2} \mathbf{w} \cdot \mathbf{w}^\top$$

• Since  $p(\mathcal{D}_m|\mathbf{w},\beta)$  and  $p(\mathbf{w}|\alpha)$  are Gaussians with quadratic forms  $E_D(\beta)$  and  $E_W(\mathbf{w})$ , we get that

$$p(\mathcal{D}_m | \alpha, \beta) = \left(\frac{\beta}{2\pi}\right)^{m/2} \left(\frac{\alpha}{2\pi}\right)^{M/2} \int \exp\{-E(\mathbf{w})\} d\mathbf{w}$$

Skutkons Institute of Science and Technology 42/43

• Let us calculate the evidence for  $(\alpha, \beta)$ 

$$p(\mathcal{D}_m|\alpha,\beta) = \int p(\mathcal{D}_m|\mathbf{w},\beta)p(\mathbf{w}|\alpha)d\mathbf{w}$$

 $\bullet$  Let us denote by  $E(\mathbf{w})$  the sum of the fit and the regularization on coefficients  $\mathbf{w}$ 

$$E(\mathbf{w}) = \beta E_D(\beta) + \alpha E_W(\mathbf{w}) = \frac{\beta}{2} \|\mathbf{Y}_m - \boldsymbol{\Phi} \cdot \mathbf{w}^\top\|^2 + \frac{\alpha}{2} \mathbf{w} \cdot \mathbf{w}^\top$$

• Since  $p(\mathcal{D}_m|\mathbf{w},\beta)$  and  $p(\mathbf{w}|\alpha)$  are Gaussians with quadratic forms  $E_D(\beta)$  and  $E_W(\mathbf{w})$ , we get that

$$p(\mathcal{D}_m|\alpha,\beta) = \left(\frac{\beta}{2\pi}\right)^{m/2} \left(\frac{\alpha}{2\pi}\right)^{M/2} \int \exp\{-E(\mathbf{w})\} d\mathbf{w}$$

(ㅁㅏㅓ@ㅏㅓㅌㅏㅓㅌㅏ ㅌ 쒸٩♡

Burnaev, ML

$$p(\mathcal{D}_m|\alpha,\beta) = \left(\frac{\beta}{2\pi}\right)^{m/2} \left(\frac{\alpha}{2\pi}\right)^{M/2} \int \exp\{-E(\mathbf{w})\} d\mathbf{w}$$

and we can get that

$$\log p(\mathcal{D}_m | \alpha, \beta) = \frac{M}{2} \log \alpha + \frac{m}{2} \log \beta$$
$$-E(\omega_N) - \frac{1}{2} \log |\mathbf{A}| - \frac{m}{2} \log(2\pi),$$

where

$$\mathbf{A} = \mathbf{S}_m^{-1} = \alpha^{-1} \mathbf{I} + \beta \mathbf{\Phi}^{\top} \mathbf{\Phi} \in \mathbb{R}^{M \times M},$$
  
$$\omega_m = \beta \mathbf{S}_m \mathbf{\Phi}^{\top} \mathbf{Y}_m$$

• **Seminar**: derivations of all formulas and an approach to optimize  $\log p(\mathcal{D}_m | \alpha, \beta)$  w.r.t.  $(\alpha, \beta)$ 

Burnaev, ML SKOITECH

43/43

$$p(\mathcal{D}_m|\alpha,\beta) = \left(\frac{\beta}{2\pi}\right)^{m/2} \left(\frac{\alpha}{2\pi}\right)^{M/2} \int \exp\{-E(\mathbf{w})\} d\mathbf{w}$$

and we can get that

$$\log p(\mathcal{D}_m | \alpha, \beta) = \frac{M}{2} \log \alpha + \frac{m}{2} \log \beta$$
$$-E(\boldsymbol{\omega}_N) - \frac{1}{2} \log |\mathbf{A}| - \frac{m}{2} \log(2\pi),$$

where

$$\mathbf{A} = \mathbf{S}_m^{-1} = \alpha^{-1} \mathbf{I} + \beta \mathbf{\Phi}^{\top} \mathbf{\Phi} \in \mathbb{R}^{M \times M},$$
  
$$\mathbf{o}_m = \beta \mathbf{S}_m \mathbf{\Phi}^{\top} \mathbf{Y}_m$$

• **Seminar**: derivations of all formulas and an approach to optimize  $\log p(\mathcal{D}_m|\alpha,\beta)$  w.r.t.  $(\alpha,\beta)$ 

◆ロ → ← 部 → ← き → を ● り へ で

43/43

Burnaev, ML Skalan Indian (Skalan Indian Ind

$$p(\mathcal{D}_m|\alpha,\beta) = \left(\frac{\beta}{2\pi}\right)^{m/2} \left(\frac{\alpha}{2\pi}\right)^{M/2} \int \exp\{-E(\mathbf{w})\} d\mathbf{w}$$

and we can get that

$$\log p(\mathcal{D}_m | \alpha, \beta) = \frac{M}{2} \log \alpha + \frac{m}{2} \log \beta$$
$$-E(\boldsymbol{\omega}_N) - \frac{1}{2} \log |\mathbf{A}| - \frac{m}{2} \log(2\pi),$$

where

$$\mathbf{A} = \mathbf{S}_m^{-1} = \alpha^{-1} \mathbf{I} + \beta \mathbf{\Phi}^{\top} \mathbf{\Phi} \in \mathbb{R}^{M \times M},$$
 $\omega_m = \beta \mathbf{S}_m \mathbf{\Phi}^{\top} \mathbf{Y}_m$ 

• **Seminar**: derivations of all formulas and an approach to optimize  $\log p(\mathcal{D}_m | \alpha, \beta)$  w.r.t.  $(\alpha, \beta)$ 

◆ロト ◆部 ト ◆ 恵 ト ◆ 恵 ・ 夕 Q C

43/43

Burnaev, ML Skanni data elikarrania

$$p(\mathcal{D}_m|\alpha,\beta) = \left(\frac{\beta}{2\pi}\right)^{m/2} \left(\frac{\alpha}{2\pi}\right)^{M/2} \int \exp\{-E(\mathbf{w})\} d\mathbf{w}$$

and we can get that

$$\log p(\mathcal{D}_m | \alpha, \beta) = \frac{M}{2} \log \alpha + \frac{m}{2} \log \beta$$
$$-E(\boldsymbol{\omega}_N) - \frac{1}{2} \log |\mathbf{A}| - \frac{m}{2} \log(2\pi),$$

where

$$\mathbf{A} = \mathbf{S}_m^{-1} = \alpha^{-1} \mathbf{I} + \beta \mathbf{\Phi}^{\top} \mathbf{\Phi} \in \mathbb{R}^{M \times M},$$
  
$$\mathbf{\omega}_m = \beta \mathbf{S}_m \mathbf{\Phi}^{\top} \mathbf{Y}_m$$

• **Seminar**: derivations of all formulas and an approach to optimize  $\log p(\mathcal{D}_m | \alpha, \beta)$  w.r.t.  $(\alpha, \beta)$ 

◆ロ → ← 部 → ← き → を ● り へ で

43/43

Burnaev, ML Salasi International Salasi Internation