Bayesian Machine Learning

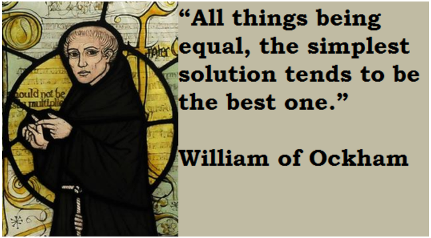
Evgeny Burnaev

Skoltech, Moscow, Russia

Outline

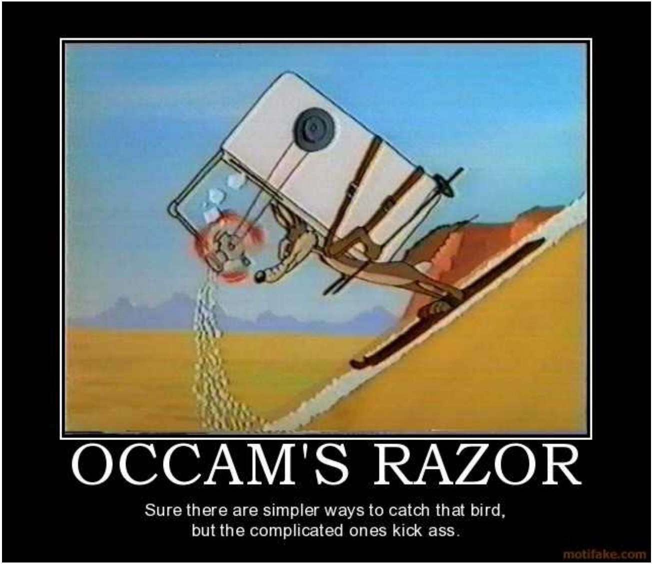
Thomas Bayes (c. 1701 – 7 April 1761) was an English statistician, philosopher and Presbyterian minister

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



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

Occam's razor



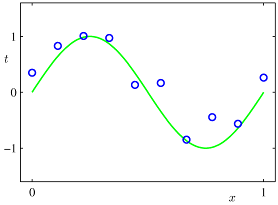


Figure – Plot of a training data

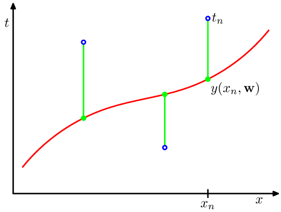


Figure – Residuals

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

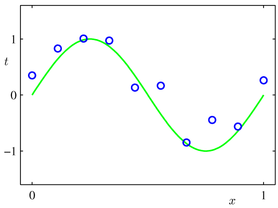


Figure – Plot of a training data

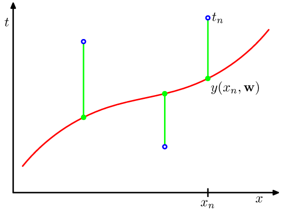


Figure – Residuals

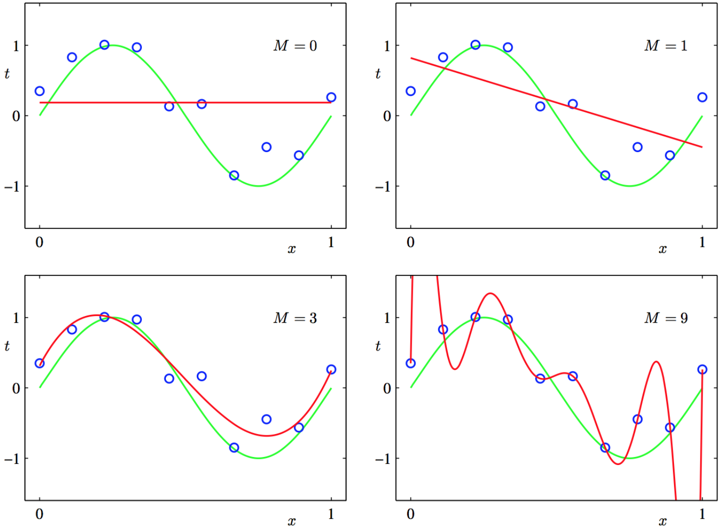
- We fit a model

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

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 M



Overfitting

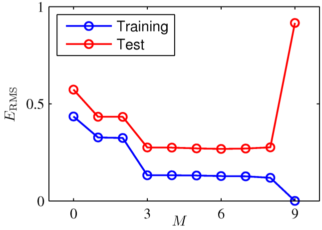


Figure – $E_{RMS} = \sqrt{2E(\mathbf{w}^*)/n}$

	$M=0$	$M=1$	$M=6$	$M=9$
$w_0^\star$	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 $\mathbf{w}^*$

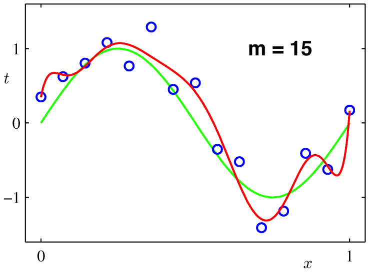


Figure – $M = 9, m = 15$

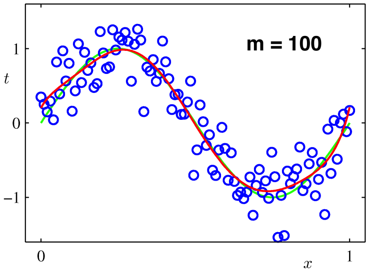


Figure – $M = 9, m = 100$

- 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!

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

- 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!

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

- 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!

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

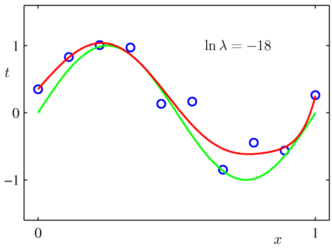


Figure – $\lambda = e^{-18} \approx 0$

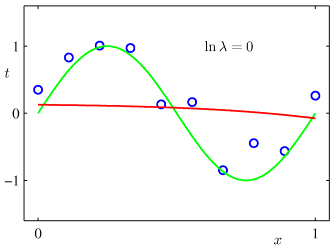


Figure – $\lambda = 1$

	$\ln\lambda = -\infty$	$\ln\lambda = -18$	$\ln\lambda = 0$
$w_0^\star$	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
$w_9^\star$	125201.43	72.68	0.01

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

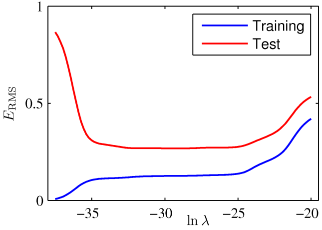


Figure – Dependence of E_{RMS} on λ

- 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 λ)? 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
$w_9^\star$	125201.43	72.68	0.01

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

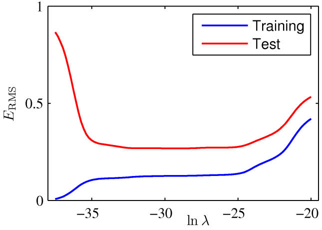
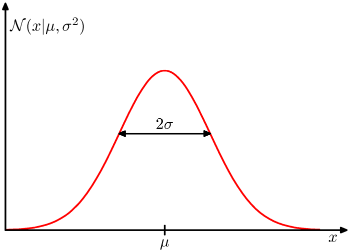


Figure – Dependence of E_{RMS} on λ

- 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 λ)? Too wasteful $\Rightarrow$ Bayesian Learning!

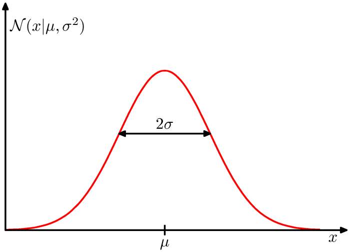


- Gaussian distribution of $x \in \mathbb{R}^1$ with $\mathbb{E}[x] = \mu$, $\mathrm{var}[x] = \sigma^2$

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

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

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



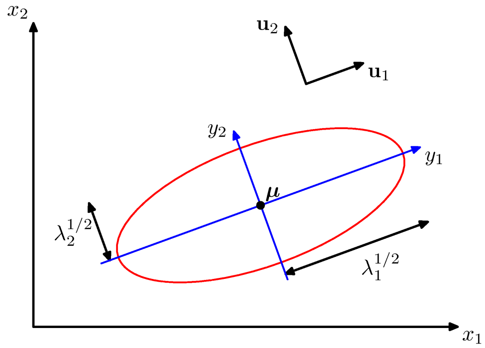
- Gaussian distribution of $x \in \mathbb{R}^1$ with $\mathbb{E}[x] = \mu$, $\mathrm{var}[x] = \sigma^2$

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

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

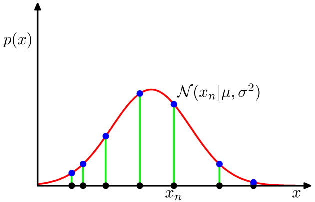
$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

Gaussian distribution



- The red curve shows the elliptical surface of constant probability density for $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $d = 2$
- 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 $\boldsymbol{\Sigma}$, with eigenvalues λ_i

Gaussian MLE



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

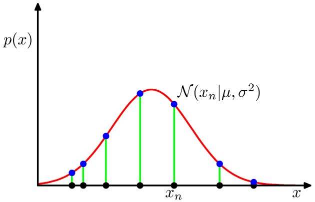
$$p(\mathbf{X}_m|\mu,\sigma^2) = \prod_{i=1}^{m} \mathcal{N}(x_i|\mu,\sigma^2)$$

- Log-likelihood is equal to

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

- MLE is equal to

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



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

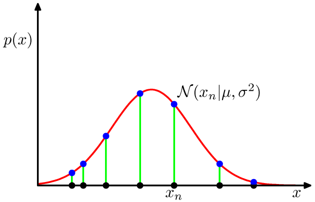
$$p(\mathbf{X}_m|\mu,\sigma^2) = \prod_{i=1}^{m} \mathcal{N}(x_i|\mu,\sigma^2)$$

- Log-likelihood is equal to

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

- MLE is equal to

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



- 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

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

- MLE is equal to

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

- 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

- 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

- 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

- 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

- 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

- 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

- 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

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

— It is almost impossible to predict random rare events $\Rightarrow$ their description is very long $\Rightarrow$ complex

— $\mathbf{w}$ defines "complexity" of the model

— $p(\mathbf{w})$ quantifies this complexity, as "small probability" $\equiv$ "complex"

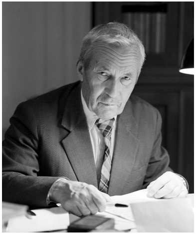


Figure – Kolmogorov A.N. (1903-1987)

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

— It is almost impossible to predict random rare events $\Rightarrow$ their description is very long $\Rightarrow$ complex

— $\mathbf{w}$ defines "complexity" of the model

— $p(\mathbf{w})$ quantifies this complexity, as "small probability" $\equiv$ "complex"

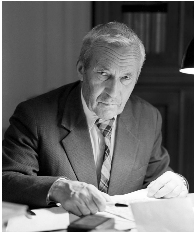


Figure – Kolmogorov A.N. (1903-1987)

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

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

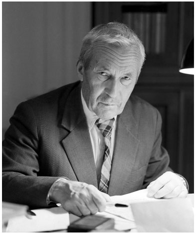


Figure – Kolmogorov A.N. (1903-1987)

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

Probability vs. complexity (Kolmogorov):

— It is almost impossible to predict random rare events $\Rightarrow$ their description is very long $\Rightarrow$ complex

— $\mathbf{w}$ defines "complexity" of the model

— $p(\mathbf{w})$ quantifies this complexity, as "small probability" $\equiv$ "complex"

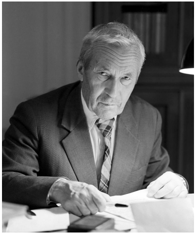


Figure – Kolmogorov A.N. (1903-1987)

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

Probability vs. complexity (Kolmogorov):

— It is almost impossible to predict random rare events $\Rightarrow$ their description is very long $\Rightarrow$ complex

— $\mathbf{w}$ defines "complexity" of the model

— $p(\mathbf{w})$ quantifies this complexity, as "small probability" $\equiv$ "complex"

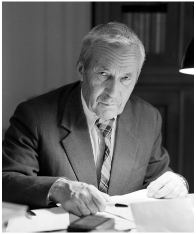


Figure – Kolmogorov A.N. (1903-1987)

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

Probability vs. complexity (Kolmogorov):

— It is almost impossible to predict random rare events ⇒ their description is very long ⇒ complex

— $\mathbf{w}$ defines "complexity" of the model

— $p(\mathbf{w})$ quantifies this complexity, as "small probability" ≡ "complex"

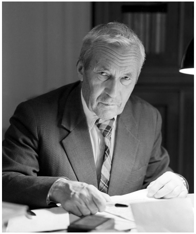


Figure – Kolmogorov A.N. (1903-1987)

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

Probability vs. complexity (Kolmogorov):

— It is almost impossible to predict random rare events $\Rightarrow$ their description is very long $\Rightarrow$ complex

— $\mathbf{w}$ defines "complexity" of the model

— $p(\mathbf{w})$ quantifies this complexity, as "small probability" $\equiv$ "complex"

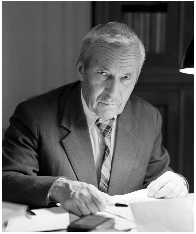


Figure – Kolmogorov A.N. (1903-1987)

- Observed data $\mathcal{D}_m = \{\mathbf{X}, \mathbf{y}\} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ 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:

$$\text{posterior} \sim \text{likelihood} \times \text{prior}$$

$$\log \text{posterior} \sim \log \text{likelihood} + \log \text{prior}$$

Example: curve fitting problem

- Observed data $\mathcal{D}_m = \{\mathbf{X}, \mathbf{y}\} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ 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:

$$\text{posterior} \sim \text{likelihood} \times \text{prior}$$

$$\log \text{posterior} \sim \log \text{likelihood} + \log \text{prior}$$

- Observed data $\mathcal{D}_m = \{\mathbf{X}, \mathbf{y}\} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ 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:

$$\text{posterior} \sim \text{likelihood} \times \text{prior}$$

$$\log \text{posterior} \sim \log \text{likelihood} + \log \text{prior}$$

- Observed data $\mathcal{D}_m = \{\mathbf{X}, \mathbf{y}\} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ 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:

$$\text{posterior} \sim \text{likelihood} \times \text{prior}$$

$$\log \text{posterior} \sim \log \text{likelihood} + \log \text{prior}$$

- Observed data $\mathcal{D}_m = \{\mathbf{X}, \mathbf{y}\} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ 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:

$$\text{posterior} \sim \text{likelihood} \times \text{prior}$$

$$\log \text{posterior} \sim \log \text{likelihood} + \log \text{prior}$$

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

- **Frequentist setting**:
  — $\mathbf{w}$ is a fixed parameter,
- **Bayesian setting**:
  — the uncertainty in the parameters is expressed through a probability distribution over $\mathbf{w}$,
  — we reduce uncertainty about $\mathbf{w}$ by observing more and more data
- The inclusion of prior knowledge arises naturally

- **Frequentist setting**:
  — $\mathbf{w}$ is a fixed parameter,
- **Bayesian setting**:
  — the uncertainty in the parameters is expressed through a probability distribution over $\mathbf{w}$,
  — we reduce uncertainty about $\mathbf{w}$ by observing more and more data
- The inclusion of prior knowledge arises naturally

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

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

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

- MLE estimate:

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

- 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:

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

- MLE estimate:

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

- 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

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

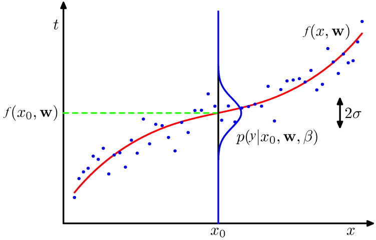
$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \log p(\mathbf{w}|\mathcal{D}_m)$$

  - MAP $\equiv$ regularized MLE:

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

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

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



- Sample $\mathcal{D}_m = \{\mathbf{X}_m, \mathbf{Y}_m\} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$,

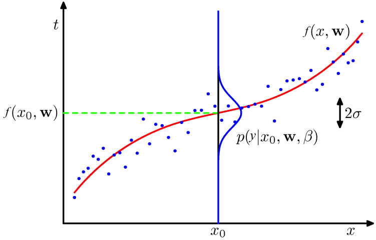
$$y_i = f(\mathbf{x}_i, \mathbf{w}) + \varepsilon_i, \text{ with i.i.d. } \varepsilon_i \sim \mathcal{N}(0, \beta^{-1})$$

- 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

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



- Sample $\mathcal{D}_m = \{\mathbf{X}_m, \mathbf{Y}_m\} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$,

$$y_i = f(\mathbf{x}_i, \mathbf{w}) + \varepsilon_i, \text{ with i.i.d. } \varepsilon_i \sim \mathcal{N}(0, \beta^{-1})$$

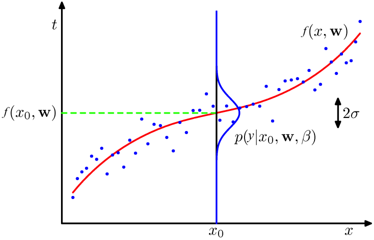
- Probabilistic model

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

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

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



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

- Log-likelihood

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

- MLE of β

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

- Predictive distribution

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

- Log-likelihood

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

- MLE of β

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

- Predictive distribution

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

Curve-fitting revisited

- A prior distribution over the polynomial coefficients $\mathbf{w}$

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

- 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

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

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

- A prior distribution over the polynomial coefficients $\mathbf{w}$

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

- 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

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

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

- A prior distribution over the polynomial coefficients $\mathbf{w}$

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

- 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

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

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

- A prior distribution over the polynomial coefficients $\mathbf{w}$

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

- 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

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

- Maximum posterior

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

- 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 $\lambda = \frac{\alpha}{\beta}$

- Maximum posterior

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

- 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 $\lambda = \frac{\alpha}{\beta}$

- Maximum posterior

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

- 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 $\lambda = \frac{\alpha}{\beta}$

- 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 $\mathbf{X}_m$ 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 $\mathbf{X}_m$ 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 $\mathbf{X}_m$ 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

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

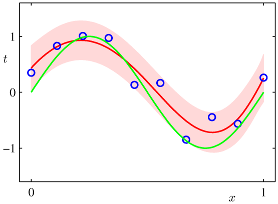


Figure – The predictive distribution for a polynomial with $M = 9$, parameters $\alpha = 5 \times 10^{-3}$ and $\beta = 11.1$ (known noise variance) are fixed

- 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

$$\phi_j(\mathbf{x}) = x_{j_1}^{j_0},\ \phi_j(\mathbf{x}) = \exp\left\{-\frac{\|\mathbf{x} - \boldsymbol{\mu}_j\|^2}{2s^2}\right\},$$

$$\phi(\mathbf{x}) = \sigma\left(\boldsymbol{\mu}_{j,1} \cdot \mathbf{x}^\top + \mu_{j,0}\right),\ \sigma(a) = \frac{1}{1 + e^{-a}}$$

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

$$\phi_j(\mathbf{x}) = x_{j_1}^{j_0},\ \phi_j(\mathbf{x}) = \exp\left\{-\frac{\|\mathbf{x} - \boldsymbol{\mu}_j\|^2}{2s^2}\right\},$$

$$\phi(\mathbf{x}) = \sigma\left(\boldsymbol{\mu}_{j,1} \cdot \mathbf{x}^\top + \mu_{j,0}\right),\ \sigma(a) = \frac{1}{1 + e^{-a}}$$

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

Least Squares = MLE

- Optimizing log-likelihood:

$$\mathbf{w}_{ML} = \left(\boldsymbol{\Phi}^\top \boldsymbol{\Phi}\right)^{-1} \boldsymbol{\Phi}^\top \mathbf{Y}_m, \;\; \boldsymbol{\Phi} = \{(\boldsymbol{\phi}_i(\mathbf{x}_j))_{j=0}^{M-1}\}_{i=1}^{m}$$

$$\frac{1}{\beta_{ML}} = \frac{1}{m}\sum_{i=1}^{m}\{y_i - \mathbf{w}_{ML}\cdot\boldsymbol{\phi}(\mathbf{x}_i)^\top\}^2$$

- Regularized Least Squares

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

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

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

- Optimizing log-likelihood:

$$\mathbf{w}_{ML} = \left(\boldsymbol{\Phi}^\top \boldsymbol{\Phi}\right)^{-1} \boldsymbol{\Phi}^\top \mathbf{Y}_m, \;\; \boldsymbol{\Phi} = \{(\phi_i(\mathbf{x}_j))_{j=0}^{M-1}\}_{i=1}^{m}$$

$$\frac{1}{\beta_{ML}} = \frac{1}{m} \sum_{i=1}^{m} \{y_i - \mathbf{w}_{ML} \cdot \boldsymbol{\phi}(\mathbf{x}_i)^\top\}^2$$

- Regularized Least Squares

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

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

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

- 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

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

- 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

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

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

Parameter distribution

- For

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

  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

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

- Thus the posterior is defined by

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

- For

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

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

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

- Thus the posterior is defined by

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

- For

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

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

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

- Thus the posterior is defined by

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

Parameter distribution

- For

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

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

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

- Thus the posterior is defined by

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

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

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

- For

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

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

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

- Thus the posterior is defined by

$$p(\mathbf{w}|\mathcal{D}_m) = \mathcal{N}(\mathbf{w}|\boldsymbol{\omega}_m, \mathbf{S}_m)$$

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

$$\boldsymbol{\omega}_m = \beta\mathbf{S}_m\boldsymbol{\Phi}^\top\mathbf{Y}_m$$

- For

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

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

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

- Thus the posterior is defined by

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

- For

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

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

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

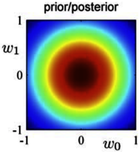
- Thus the posterior is defined by

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

likelihood

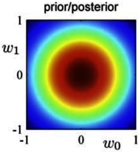
The Model $f(x, \mathbf{w}) = w_0 + w_1 x$

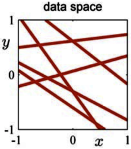
likelihood



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

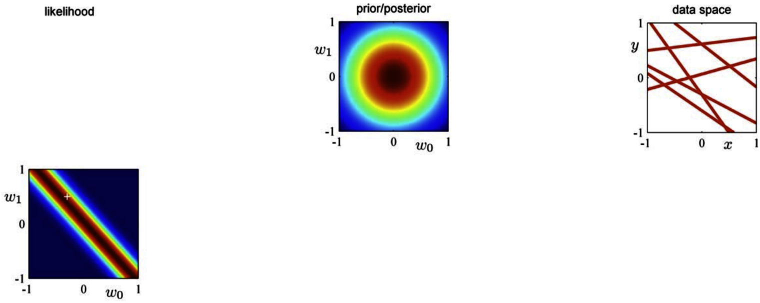
likelihood





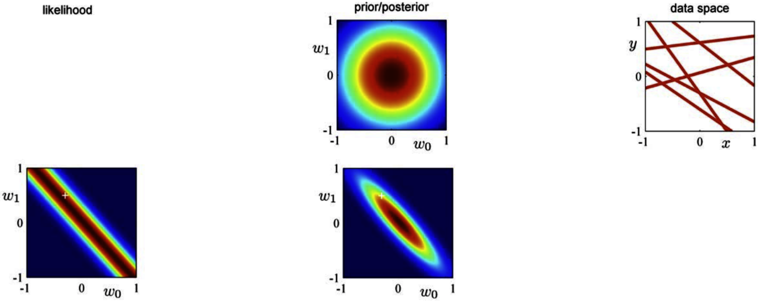
The Model $f(x, \mathbf{w}) = w_0 + w_1 x$

Sequential Bayesian Learning



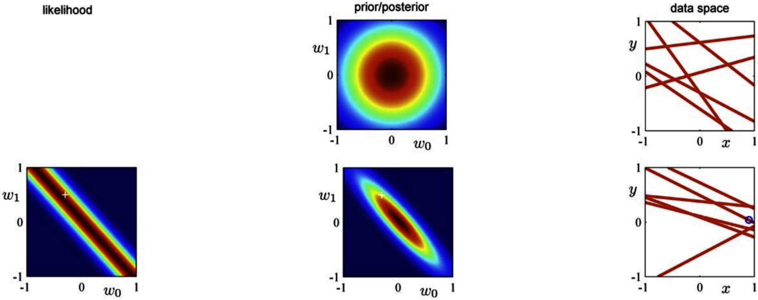
The Model $f(x, \mathbf{w}) = w_0 + w_1 x$

Sequential Bayesian Learning



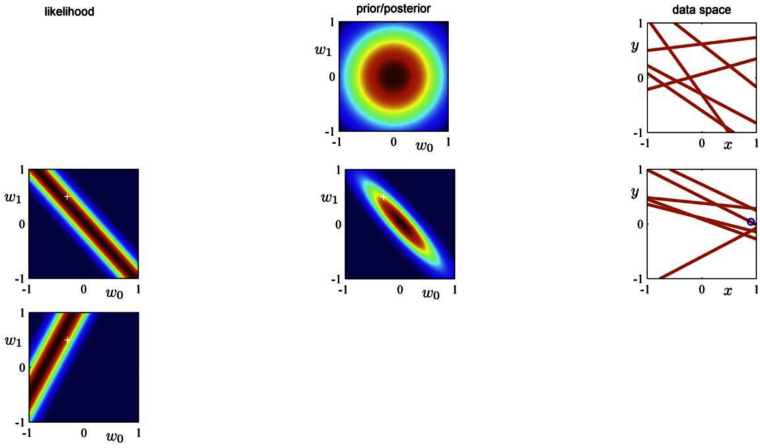
The Model $f(x, \mathbf{w}) = w_0 + w_1 x$

Sequential Bayesian Learning



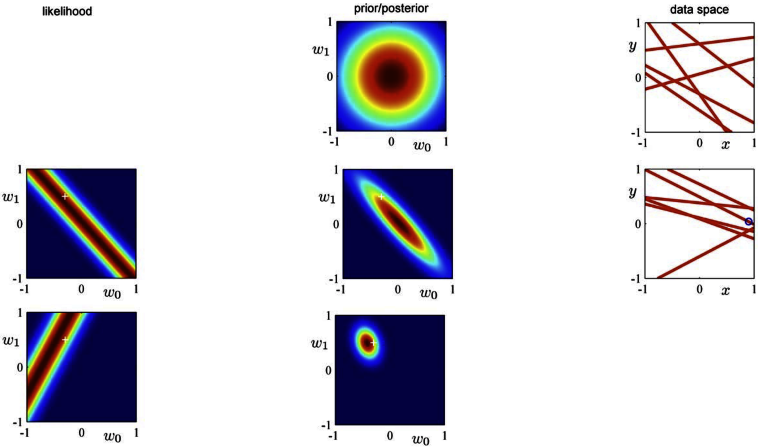
The Model $f(x, \mathbf{w}) = w_0 + w_1 x$

Sequential Bayesian Learning



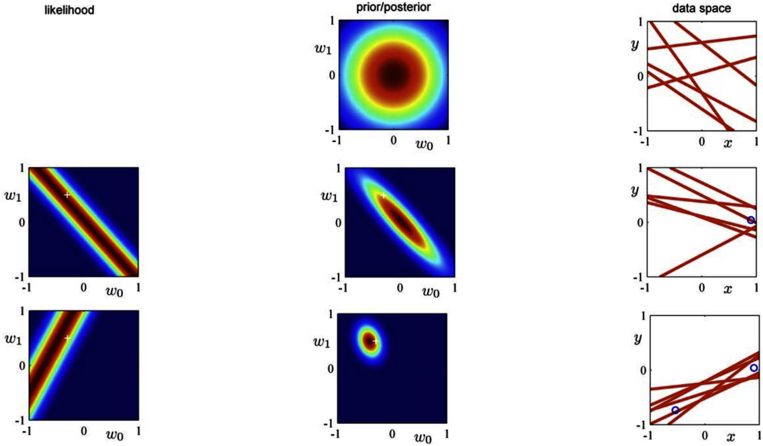
The Model $f(x, \mathbf{w}) = w_0 + w_1 x$

Sequential Bayesian Learning

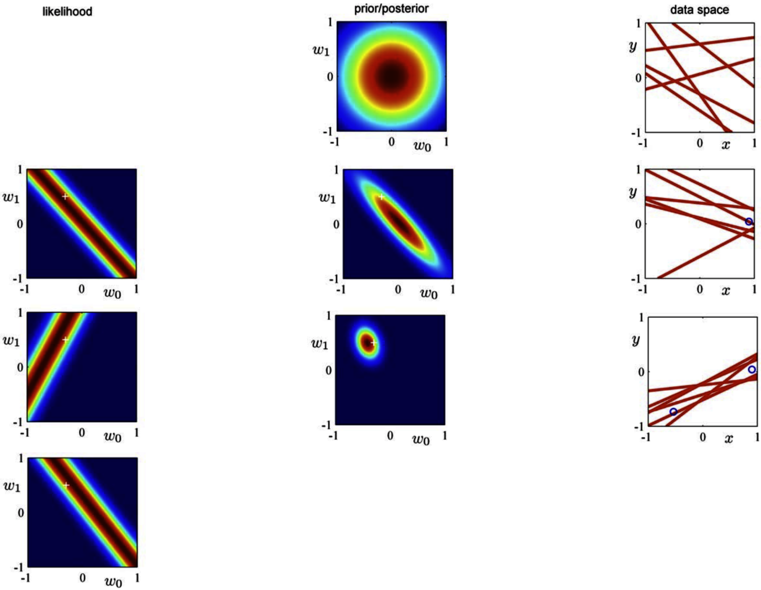


The Model $f(x, \mathbf{w}) = w_0 + w_1 x$

Sequential Bayesian Learning

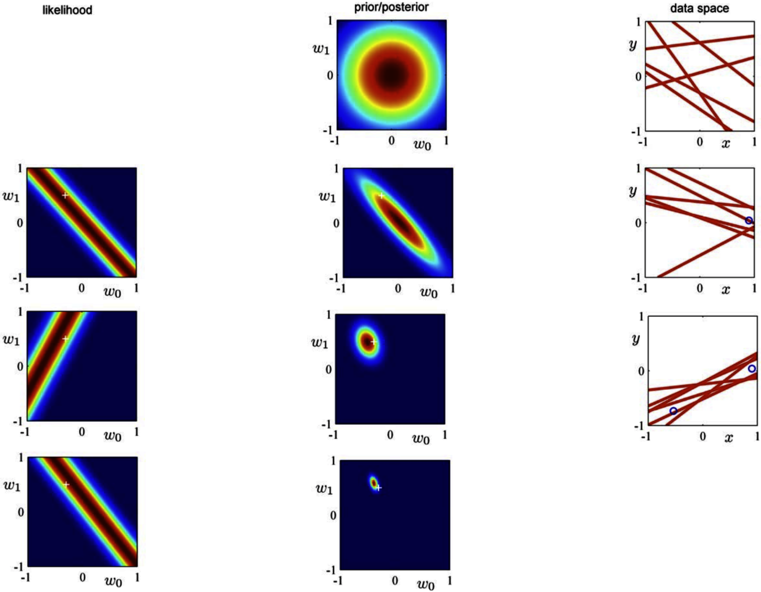


The Model $f(x, \mathbf{w}) = w_0 + w_1 x$



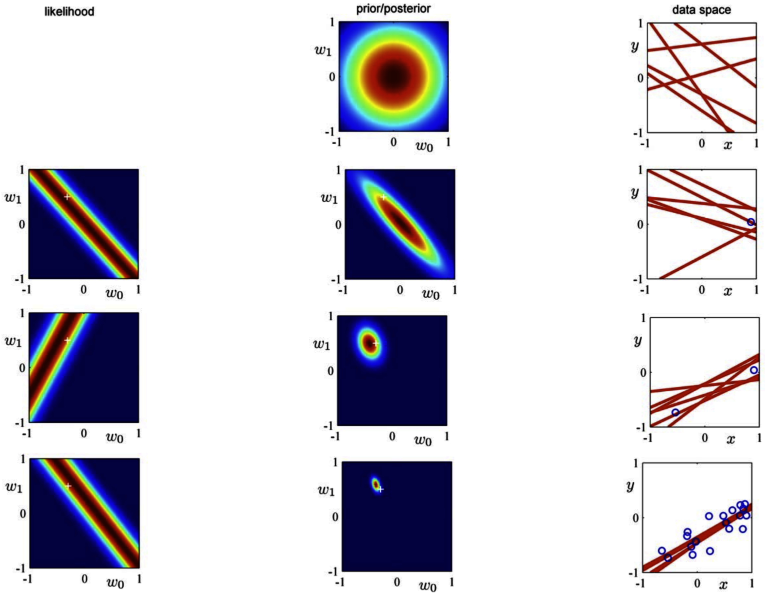
The Model $f(x, \mathbf{w}) = w_0 + w_1 x$

Sequential Bayesian Learning



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

Sequential Bayesian Learning



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

- Make prediction of y for new value of $\mathbf{x}$:

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

- Actually, posterior of $\mathbf{w}$ is $p(\mathbf{w}|\mathcal{D}_m) = \mathcal{N}(\mathbf{w}|\boldsymbol{\omega}_m, \mathbf{S}_m)$ with
  - $\mathbf{S}_m = \left(\alpha^{-1}\mathbf{I} + \beta\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)^{-1}$ — posterior covariance of $\mathbf{w}$
  - $\boldsymbol{\omega}_m = \beta\mathbf{S}_m\boldsymbol{\Phi}^\top\mathbf{Y}_m$ — posterior mean of $\mathbf{w}$
- Since $p(y|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(y|f(\mathbf{x}, \mathbf{w}), \beta^{-1})$ and the posterior $p(\mathbf{w}|\mathcal{D}_m, \alpha, \beta)$ is Gaussian, then

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

  Here

- We can use posterior mean for point prediction

$$\widehat{f}(\mathbf{x}, \mathbf{w}) = \boldsymbol{\omega}_m \cdot \boldsymbol{\phi}(\mathbf{x})^\top$$

  and posterior variance $\sigma_m^2(\mathbf{x})$ for its uncertainty estimate

- Make prediction of y for new value of $\mathbf{x}$:

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

- Actually, posterior of $\mathbf{w}$ is $p(\mathbf{w}|\mathcal{D}_m) = \mathcal{N}(\mathbf{w}|\boldsymbol{\omega}_m, \mathbf{S}_m)$ with
  - $\mathbf{S}_m = \left(\alpha^{-1}\mathbf{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi}\right)^{-1}$ — posterior covariance of $\mathbf{w}$
  - $\boldsymbol{\omega}_m = \beta \mathbf{S}_m \boldsymbol{\Phi}^\top \mathbf{Y}_m$ — posterior mean of $\mathbf{w}$
- Since $p(y|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(y|f(\mathbf{x}, \mathbf{w}), \beta^{-1})$ and the posterior $p(\mathbf{w}|\mathcal{D}_m, \alpha, \beta)$ is Gaussian, then

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

Here

- We can use posterior mean for point prediction

$$\widehat{f}(\mathbf{x}, \mathbf{w}) = \boldsymbol{\omega}_m \cdot \boldsymbol{\phi}(\mathbf{x})^\top$$

and posterior variance $\sigma_m^2(\mathbf{x})$ for its uncertainty estimate

- Make prediction of y for new value of $\mathbf{x}$:

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

- Actually, posterior of $\mathbf{w}$ is $p(\mathbf{w}|\mathcal{D}_m) = \mathcal{N}(\mathbf{w}|\boldsymbol{\omega}_m, \mathbf{S}_m)$ with
  - $\mathbf{S}_m = \left(\alpha^{-1}\mathbf{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi}\right)^{-1}$ — posterior covariance of $\mathbf{w}$
  - $\boldsymbol{\omega}_m = \beta \mathbf{S}_m \boldsymbol{\Phi}^\top \mathbf{Y}_m$ — posterior mean of $\mathbf{w}$
- Since $p(y|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(y|f(\mathbf{x}, \mathbf{w}), \beta^{-1})$ and the posterior $p(\mathbf{w}|\mathcal{D}_m, \alpha, \beta)$ is Gaussian, then

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

Here

- We can use posterior mean for point prediction

$$\widehat{f}(\mathbf{x}, \mathbf{w}) = \boldsymbol{\omega}_m \cdot \boldsymbol{\phi}(\mathbf{x})^\top$$

and posterior variance $\sigma_m^2(\mathbf{x})$ for its uncertainty estimate

- Make prediction of y for new value of $\mathbf{x}$:

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

- Actually, posterior of $\mathbf{w}$ is $p(\mathbf{w}|\mathcal{D}_m) = \mathcal{N}(\mathbf{w}|\boldsymbol{\omega}_m, \mathbf{S}_m)$ with
  - $\mathbf{S}_m = \left(\alpha^{-1}\mathbf{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi}\right)^{-1}$ — posterior covariance of $\mathbf{w}$
  - $\boldsymbol{\omega}_m = \beta \mathbf{S}_m \boldsymbol{\Phi}^\top \mathbf{Y}_m$ — posterior mean of $\mathbf{w}$
- Since $p(y|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(y|f(\mathbf{x}, \mathbf{w}), \beta^{-1})$ and the posterior $p(\mathbf{w}|\mathcal{D}_m, \alpha, \beta)$ is Gaussian, then

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

  Here

- We can use posterior mean for point prediction

$$\widehat{f}(\mathbf{x}, \mathbf{w}) = \boldsymbol{\omega}_m \cdot \boldsymbol{\phi}(\mathbf{x})^\top$$

  and posterior variance $\sigma_m^2(\mathbf{x})$ for its uncertainty estimate

- Make prediction of y for new value of $\mathbf{x}$:

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

- Actually, posterior of $\mathbf{w}$ is $p(\mathbf{w}|\mathcal{D}_m) = \mathcal{N}(\mathbf{w}|\boldsymbol{\omega}_m, \mathbf{S}_m)$ with
  - $\mathbf{S}_m = \left(\alpha^{-1}\mathbf{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi}\right)^{-1}$ — posterior covariance of $\mathbf{w}$
  - $\boldsymbol{\omega}_m = \beta \mathbf{S}_m \boldsymbol{\Phi}^\top \mathbf{Y}_m$ — posterior mean of $\mathbf{w}$
- Since $p(y|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(y|f(\mathbf{x}, \mathbf{w}), \beta^{-1})$ and the posterior $p(\mathbf{w}|\mathcal{D}_m, \alpha, \beta)$ is Gaussian, then

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

Here

- We can use posterior mean for point prediction

$$\widehat{f}(\mathbf{x}, \mathbf{w}) = \boldsymbol{\omega}_m \cdot \boldsymbol{\phi}(\mathbf{x})^\top$$

and posterior variance $\sigma_m^2(\mathbf{x})$ for its uncertainty estimate

- Make prediction of y for new value of $\mathbf{x}$:

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

- Actually, posterior of $\mathbf{w}$ is $p(\mathbf{w}|\mathcal{D}_m) = \mathcal{N}(\mathbf{w}|\boldsymbol{\omega}_m, \mathbf{S}_m)$ with
  - $\mathbf{S}_m = \left(\alpha^{-1}\mathbf{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi}\right)^{-1}$ — posterior covariance of $\mathbf{w}$
  - $\boldsymbol{\omega}_m = \beta \mathbf{S}_m \boldsymbol{\Phi}^\top \mathbf{Y}_m$ — posterior mean of $\mathbf{w}$
- Since $p(y|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(y|f(\mathbf{x}, \mathbf{w}), \beta^{-1})$ and the posterior $p(\mathbf{w}|\mathcal{D}_m, \alpha, \beta)$ is Gaussian, then

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

  Here

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

- We can use posterior mean for point prediction

$$\widehat{f}(\mathbf{x}, \mathbf{w}) = \boldsymbol{\omega}_m \cdot \boldsymbol{\phi}(\mathbf{x})^\top$$

  and posterior variance $\sigma_m^2(\mathbf{x})$ for its uncertainty estimate

- Make prediction of y for new value of $\mathbf{x}$:

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

- Actually, posterior of $\mathbf{w}$ is $p(\mathbf{w}|\mathcal{D}_m) = \mathcal{N}(\mathbf{w}|\boldsymbol{\omega}_m, \mathsf{S}_m)$ with
  - $\mathsf{S}_m = \left(\alpha^{-1}\mathbf{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi}\right)^{-1}$ — posterior covariance of $\mathbf{w}$
  - $\boldsymbol{\omega}_m = \beta \mathsf{S}_m \boldsymbol{\Phi}^\top \mathbf{Y}_m$ — posterior mean of $\mathbf{w}$
- Since $p(y|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(y|f(\mathbf{x}, \mathbf{w}), \beta^{-1})$ and the posterior $p(\mathbf{w}|\mathcal{D}_m, \alpha, \beta)$ is Gaussian, then

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

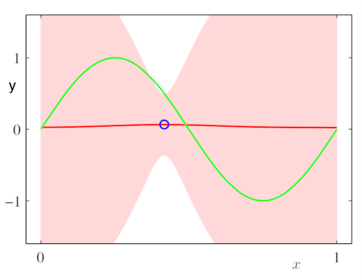
  Here

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

- We can use posterior mean for point prediction

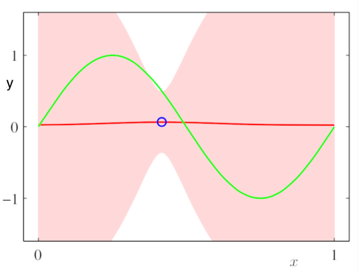
$$\widehat{f}(\mathbf{x}, \mathbf{w}) = \boldsymbol{\omega}_m \cdot \boldsymbol{\phi}(\mathbf{x})^\top$$

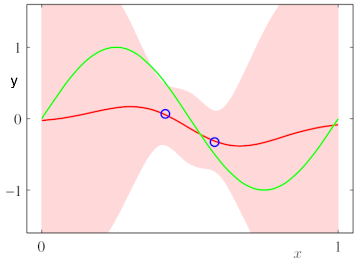
  and posterior variance $\sigma_m^2(\mathbf{x})$ for its uncertainty estimate



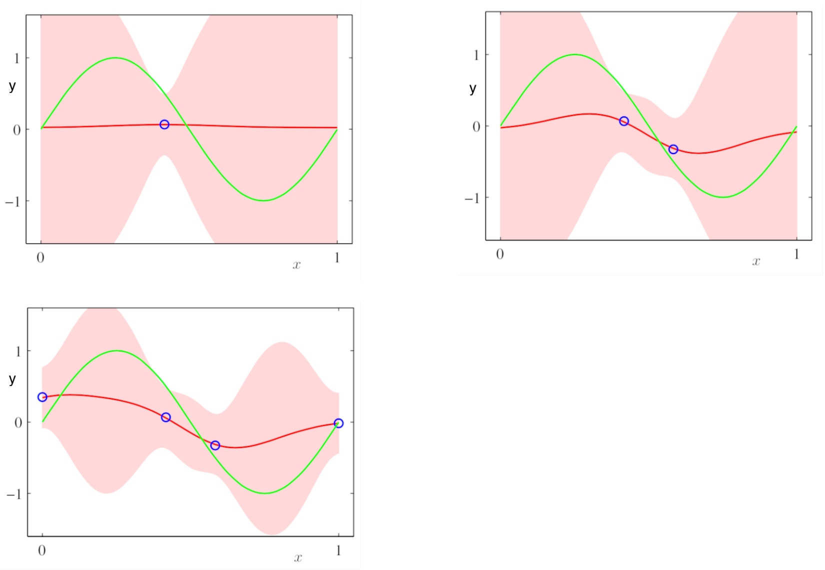
$M = 9$ Gaussian basis functions were used as $\phi(\mathbf{x})$

Predictive Distribution

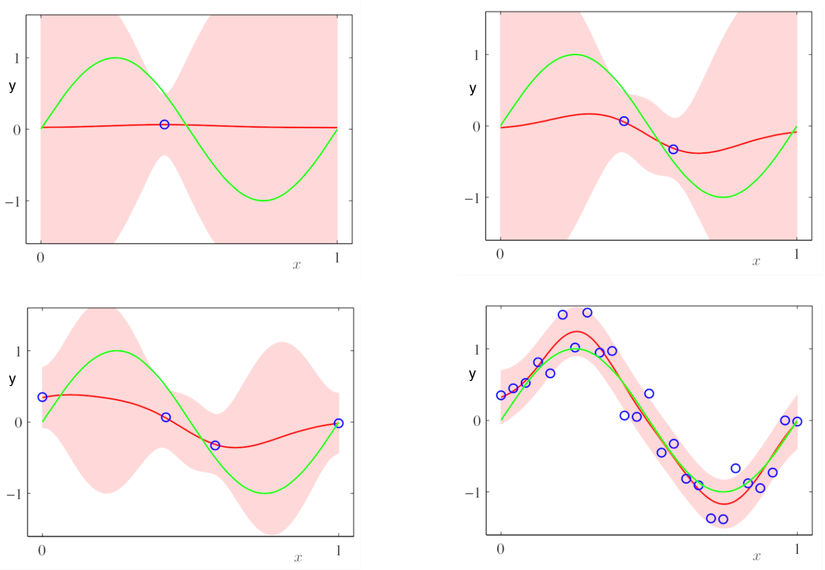




$M = 9$ Gaussian basis functions were used as $\phi(\mathbf{x})$

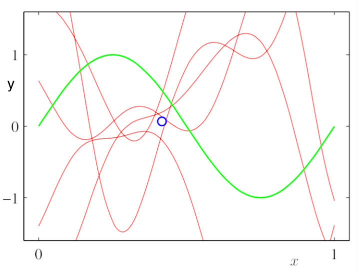


$M = 9$ Gaussian basis functions were used as $\phi(\mathbf{x})$



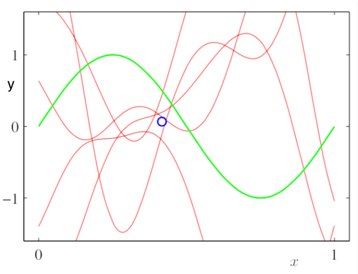
$M = 9$ Gaussian basis functions were used as $\phi(\mathbf{x})$

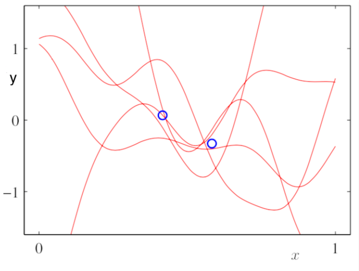
Samples from the Predictive Distribution



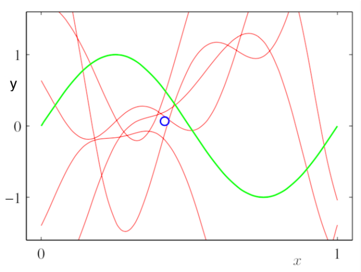
Plots of $f(\mathbf{x}, \mathbf{w})$ using samples from the posterior distributions over $\mathbf{w} \sim p(\mathbf{w}|\mathcal{D}_m, \alpha, \beta)$ for some α and β

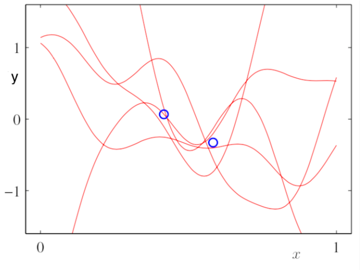
Samples from the Predictive Distribution

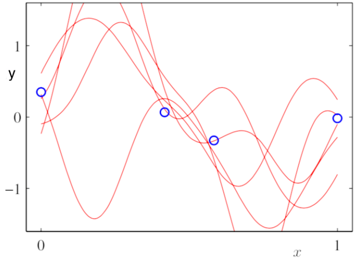




Plots of $f(\mathbf{x}, \mathbf{w})$ using samples from the posterior distributions over $\mathbf{w} \sim p(\mathbf{w}|\mathcal{D}_m, \alpha, \beta)$ for some α and β

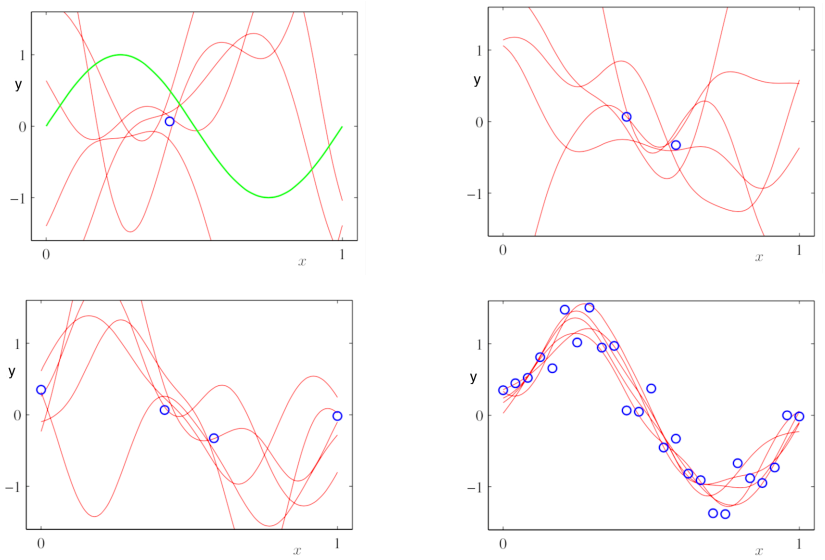






Plots of $f(\mathbf{x}, \mathbf{w})$ using samples from the posterior distributions over $\mathbf{w} \sim p(\mathbf{w}|\mathcal{D}_m, \alpha, \beta)$ for some α and β

Samples from the Predictive Distribution



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 α and β

- Distribution of y given new value of $\mathbf{x}$:

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

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

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

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

- $p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta)$ depends on α and β! How to define them? $\Rightarrow$ Full Bayesian approach! We introduce hyperpriors over α and β!

- Distribution of y given new value of $\mathbf{x}$:

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

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

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

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

- $p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta)$ depends on α and β! How to define them? $\Rightarrow$ Full Bayesian approach! We introduce hyperpriors over α and β!

- Distribution of y given new value of $\mathbf{x}$:

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

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

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

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

- $p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta)$ depends on α and β! How to define them? $\Rightarrow$ Full Bayesian approach! We introduce hyperpriors over α and β!

- Distribution of y given new value of $\mathbf{x}$:

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

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

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

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

- $p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta)$ depends on α and β! How to define them? $\Rightarrow$ Full Bayesian approach! We introduce hyperpriors over α and β!

- Distribution of y given new value of $\mathbf{x}$:

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

- $p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta)$ depends on α and β! How to define them? $\Rightarrow$ Full Bayesian approach! We introduce hyperpriors over α and β!

- Distribution of y given new value of $\mathbf{x}$:

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

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

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

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

- $p(y|\mathbf{x}, \mathcal{D}_m, \alpha, \beta)$ depends on α and β! How to define them? $\Rightarrow$ Full Bayesian approach! We introduce hyperpriors over α and β!

- We introduce hyperpriors over α and β

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

- 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 α and β are fixed to the values $\widehat{\alpha}$ and $\widehat{\beta}$, so that

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

- We introduce hyperpriors over α and β

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

- 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 α and β are fixed to the values $\widehat{\alpha}$ and $\widehat{\beta}$, so that

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

- We introduce hyperpriors over α and β

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

- 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 α and β are fixed to the values $\widehat{\alpha}$ and $\widehat{\beta}$, so that

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

- The posterior for α and β 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

- The posterior for α and β 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

- The posterior for α and β 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

- Let us calculate the evidence for (α, β)

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

- 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

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

- Let us calculate the evidence for (α, β)

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

- 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

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

- Let us calculate the evidence for (α, β)

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

- 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

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

- So

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

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

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

- **Bayesian ML course**: derivations of all formulas and an approach to optimize $\log p(\mathcal{D}_m|\alpha,\beta)$ w.r.t. (α,β)

- So

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

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

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

- **Bayesian ML course**: derivations of all formulas and an approach to optimize $\log p(\mathcal{D}_m|\alpha,\beta)$ w.r.t. (α,β)

- So

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

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

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

$$\omega_m = \beta\mathbf{S}_m\boldsymbol{\Phi}^\top\mathbf{Y}_m$$

- **Bayesian ML course**: derivations of all formulas and an approach to optimize $\log p(\mathcal{D}_m|\alpha,\beta)$ w.r.t. (α,β)

- So

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

and we can get that

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

where

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

- **Bayesian ML course**: derivations of all formulas and an approach to optimize $\log p(\mathcal{D}_m|\alpha,\beta)$ w.r.t. (α,β)