Conjugate Distributions. Bayesian Linear Regression

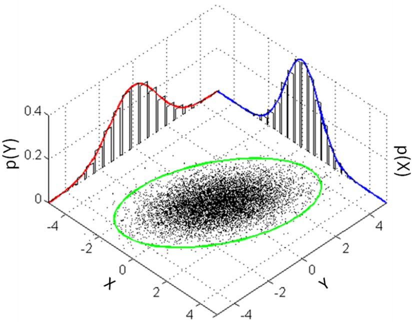
Evgeny Burnaev

Skoltech, Moscow, Russia

Skoltech

Skolkovo Institute of Science and Technology

Outline

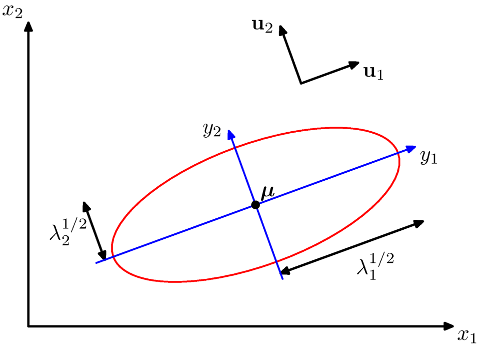


- In case of a single variable x

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

- For $\mathbf{x} \in \mathbb{R}^d$ with $\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$ and $\mathrm{cov}[\mathbf{x}] = \boldsymbol{\Sigma}$

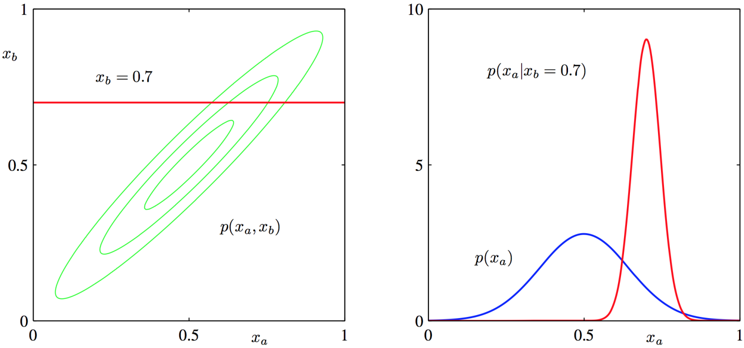
$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}} \frac{1}{|\boldsymbol{\Sigma}|^{d/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$



- The red curve — 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

- $\mathbf{x}$ is distributed as $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- We divide $\mathbf{x}$ in two subvectors $\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}$
- Let us also partition the mean and the covariance

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}$$



We can prove that

$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b}),$$

where

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b)$$
$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}$$

We assume that

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

Then we can prove that

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

where

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

$$p(\mathbf{x}) = h(\mathbf{x}) \cdot e^{\boldsymbol{\theta}^\top T(\mathbf{x}) - A(\boldsymbol{\theta})}$$

- $\boldsymbol{\theta}$ — vector of parameters
- $T(\mathbf{x})$ — vector of sufficient statistics
- $A(\boldsymbol{\theta})$ — cumulant generating function

Key point: $\mathbf{x}$ and $\boldsymbol{\theta}$ only "mix" in $e^{\boldsymbol{\theta}^\top T(\mathbf{x})}$

$$p(\mathbf{x}) = h(\mathbf{x}) \cdot e^{\boldsymbol{\theta}^\top T(\mathbf{x}) - A(\boldsymbol{\theta})}$$

To get a normalized distribution for any $\boldsymbol{\theta}$

$$\int p(\mathbf{x}) d\mathbf{x} = e^{-A(\boldsymbol{\theta})} \int h(\mathbf{x}) e^{\boldsymbol{\theta}^\top T(\mathbf{x})} d\mathbf{x} = 1$$

so

$$e^{A(\boldsymbol{\theta})} = \int h(\mathbf{x}) e^{\boldsymbol{\theta}^\top T(\mathbf{x})} d\mathbf{x}$$

E.g. for $T(\mathbf{x}) = x$, $A(\boldsymbol{\theta})$ is the $\log$ of Laplace transform of $h(\mathbf{x})$

- Gaussian $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, $x \in \mathbb{R}$
- Bernoulli $p(x) = \alpha^x (1-\alpha)^{1-x}$, $x \in \{0, 1\}$
- Binomial $p(x) = C_n^x \alpha^x (1-\alpha)^{n-x}$, $x \in \{0, 1, 2, \ldots, n\}$
- Multinomial $p(\mathbf{x}) = \frac{n!}{x_1! x_2! \ldots x_n!} \prod_{i=1}^n \alpha_i^{x_i}$, $x_i \in \{0, 1, 2, \ldots, n\}$, $\sum_i x_i = n$
- Exponential $p(x) = \lambda e^{-\lambda x}$, $x \in \mathbb{R}^+$
- Poisson $p(x) = \frac{e^{-\lambda}}{x!} \lambda^x$, $x \in \{0, 1, 2, \ldots\}$
- Dirichlet $p(\mathbf{x}) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_i x_i^{\alpha_i - 1}$, $x_i \in [0, 1]$, $\sum_i x_i = 1$

Natural Parameter Form for Bernoulli

$$p(\mathbf{x}) = h(\mathbf{x}) \cdot e^{\boldsymbol{\theta}^\top T(\mathbf{x}) - A(\boldsymbol{\theta})}$$

$$\begin{aligned} p(x) = \alpha^x (1-\alpha)^{1-x} &= \exp\left[x \log \frac{\alpha}{1-\alpha} + \log(1-\alpha)\right] \\ &= \exp\left[x\theta - \log(1+e^\theta)\right] \end{aligned}$$

Thus

$$T(x) = x, \;\; \theta = \log \frac{\alpha}{1-\alpha}, \;\; A(\theta) = \log(1+e^\theta)$$

Natural Parameter Form for Gaussian

$$\begin{aligned} p(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)} \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\log\sigma - \frac{x^2}{2\sigma^2} + \frac{\mu x}{\sigma} - \frac{\mu^2}{2\sigma^2}\right) \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(\boldsymbol{\theta}^\top T(x) - \log(\sigma) - \frac{\mu^2}{2\sigma^2}\right) \end{aligned}$$

Thus

$$T(\mathbf{x}) = \begin{pmatrix} x \\ x^2 \end{pmatrix} \quad \boldsymbol{\theta} = \begin{pmatrix} \mu/\sigma^2 \\ -1/(2\sigma^2) \end{pmatrix}$$

$$A(\boldsymbol{\theta}) = \frac{\mu^2}{2\sigma^2} + \log\sigma = -\frac{[\boldsymbol{\theta}]_1^2}{4[\boldsymbol{\theta}]_2} - \frac{1}{2}\log(-2[\boldsymbol{\theta}]_2)$$

- Posterior

$$p(\mathbf{w}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{w})p(\mathbf{w})}{\int p(\mathbf{x}|\mathbf{w})p(\mathbf{w})d\mathbf{w}}$$

- Note: denominator not a function of $\mathbf{w}$ $\rightarrow$ just normalizing term
- Type of a posterior given prior?

$$\underbrace{p(\mathbf{w})}_{parametric} \quad \Rightarrow \quad \underbrace{p(\mathbf{x}|\mathbf{w})}_{parametric} \cdot p(\mathbf{w}) \Rightarrow \quad \text{we get} \quad p(\mathbf{w}|\mathbf{x}) \sim \underbrace{p(\mathbf{x}|\mathbf{w}) \cdot p(\mathbf{w})}_{???}$$

- Conjugacy: require $p(\mathbf{w})$ and $p(\mathbf{w}|\mathbf{x})$ to be of the same form. E.g.

$$\underbrace{p(\mathbf{w})}_{Dirichlet} \quad \Rightarrow \quad \underbrace{p(\mathbf{x}|\mathbf{w})}_{Multinomial} \cdot p(\mathbf{w}) \Rightarrow \underbrace{p(\mathbf{w}|\mathbf{x})}_{Dirichlet}$$

- $p(\mathbf{w})$ and $p(\mathbf{x}|\mathbf{w})$ are then called conjugate distributions

Example: Dirichlet and Multinomial

$$p(\mathbf{w}) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_i w_i^{\alpha_i - 1} \text{ — Dirichlet in } \mathbf{w}, \;\; \Gamma(n) = (n-1)!$$

$$p(\mathbf{x}|\mathbf{w}) = \frac{(\sum_i x_i)!}{x_1! x_2! \ldots x_d!} \prod_{i=1}^{d} w_i^{x_i} \text{ — Multinomial in } \mathbf{x}$$

$$p(\mathbf{w}|\mathbf{x}) \sim p(\mathbf{x}|\mathbf{w}) p(\mathbf{w}) = \text{const} \times \prod_i w_i^{x_i + \alpha_i - 1},$$

which is again Dirichlet, so we must have

$$p(\mathbf{w}|\mathbf{x}) = \frac{\Gamma(\sum_i \alpha_i + x_i)}{\prod_i \Gamma(\alpha_i + x_i)} \prod_i w_i^{x_i + \alpha_i - 1}$$

- **Prior**: Gaussian $e^{-\|\boldsymbol{\mu}-\boldsymbol{\mu}_0\|^2/(2\sigma^2)}$; **Conditional**: $e^{-\|\mathbf{x}-\boldsymbol{\mu}\|^2/(2\sigma^2)}$
- **Prior**: Beta $\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} w^{r-1}(1-w)^{s-1}$; **Conditional**: Bernoulli $w^x(1-w)^{1-x}$
- **Prior**: Dirichlet $\frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod w_i^{\alpha_i - 1}$; **Conditional**: Multinomial $\frac{(\sum_i x_i)!}{\prod x_i!} \prod w_i^{x_i}$
- **Prior**: Inv. Wishart; **Conditional**: Gaussian (cov)

Note: Conjugacy is mutual, e.g.

$$Dirichlet \Rightarrow Multinomial \Rightarrow Dirichlet$$

$$Multinomial \Rightarrow Dirichlet \Rightarrow Multinomial$$

- Linear Basis Function Models

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

  where $\boldsymbol{\phi}(\mathbf{x})$ is a vector of known basis functions $\phi_j(\mathbf{x})$
- 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

- Data model for y (ε is a Gaussian white noise with variance β^{-1})

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

- For $\mathbf{Y}_m = \{y_1, \dots, y_m\}$ and $\mathbf{X}_m = \{\mathbf{x}_1, \dots, \mathbf{x}_m\}$ data likelihood

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

- Data log-likelihood has the form

$$\log p(\mathbf{Y}_m|\mathbf{X}_m, \mathbf{w}, \beta) = \sum_{i=1}^{m} \log \mathcal{N}(y_i|\mathbf{w} \cdot \boldsymbol{\phi}(\mathbf{x}_i), \beta^{-1})$$
$$= \frac{m}{2} \log \beta - \frac{m}{2} \log(2\pi) - \beta E_D(\mathbf{w})$$

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

- Maximizing log-likelihood $\equiv$ minimizing $E_D(\mathbf{w})$:

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

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

- Regularized Least Squares

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

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

Solution has the form

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

- We have a data sample $\mathcal{D}_m = (\mathbf{X}_m, \mathbf{Y}_m)$ from a linear basis function model
- 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

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

- For

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

we get that

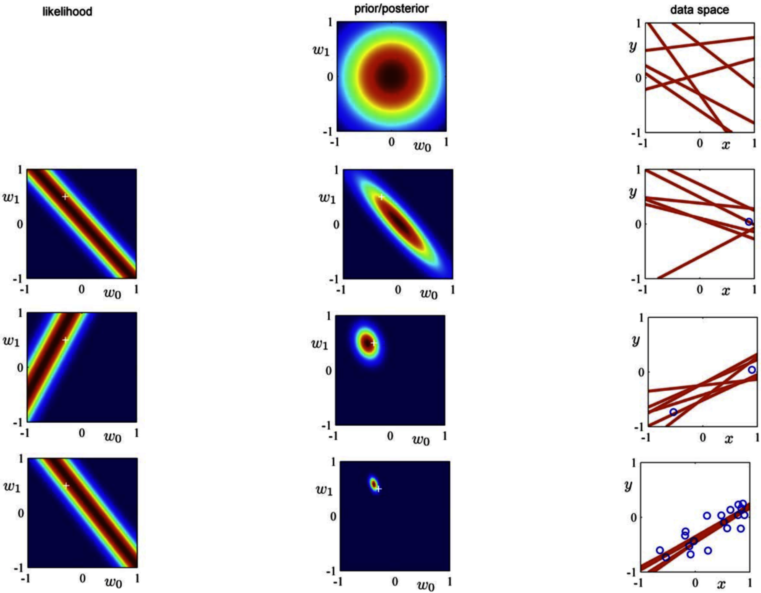
$$p(\mathbf{z}|\mathbf{y}) = \mathcal{N}\left(\mathbf{z}|\boldsymbol{\Sigma}\left[\mathbf{A}^\top \mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\right], \boldsymbol{\Sigma}\right),$$

where

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

- Thus the posterior is defined by

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



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

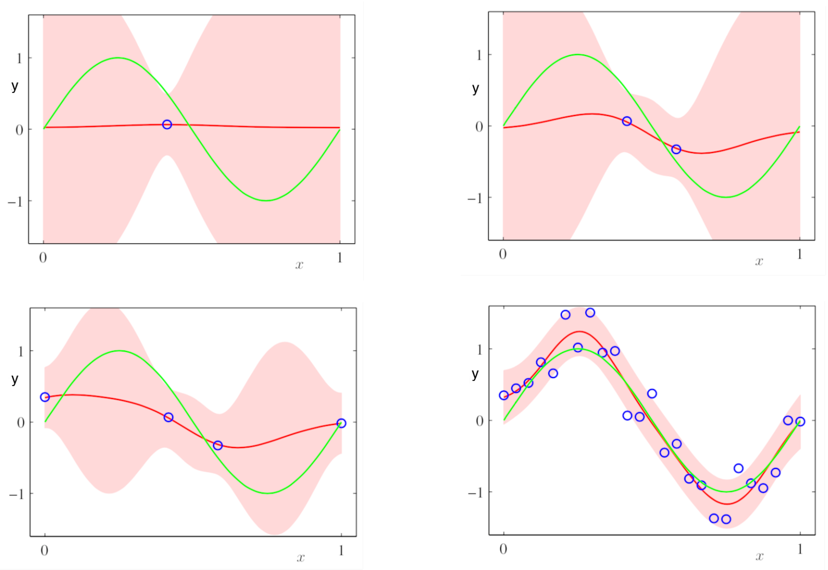
$$\sigma_m^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^\top \mathbf{S}_m \boldsymbol{\phi}(\mathbf{x})$$

- We can use posterior mean for point prediction

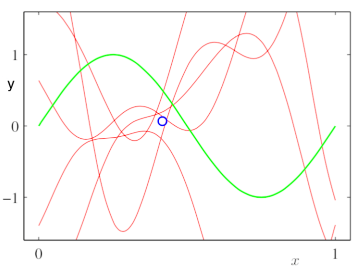
$$\widehat{f}(\mathbf{x}, \mathbf{w}) = \boldsymbol{\omega}_m \cdot \boldsymbol{\phi}(\mathbf{x})^\top$$

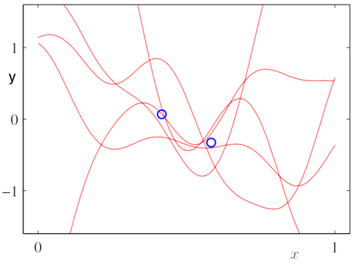
  and posterior variance $\sigma_m^2(\mathbf{x})$ for its uncertainty estimate

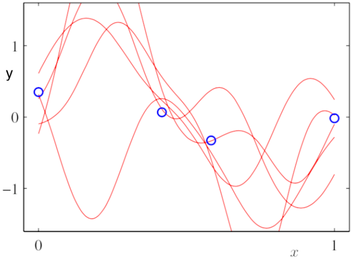
Predictive Distribution

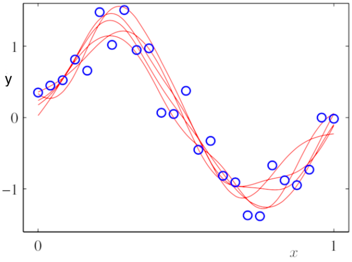


$M = 9$ Gaussian basis functions were used as $\phi(\mathbf{x})$









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

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

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

  Depends on α and β! How to define them? $\Rightarrow$ Full Bayesian approach!

- 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 approximately

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

- 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

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

- So

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

- We denote

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

- We get that

$$\begin{aligned} E(\mathbf{w}) &= E(\mathbf{w} - \boldsymbol{\omega}_m + \boldsymbol{\omega}_m) \\ &= E(\boldsymbol{\omega}_m) + (\mathbf{w} - \boldsymbol{\omega}_m)^\top\mathbf{A}(\mathbf{w} - \boldsymbol{\omega}_m)/2, \end{aligned}$$

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

- Thus

$$\begin{aligned}&\int e^{-E(\mathbf{w})}d\mathbf{w}\\&= e^{-E(\boldsymbol{\omega}_m)}\int e^{\left\{-\frac{1}{2}(\mathbf{w}-\boldsymbol{\omega}_m)^\top \mathbf{A}(\mathbf{w}-\boldsymbol{\omega}_m)\right\}}d\mathbf{w}\\&= e^{-E(\boldsymbol{\omega}_m)}\cdot(2\pi)^{M/2}|\mathbf{A}|^{-1/2}\end{aligned}$$

- Therefore the log-evidence is equal to

$$\begin{aligned}\log p(\mathcal{D}_m|\alpha,\beta) &= \log\left[\left(\frac{\beta}{2\pi}\right)^{\frac{m}{2}}\left(\frac{\alpha}{2\pi}\right)^{\frac{M}{2}} e^{-E(\boldsymbol{\omega}_m)}\cdot(2\pi)^{\frac{M}{2}}|\mathbf{A}|^{-1/2}\right]\\&= \frac{M}{2}\log\alpha + \frac{m}{2}\log\beta - E(\boldsymbol{\omega}_m) - \frac{1}{2}\log|\mathbf{A}| - \frac{m}{2}\log(2\pi)\end{aligned}$$

where

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

- We can maximize $p(\mathcal{D}_m|\alpha,\beta)$ w.r.t. (α,β)

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

- Here

$$\begin{aligned} \mathbf{A} = \mathsf{S}_m^{-1} &= \alpha^{-1}\mathbf{I} + \beta\boldsymbol{\Phi}^\top\boldsymbol{\Phi} \in \mathbb{R}^{M\times M}, \\ \boldsymbol{\omega}_m &= \beta\mathsf{S}_m\boldsymbol{\Phi}^\top\mathbf{Y}_m \\ E(\boldsymbol{\omega}_m) &= \frac{\beta}{2}\|\mathbf{Y}_m - \boldsymbol{\Phi}\cdot\boldsymbol{\omega}_m^\top\|^2 + \frac{\alpha}{2}\boldsymbol{\omega}_m\cdot\boldsymbol{\omega}_m^\top \end{aligned}$$

- Also we can estimate model complexity (e.g. order of a polynomial M) by optimizing $\log p(\mathcal{D}_m|\alpha,\beta)$

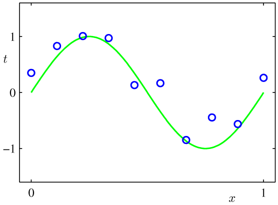


Figure – Plot of a training data

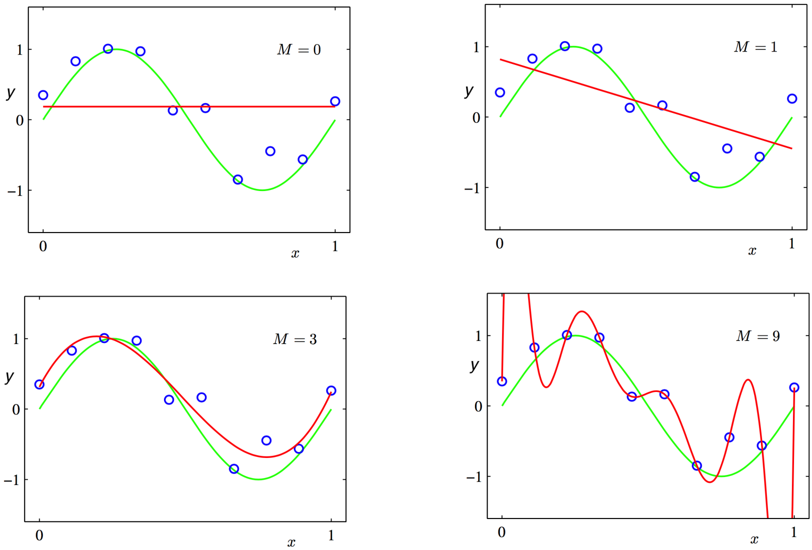
- We fit a model

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

by minimizing the error

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

Plots of polynomials having various orders M



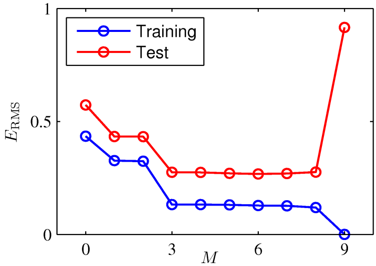


Figure – $E_{RMS} = \sqrt{2E(\mathbf{w}^*)/m}$ versus M

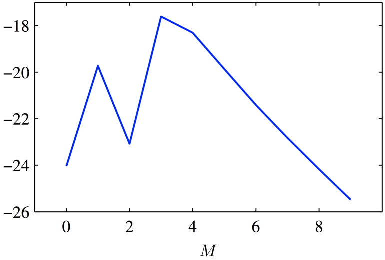


Figure – Plot of log-evidence $\log p(\mathcal{D}_m|\alpha, \beta)$ versus M for a fixed $\alpha = 5 \times 10^{-3}$

- Let us illustrate log-evidence $\log p(\mathcal{D}_m|\alpha, \beta)$ optimization w.r.t. α
- We set β to its true value $(= 11.1)$
- We consider a polynomial model of order $M = 9$
- We plot dependence of
  - log-evidence $\log p(\mathcal{D}_m|\alpha, \beta)$
  - test error

  on α

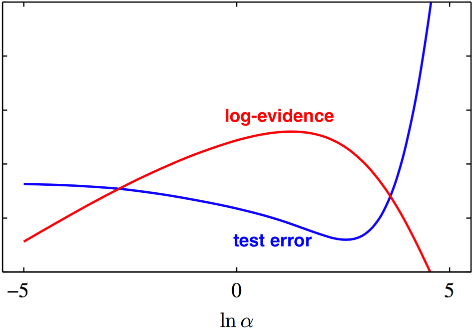


Figure – Dependence of $\log p(\mathcal{D}_m|\alpha, \beta)$ and test error on α

Maximizing the evidence function

- Efficient optimization of $p(\mathcal{D}_m|\alpha,\beta)$?
- Let us first maximize $\log p(\mathcal{D}_m|\alpha,\beta)$ w.r.t. α for a fixed β

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

- Let us differentiate $\log p(\mathcal{D}|\alpha,\beta)$ w.r.t. α
- Let us consider the eigenvector equation

$$(\beta\boldsymbol{\Phi}^\top\boldsymbol{\Phi})\mathbf{u}_i = \lambda_i\mathbf{u}_i$$

- $\mathbf{A} = \alpha\mathbf{I} + \beta\boldsymbol{\Phi}^\top\boldsymbol{\Phi}$ has eigenvalues $\alpha+\lambda_i$
- We get that

$$|\mathbf{A}| = \prod_{i=1}^{M}(\lambda_i+\alpha)$$

$$\frac{d}{d\alpha}\log|\mathbf{A}| = \frac{d}{d\alpha}\log\prod_i(\lambda_i+\alpha) =$$

$$= \frac{d}{d\alpha}\sum_i\log(\lambda_i+\alpha) = \sum_i\frac{1}{\lambda_i+\alpha}$$

- The stationary points of $\log p(\mathcal{D}|\alpha, \beta)$ w.r.t. α satisfy

$$0 = \frac{M}{2\alpha} - \frac{1}{2}\boldsymbol{\omega}_m \cdot \boldsymbol{\omega}_m^\top - \frac{1}{2}\sum_i \frac{1}{\lambda_i + \alpha}$$

$$\alpha \boldsymbol{\omega}_m \cdot \boldsymbol{\omega}_m^\top = M - \alpha \sum_i \frac{1}{(\lambda_i + \alpha)}$$

- Let us denote

$$\gamma = M - \sum_{i=1}^{M} \frac{\alpha}{\lambda_i + \alpha}$$

$$\gamma = \sum_{i=1}^{M} \frac{\lambda_i + \alpha}{\lambda_i + \alpha} - \sum_{i=1}^{M} \frac{\alpha}{\alpha + \lambda_i} = \sum_{i=1}^{M} \frac{\lambda_i}{\lambda_i + \alpha}$$

- Thus we get that

$$\alpha \boldsymbol{\omega}_m \cdot \boldsymbol{\omega}_m^\top = \gamma, \ \gamma = \sum_{i=1}^{M} \frac{\lambda_i}{\lambda_i + \alpha}$$

$$\alpha = \frac{\gamma}{\boldsymbol{\omega}_m \cdot \boldsymbol{\omega}_m^\top}$$

Maximizing the evidence function

- We adopt an iterative process:
  - We make an initial choice for α
  - We use this to find $\boldsymbol{\omega}_m$

$$\boldsymbol{\omega}_m = \beta \mathbf{S}_m \boldsymbol{\Phi}^\top \mathbf{Y}_m, \text{ with}$$
$$\mathbf{A} = \mathbf{S}_m^{-1} = \alpha^{-1}\mathbf{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi}$$

  - We evaluate γ

$$\gamma = \sum_{i=1}^{M} \frac{\lambda_i}{\lambda_i + \alpha},$$

  - We re-estimate α

$$\alpha = \frac{\gamma}{\boldsymbol{\omega}_m \cdot \boldsymbol{\omega}_m^\top},$$

etc.

- Let us consider optimization w.r.t. β
- Recall the eigenvector equation

$$(\beta\boldsymbol{\Phi}^\top\boldsymbol{\Phi})\mathbf{u}_i = \lambda_i\mathbf{u}_i$$
$$\frac{d\lambda_i}{d\beta}\mathbf{u}_i = \frac{1}{\beta}(\beta\boldsymbol{\Phi}^\top\boldsymbol{\Phi})\mathbf{u}_i = \frac{1}{\beta}\lambda_i\mathbf{u}_i$$

- Thus we get that $\frac{d\lambda_i}{d\beta} = \frac{\lambda_i}{\beta}$. Then

$$\frac{d}{d\beta}\log|\mathbf{A}| = \frac{d}{d\beta}\sum_i \log(\lambda_i+\alpha) = \frac{1}{\beta}\sum_i \frac{\lambda_i}{\lambda_i+\alpha} = \frac{\gamma}{\beta}$$

- We know that

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

and

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

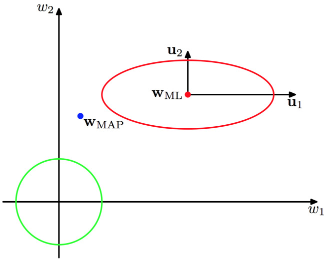
- The stationary points of $\log p(\mathcal{D}|\alpha, \beta)$ w.r.t. β

$$0 = \frac{m}{2\beta} - \frac{1}{2}\sum_{i=1}^{m}(y_i - \boldsymbol{\omega}_m \cdot \boldsymbol{\phi}(\mathbf{x}_i)^\top)^2 - \frac{\gamma}{2\beta}$$

$$\frac{1}{\beta} = \frac{1}{m-\gamma}\sum_{i=1}^{m}(y_i - \boldsymbol{\omega}_m \cdot \boldsymbol{\phi}(\mathbf{x}_i)^\top)^2$$

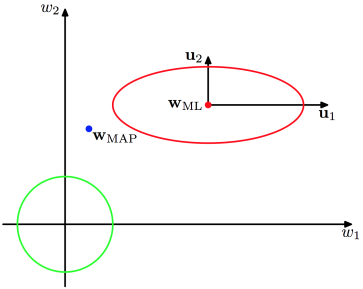
- We adopt an iterative process:
  - We make an initial choice for β
  - We use this to find $\boldsymbol{\omega}_m$ and γ
  - We re-estimate β, etc.

Effective number of parameters



- Contours of the likelihood function (red) and the prior (green) in which the axes in parameter space have been rotated to align with the eigenvectors $\mathbf{u}_i$ of the Hessian
- For $\alpha = 0$ the mode of the posterior $\mathbf{w}_{MAP} = \mathbf{w}_{ML}$; for non-zero α the mode is at $\mathbf{w}_{MAP} = \boldsymbol{\omega}_m$

Effective number of parameters



- Recall that $\boldsymbol{\omega}_m = \beta \mathbf{S}_m \boldsymbol{\Phi}^\top \mathbf{Y}_m$ with $\mathbf{S}_m^{-1} = \alpha \mathbf{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi} \in \mathbb{R}^{M \times M}$
- Variance of the components of the estimate $\mathbf{w}_{ML}$ is inversely proportional to eigenvalues of $\lambda_i(\boldsymbol{\Phi}^\top \cdot \boldsymbol{\Phi})$. Sizes of the axes of the ellipsoid is inversely proportional to λ_i
- In the direction w_1 the eigenvalue λ_1 is small compared with α and so $\lambda_1/(\lambda_1 + \alpha)$ is ≈ 0, and so $w_{1,MAP} \approx 0$
- In the direction w_2 the eigenvalue $\lambda_2 \gg \alpha$ and so $\lambda_2/(\lambda_2 + \alpha) \approx 1$, i.e. $w_{2,MAP} \approx w_{2,MLE}$
- Thus $0 \le \gamma \le M$. The effective number of parameters determined by the data is γ, with remaining $M - \gamma$ param. set to small values by the prior

Effective number of parameters

- Let us consider the limit $m \gg M$
- Recall that

$$\gamma = \sum_{i=1}^{M} \frac{\lambda_i}{\lambda_i + \alpha}$$

- Since $\boldsymbol{\Phi}^\top \boldsymbol{\Phi} = \sum_{i=1}^{m} \boldsymbol{\phi}(\mathbf{x}_i)^\top \boldsymbol{\phi}(\mathbf{x}_i)$ involves an implicit sum over data points, so $\lambda_i = \lambda_i(\boldsymbol{\Phi}^\top \boldsymbol{\Phi})$ increase with the size of the data set.
- In this case $\lambda_i \gg \alpha \ \forall i$ and $\gamma \approx \sum_{i=1}^{M} 1 = M$.
- Since

$$\alpha = \frac{\gamma}{\boldsymbol{\omega}_m \cdot \boldsymbol{\omega}_m^\top},$$

$$\frac{1}{\beta} = \frac{1}{m - \gamma} \sum_{i=1}^{m} (y_i - \boldsymbol{\omega}_m \cdot \boldsymbol{\phi}(\mathbf{x}_i)^\top)^2$$

we get the re-estimation equations

$$\alpha = \frac{M}{2E_W(\boldsymbol{\omega}_m)}, \ \beta = \frac{m}{2E_D(\boldsymbol{\omega}_m)}$$

with

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