02477 – Bayesian Machine Learning: Lecture 3

Michael Riis Andersen

Technical University of Denmark,
DTU Compute, Department of Applied Math and Computer Science

Outline

1. Recap: Probabilistic machine learning
2. Recap: Multivariate Gaussian distributions
3. Recap: Linear regression and supervised learning
4. Bayesian linear regression
5. The posterior predictive distribution
6. Dealing with hyperparameters

Recap: Probabilistic machine learning

A probabilistic perspective on making predictions

Product rule
$p(\boldsymbol{a}, \boldsymbol{b}) = p(\boldsymbol{b}|\boldsymbol{a})p(\boldsymbol{a})$

Sum rule
$p(\boldsymbol{b}) = \int p(\boldsymbol{a}, \boldsymbol{b})\mathrm{d}\boldsymbol{a}$

Conditional
$p(\boldsymbol{a}|\boldsymbol{b}) = \frac{p(\boldsymbol{a},\boldsymbol{b})}{p(\boldsymbol{b})}$

Conditional independence
$p(\boldsymbol{a}, \boldsymbol{b}|\boldsymbol{c}) = p(\boldsymbol{a}|\boldsymbol{c})p(\boldsymbol{b}|\boldsymbol{c})$

Goal: Given some data y, what can we say about a new observation y^*?

A probabilistic perspective on making predictions

Product rule
$p(\boldsymbol{a}, \boldsymbol{b}) = p(\boldsymbol{b}|\boldsymbol{a})p(\boldsymbol{a})$

Sum rule
$p(\boldsymbol{b}) = \int p(\boldsymbol{a}, \boldsymbol{b})\mathrm{d}\boldsymbol{a}$

Conditional
$p(\boldsymbol{a}|\boldsymbol{b}) = \frac{p(\boldsymbol{a},\boldsymbol{b})}{p(\boldsymbol{b})}$

Conditional independence
$p(\boldsymbol{a}, \boldsymbol{b}|\boldsymbol{c}) = p(\boldsymbol{a}|\boldsymbol{c})p(\boldsymbol{b}|\boldsymbol{c})$

Goal: Given some data y, what can we say about a new observation y^*?

- Step 1: Formulate *joint distribution* for *all variables* of interests

$$p(y^*, y, \theta) = p(y^*, y|\theta)p(\theta) =$$

A probabilistic perspective on making predictions

Product rule
$p(\boldsymbol{a}, \boldsymbol{b}) = p(\boldsymbol{b}|\boldsymbol{a})p(\boldsymbol{a})$

Sum rule
$p(\boldsymbol{b}) = \int p(\boldsymbol{a}, \boldsymbol{b})\mathrm{d}\boldsymbol{a}$

Conditional
$p(\boldsymbol{a}|\boldsymbol{b}) = \frac{p(\boldsymbol{a},\boldsymbol{b})}{p(\boldsymbol{b})}$

Conditional independence
$p(\boldsymbol{a}, \boldsymbol{b}|\boldsymbol{c}) = p(\boldsymbol{a}|\boldsymbol{c})p(\boldsymbol{b}|\boldsymbol{c})$

Goal: Given some data y, what can we say about a new observation y^*?

- Step 1: Formulate *joint distribution* for *all variables* of interests

$$p(y^*, y, \theta) = p(y^*, y|\theta)p(\theta) = p(y^*|\theta)p(y|\theta)p(\theta)$$

A probabilistic perspective on making predictions

Product rule
$p(\boldsymbol{a}, \boldsymbol{b}) = p(\boldsymbol{b}|\boldsymbol{a})p(\boldsymbol{a})$

Sum rule
$p(\boldsymbol{b}) = \int p(\boldsymbol{a}, \boldsymbol{b})\mathrm{d}\boldsymbol{a}$

Conditional
$p(\boldsymbol{a}|\boldsymbol{b}) = \frac{p(\boldsymbol{a}, \boldsymbol{b})}{p(\boldsymbol{b})}$

Conditional independence
$p(\boldsymbol{a}, \boldsymbol{b}|\boldsymbol{c}) = p(\boldsymbol{a}|\boldsymbol{c})p(\boldsymbol{b}|\boldsymbol{c})$

Goal: Given some data y, what can we say about a new observation y^*?

- Step 1: Formulate *joint distribution* for *all variables* of interests

$$p(y^*, y, \theta) = p(y^*, y|\theta)p(\theta) = p(y^*|\theta)p(y|\theta)p(\theta)$$

- Step 2: *Condition* on the *observed data* y

$$p(y^*, \theta|y) = \frac{p(y*, y, \theta)}{p(y)} =$$

A probabilistic perspective on making predictions

Product rule
$p(\boldsymbol{a}, \boldsymbol{b}) = p(\boldsymbol{b}|\boldsymbol{a})p(\boldsymbol{a})$

Sum rule
$p(\boldsymbol{b}) = \int p(\boldsymbol{a}, \boldsymbol{b})\mathrm{d}\boldsymbol{a}$

Conditional
$p(\boldsymbol{a}|\boldsymbol{b}) = \frac{p(\boldsymbol{a},\boldsymbol{b})}{p(\boldsymbol{b})}$

Conditional independence
$p(\boldsymbol{a}, \boldsymbol{b}|\boldsymbol{c}) = p(\boldsymbol{a}|\boldsymbol{c})p(\boldsymbol{b}|\boldsymbol{c})$

Goal: Given some data y, what can we say about a new observation y^*?

- Step 1: Formulate *joint distribution* for *all variables* of interests

$$p(y^*, y, \theta) = p(y^*, y|\theta)p(\theta) = p(y^*|\theta)p(y|\theta)p(\theta)$$

- Step 2: *Condition* on the *observed data* y

$$p(y^*, \theta|y) = \frac{p(y*, y, \theta)}{p(y)} = \frac{p(y^*|\theta)p(y|\theta)p(\theta)}{p(y)}$$

A probabilistic perspective on making predictions

Product rule
$p(\boldsymbol{a}, \boldsymbol{b}) = p(\boldsymbol{b}|\boldsymbol{a})p(\boldsymbol{a})$

Sum rule
$p(\boldsymbol{b}) = \int p(\boldsymbol{a}, \boldsymbol{b})\mathrm{d}\boldsymbol{a}$

Conditional
$p(\boldsymbol{a}|\boldsymbol{b}) = \frac{p(\boldsymbol{a},\boldsymbol{b})}{p(\boldsymbol{b})}$

Conditional independence
$p(\boldsymbol{a}, \boldsymbol{b}|\boldsymbol{c}) = p(\boldsymbol{a}|\boldsymbol{c})p(\boldsymbol{b}|\boldsymbol{c})$

Goal: Given some data y, what can we say about a new observation y^*?

- Step 1: Formulate *joint distribution* for *all variables* of interests

$$p(y^*, y, \theta) = p(y^*, y|\theta)p(\theta) = p(y^*|\theta)p(y|\theta)p(\theta)$$

- Step 2: *Condition* on the *observed data* y

$$p(y^*, \theta|y) = \frac{p(y*, y, \theta)}{p(y)} = \frac{p(y^*|\theta)p(y|\theta)p(\theta)}{p(y)}$$

- Step 3: *Marginalize* out parameter θ using the *sum rule* to get the *posterior predictive distribution*

$$p(y^*|y) = \int p(y^*, \theta|y)\mathrm{d}\theta =$$

A probabilistic perspective on making predictions

Product rule
$p(\boldsymbol{a}, \boldsymbol{b}) = p(\boldsymbol{b}|\boldsymbol{a})p(\boldsymbol{a})$

Sum rule
$p(\boldsymbol{b}) = \int p(\boldsymbol{a}, \boldsymbol{b})\mathrm{d}\boldsymbol{a}$

Conditional
$p(\boldsymbol{a}|\boldsymbol{b}) = \frac{p(\boldsymbol{a},\boldsymbol{b})}{p(\boldsymbol{b})}$

Conditional independence
$p(\boldsymbol{a}, \boldsymbol{b}|\boldsymbol{c}) = p(\boldsymbol{a}|\boldsymbol{c})p(\boldsymbol{b}|\boldsymbol{c})$

Goal: Given some data y, what can we say about a new observation y^*?

- Step 1: Formulate *joint distribution* for *all variables* of interests

$$p(y^*, y, \theta) = p(y^*, y|\theta)p(\theta) = p(y^*|\theta)p(y|\theta)p(\theta)$$

- Step 2: *Condition* on the *observed data* y

$$p(y^*, \theta|y) = \frac{p(y*, y, \theta)}{p(y)} = \frac{p(y^*|\theta)p(y|\theta)p(\theta)}{p(y)}$$

- Step 3: *Marginalize* out parameter θ using the *sum rule* to get the *posterior predictive distribution*

$$p(y^*|y) = \int p(y^*, \theta|y)\mathrm{d}\theta = \int \frac{p(y^*|\theta)p(y|\theta)p(\theta)}{p(y)}\mathrm{d}\theta =$$

A probabilistic perspective on making predictions

Product rule
$p(\boldsymbol{a}, \boldsymbol{b}) = p(\boldsymbol{b}|\boldsymbol{a})p(\boldsymbol{a})$

Sum rule
$p(\boldsymbol{b}) = \int p(\boldsymbol{a}, \boldsymbol{b})\mathrm{d}\boldsymbol{a}$

Conditional
$p(\boldsymbol{a}|\boldsymbol{b}) = \frac{p(\boldsymbol{a},\boldsymbol{b})}{p(\boldsymbol{b})}$

Conditional independence
$p(\boldsymbol{a}, \boldsymbol{b}|\boldsymbol{c}) = p(\boldsymbol{a}|\boldsymbol{c})p(\boldsymbol{b}|\boldsymbol{c})$

Goal: Given some data y, what can we say about a new observation y^*?

- Step 1: Formulate *joint distribution* for *all variables* of interests

$$p(y^*, y, \theta) = p(y^*, y|\theta)p(\theta) = p(y^*|\theta)p(y|\theta)p(\theta)$$

- Step 2: *Condition* on the *observed data* y

$$p(y^*, \theta|y) = \frac{p(y*, y, \theta)}{p(y)} = \frac{p(y^*|\theta)p(y|\theta)p(\theta)}{p(y)}$$

- Step 3: *Marginalize* out parameter θ using the *sum rule* to get the *posterior predictive distribution*

$$p(y^*|y) = \int p(y^*, \theta|y)\mathrm{d}\theta = \int \frac{p(y^*|\theta)p(y|\theta)p(\theta)}{p(y)}\mathrm{d}\theta = \int p(y^*|\theta)p(\theta|y)\mathrm{d}\theta$$

A probabilistic perspective on making predictions

Product rule
$p(\boldsymbol{a}, \boldsymbol{b}) = p(\boldsymbol{b}|\boldsymbol{a})p(\boldsymbol{a})$

Sum rule
$p(\boldsymbol{b}) = \int p(\boldsymbol{a}, \boldsymbol{b})\mathrm{d}\boldsymbol{a}$

Conditional
$p(\boldsymbol{a}|\boldsymbol{b}) = \frac{p(\boldsymbol{a},\boldsymbol{b})}{p(\boldsymbol{b})}$

Conditional independence
$p(\boldsymbol{a}, \boldsymbol{b}|\boldsymbol{c}) = p(\boldsymbol{a}|\boldsymbol{c})p(\boldsymbol{b}|\boldsymbol{c})$

Goal: Given some data y, what can we say about a new observation y^*?

- Step 1: Formulate *joint distribution* for *all variables* of interests

$$p(y^*, y, \theta) = p(y^*, y|\theta)p(\theta) = p(y^*|\theta)p(y|\theta)p(\theta)$$

- Step 2: *Condition* on the *observed data* y

$$p(y^*, \theta|y) = \frac{p(y*, y, \theta)}{p(y)} = \frac{p(y^*|\theta)p(y|\theta)p(\theta)}{p(y)}$$

- Step 3: *Marginalize* out parameter θ using the *sum rule* to get the *posterior predictive distribution*

$$p(y^*|y) = \int p(y^*, \theta|y)\mathrm{d}\theta = \int \frac{p(y^*|\theta)p(y|\theta)p(\theta)}{p(y)}\mathrm{d}\theta = \int p(y^*|\theta)p(\theta|y)\mathrm{d}\theta = \mathbb{E}_{p(\theta|y)}\left[p(y^*|\theta)\right]$$

A probabilistic perspective on making predictions

Product rule
$p(\boldsymbol{a}, \boldsymbol{b}) = p(\boldsymbol{b}|\boldsymbol{a})p(\boldsymbol{a})$

Sum rule
$p(\boldsymbol{b}) = \int p(\boldsymbol{a}, \boldsymbol{b})\mathrm{d}\boldsymbol{a}$

Conditional
$p(\boldsymbol{a}|\boldsymbol{b}) = \frac{p(\boldsymbol{a},\boldsymbol{b})}{p(\boldsymbol{b})}$

Conditional independence
$p(\boldsymbol{a}, \boldsymbol{b}|\boldsymbol{c}) = p(\boldsymbol{a}|\boldsymbol{c})p(\boldsymbol{b}|\boldsymbol{c})$

Goal: Given some data y, what can we say about a new observation y^*?

- Step 1: Formulate *joint distribution* for *all variables* of interests

$$p(y^*, y, \theta) = p(y^*, y|\theta)p(\theta) = p(y^*|\theta)p(y|\theta)p(\theta)$$

- Step 2: *Condition* on the *observed data* y

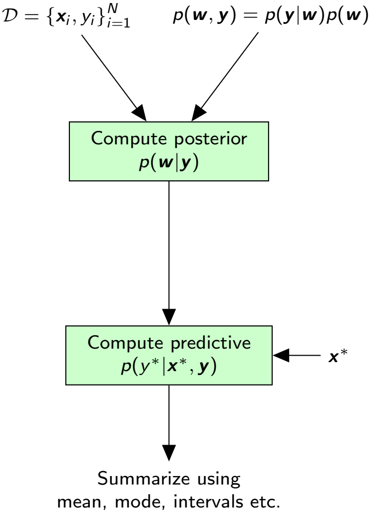
$$p(y^*, \theta|y) = \frac{p(y*, y, \theta)}{p(y)} = \frac{p(y^*|\theta)p(y|\theta)p(\theta)}{p(y)}$$

- Step 3: *Marginalize* out parameter θ using the *sum rule* to get the *posterior predictive distribution*

$$p(y^*|y) = \int p(y^*, \theta|y)\mathrm{d}\theta = \int \frac{p(y^*|\theta)p(y|\theta)p(\theta)}{p(y)}\mathrm{d}\theta = \int p(y^*|\theta)p(\theta|y)\mathrm{d}\theta = \mathbb{E}_{p(\theta|y)}\left[p(y^*|\theta)\right]$$

- **Key take-away**: To reason about *y^* given y*, we need to *average the likelihood* for y^* wrt. to the *posterior distribution* $p(\theta|y)$.

Bayesian inference for supervised learning



- Same principles for linear regression, logistic regression, neural networks etc. etc.

Recap: Multivariate Gaussian distributions

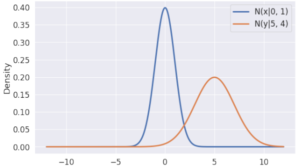
Univariate normal distribution

- The *normal distribution* (also known as the Gaussian)

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

- Two parameters: $\mu = \mathbb{E}[x]$ and $\sigma^2 = \mathbb{V}[x]$

- Widely due to Central limit theorems, maximum entropy principle, relation to least squares minimization, nice mathematical properties



Univariate normal distribution

- The *normal distribution* (also known as the Gaussian)

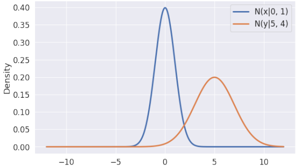
$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- Two parameters: $\mu = \mathbb{E}[x]$ and $\sigma^2 = \mathbb{V}[x]$

- Widely due to Central limit theorems, maximum entropy principle, relation to least squares minimization, nice mathematical properties

- Closed under affine transformations

$$x \sim \mathcal{N}(m, v) \quad \Rightarrow \quad a + bx \sim \mathcal{N}(a + bm, b^2 v)$$



Univariate normal distribution

- The *normal distribution* (also known as the Gaussian)

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

- Two parameters: $\mu = \mathbb{E}[x]$ and $\sigma^2 = \mathbb{V}[x]$

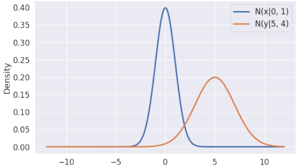
- Widely due to Central limit theorems, maximum entropy principle, relation to least squares minimization, nice mathematical properties

- Closed under affine transformations

$$x \sim \mathcal{N}(m, v) \quad \Rightarrow \quad a + bx \sim \mathcal{N}(a + bm, b^2 v)$$

- Let $x \sim \mathcal{N}(m_x, v_x)$ and $y \sim \mathcal{N}(m_y, v_y)$ for $x \perp y$, then

$$x + y \sim \mathcal{N}(m_x + m_y, v_x + v_y)$$



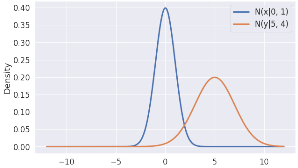
Univariate normal distribution

- The *normal distribution* (also known as the Gaussian)

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

- Two parameters: $\mu = \mathbb{E}[x]$ and $\sigma^2 = \mathbb{V}[x]$

- Widely due to Central limit theorems, maximum entropy principle, relation to least squares minimization, nice mathematical properties



- Closed under affine transformations

$$x \sim \mathcal{N}(m, v) \quad \Rightarrow \quad a + bx \sim \mathcal{N}(a + bm, b^2 v)$$

$$x \sim N(0,1)$$
$$y = 5 + 2x$$

- Let $x \sim \mathcal{N}(m_x, v_x)$ and $y \sim \mathcal{N}(m_y, v_y)$ for $x \perp y$, then

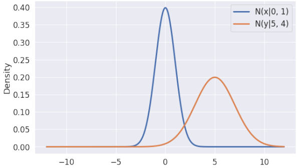
$$x + y \sim \mathcal{N}(m_x + m_y, v_x + v_y)$$

Univariate normal distribution

- The *normal distribution* (also known as the Gaussian)

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

- Two parameters: $\mu = \mathbb{E}[x]$ and $\sigma^2 = \mathbb{V}[x]$



- Widely due to Central limit theorems, maximum entropy principle, relation to least squares minimization, nice mathematical properties

- Closed under affine transformations

$$x \sim \mathcal{N}(m, v) \quad \Rightarrow \quad a + bx \sim \mathcal{N}(a + bm, b^2 v)$$

$$x \sim N(0, 1)$$
$$y = 5 + 2x$$

- Let $x \sim \mathcal{N}(m_x, v_x)$ and $y \sim \mathcal{N}(m_y, v_y)$ for $x \perp y$, then

$$x + y \sim \mathcal{N}(m_x + m_y, v_x + v_y)$$

- Functional form: The *logarithm of a Gaussian density* is a *second order polynomial*

$$\ln \mathcal{N}(x|\mu, \sigma^2) = -\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x + K$$

The functional form of a Gaussian distribution I

- Recall we discussed the *functional form* of a Beta distribution

The functional form of a Gaussian distribution I

- Recall we discussed the *functional form* of a Beta distribution
- Let's derive the functional form of the Gaussian

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

The functional form of a Gaussian distribution II: example

- Take-away: the *the functional form* of a univariate Gaussian is

$$\ln \mathcal{N}(x|\mu, \sigma^2) = -\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x + K$$

The functional form of a Gaussian distribution II: example

- Take-away: the *the functional form* of a univariate Gaussian is

$$\ln \mathcal{N}(x|\mu, \sigma^2) = -\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x + K$$

- Example: suppose we are given the following log density for a random variable x

$$\ln p(x) = -\frac{1}{4}x^2 + 2x + K$$

The functional form of a Gaussian distribution II: example

- Take-away: the *the functional form* of a univariate Gaussian is

$$\ln \mathcal{N}(x|\mu, \sigma^2) = -\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x + K$$

- Example: suppose we are given the following log density for a random variable x

$$\ln p(x) = -\frac{1}{4}x^2 + 2x + K$$

- We recognize the 2nd order polynomial and conclude that $p(x)$ must be Gaussian

The functional form of a Gaussian distribution II: example

- Take-away: the *the functional form* of a univariate Gaussian is

$$\ln \mathcal{N}(x|\mu, \sigma^2) = -\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x + K$$

- Example: suppose we are given the following log density for a random variable x

$$\ln p(x) = -\frac{1}{4}x^2 + 2x + K$$

- We recognize the 2nd order polynomial and conclude that $p(x)$ must be Gaussian

- We determine the *variance* by matching the coefficient for 2nd order term

$$-\frac{1}{2\sigma^2} = -\frac{1}{4}$$

The functional form of a Gaussian distribution II: example

- Take-away: the *the functional form* of a univariate Gaussian is

$$\ln \mathcal{N}(x|\mu, \sigma^2) = -\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x + K$$

- Example: suppose we are given the following log density for a random variable x

$$\ln p(x) = -\frac{1}{4}x^2 + 2x + K$$

- We recognize the 2nd order polynomial and conclude that $p(x)$ must be Gaussian

- We determine the *variance* by matching the coefficient for 2nd order term

$$-\frac{1}{2\sigma^2} = -\frac{1}{4} \quad \Rightarrow \quad \frac{1}{\sigma^2} = \frac{1}{2}$$

The functional form of a Gaussian distribution II: example

- Take-away: the *the functional form* of a univariate Gaussian is

$$\ln \mathcal{N}(x|\mu, \sigma^2) = -\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x + K$$

- Example: suppose we are given the following log density for a random variable x

$$\ln p(x) = -\frac{1}{4}x^2 + 2x + K$$

- We recognize the 2nd order polynomial and conclude that $p(x)$ must be Gaussian

- We determine the *variance* by matching the coefficient for 2nd order term

$$-\frac{1}{2\sigma^2} = -\frac{1}{4} \quad \Rightarrow \quad \frac{1}{\sigma^2} = \frac{1}{2} \quad \Rightarrow \quad \sigma^2 = 2$$

The functional form of a Gaussian distribution II: example

- Take-away: the *the functional form* of a univariate Gaussian is

$$\ln \mathcal{N}(x|\mu, \sigma^2) = -\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x + K$$

- Example: suppose we are given the following log density for a random variable x

$$\ln p(x) = -\frac{1}{4}x^2 + 2x + K$$

- We recognize the 2nd order polynomial and conclude that $p(x)$ must be Gaussian

- We determine the *variance* by matching the coefficient for 2nd order term

$$-\frac{1}{2\sigma^2} = -\frac{1}{4} \quad \Rightarrow \quad \frac{1}{\sigma^2} = \frac{1}{2} \quad \Rightarrow \quad \sigma^2 = 2$$

- We determine the *mean* by matching coefficient for 1st order term

$$\frac{\mu}{\sigma^2} = 2$$

The functional form of a Gaussian distribution II: example

- Take-away: the *the functional form* of a univariate Gaussian is

$$\ln \mathcal{N}(x|\mu, \sigma^2) = -\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x + K$$

- Example: suppose we are given the following log density for a random variable x

$$\ln p(x) = -\frac{1}{4}x^2 + 2x + K$$

- We recognize the 2nd order polynomial and conclude that $p(x)$ must be Gaussian

- We determine the *variance* by matching the coefficient for 2nd order term

$$-\frac{1}{2\sigma^2} = -\frac{1}{4} \quad \Rightarrow \quad \frac{1}{\sigma^2} = \frac{1}{2} \quad \Rightarrow \quad \sigma^2 = 2$$

- We determine the *mean* by matching coefficient for 1st order term

$$\frac{\mu}{\sigma^2} = 2$$

The functional form of a Gaussian distribution II: example

- Take-away: the *the functional form* of a univariate Gaussian is

$$\ln \mathcal{N}(x|\mu, \sigma^2) = -\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x + K$$

- Example: suppose we are given the following log density for a random variable x

$$\ln p(x) = -\frac{1}{4}x^2 + 2x + K$$

- We recognize the 2nd order polynomial and conclude that $p(x)$ must be Gaussian

- We determine the *variance* by matching the coefficient for 2nd order term

$$-\frac{1}{2\sigma^2} = -\frac{1}{4} \quad \Rightarrow \quad \frac{1}{\sigma^2} = \frac{1}{2} \quad \Rightarrow \quad \sigma^2 = 2$$

- We determine the *mean* by matching coefficient for 1st order term

$$\frac{\mu}{\sigma^2} = 2 \quad \Rightarrow \quad \mu = 4$$

The functional form of a Gaussian distribution II: example

- Take-away: the *the functional form* of a univariate Gaussian is

$$\ln \mathcal{N}(x|\mu, \sigma^2) = -\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x + K$$

- Example: suppose we are given the following log density for a random variable x

$$\ln p(x) = -\frac{1}{4}x^2 + 2x + K$$

- We recognize the 2nd order polynomial and conclude that $p(x)$ must be Gaussian

- We determine the *variance* by matching the coefficient for 2nd order term

$$-\frac{1}{2\sigma^2} = -\frac{1}{4} \quad \Rightarrow \quad \frac{1}{\sigma^2} = \frac{1}{2} \quad \Rightarrow \quad \sigma^2 = 2$$

- We determine the *mean* by matching coefficient for 1st order term

$$\frac{\mu}{\sigma^2} = 2 \quad \Rightarrow \quad \mu = 4$$

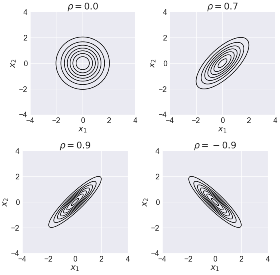
- Therefore, we conclude $p(x) = \mathcal{N}(x|4, 2)$.

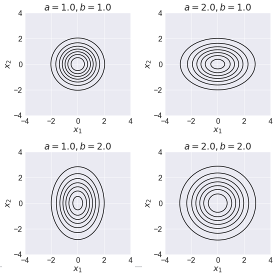
The multivariate normal distribution

- The *multivariate normal distribution*

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \mu)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \mu)\right)$$

- Two parameters: $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{x}]$ and $\boldsymbol{\Sigma} = \text{cov}[\boldsymbol{x}]$





The multivariate normal distribution

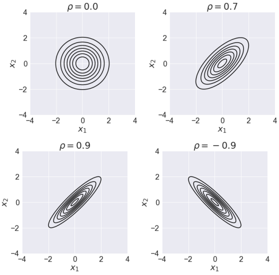
- The *multivariate normal distribution*

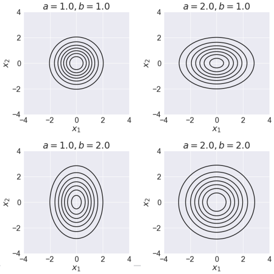
$$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\mu)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\mu)\right)$$

- Two parameters: $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{x}]$ and $\boldsymbol{\Sigma} = \text{cov}[\boldsymbol{x}]$

- Covariance matrix for $D = 2$

$$\boldsymbol{\Sigma} = \begin{bmatrix} a & c \\ c & b \end{bmatrix}$$





The multivariate normal distribution

- The *multivariate normal distribution*

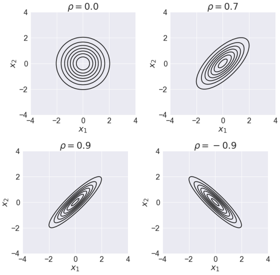
$$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\mu)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\mu)\right)$$

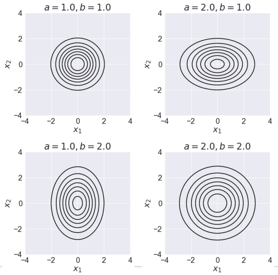
- Two parameters: $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{x}]$ and $\boldsymbol{\Sigma} = \text{cov}[\boldsymbol{x}]$

- Covariance matrix for $D = 2$

$$\boldsymbol{\Sigma} = \begin{bmatrix} a & c \\ c & b \end{bmatrix}$$

- Correlation coefficient $\rho = \frac{c}{\sqrt{ab}}$





The multivariate normal distribution

- The *multivariate normal distribution*

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\mu)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\mu)\right)$$

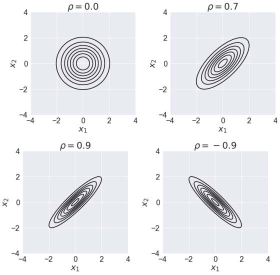
- Two parameters: $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{x}]$ and $\boldsymbol{\Sigma} = \text{cov}[\boldsymbol{x}]$

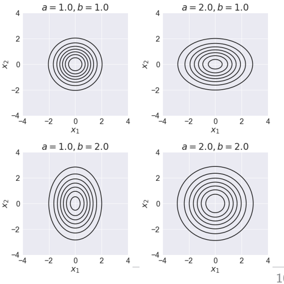
- Covariance matrix for $D = 2$

$$\boldsymbol{\Sigma} = \begin{bmatrix} a & c \\ c & b \end{bmatrix}$$

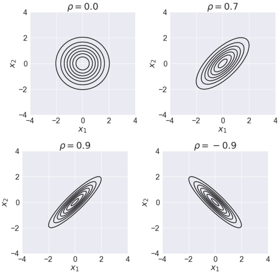
- Correlation coefficient $\rho = \frac{c}{\sqrt{ab}}$

- Let $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{m}_x, \boldsymbol{V}_x)$ and $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{m}_y, \boldsymbol{V}_y)$ for $\boldsymbol{x} \perp \boldsymbol{y}$, then





The multivariate normal distribution



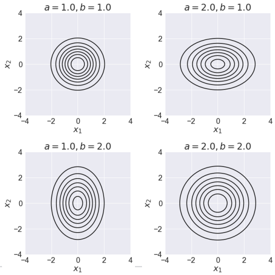
- The *multivariate normal distribution*

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\mu)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\mu)\right)$$

- Two parameters: $\boldsymbol{\mu} = \mathbb{E}\left[\boldsymbol{x}\right]$ and $\boldsymbol{\Sigma} = \text{cov}\left[\boldsymbol{x}\right]$

- Covariance matrix for $D = 2$

$$\boldsymbol{\Sigma} = \begin{bmatrix} a & c \\ c & b \end{bmatrix}$$



- Correlation coefficient $\rho = \frac{c}{\sqrt{ab}}$

- Let $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{m}_x, \boldsymbol{V}_x)$ and $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{m}_y, \boldsymbol{V}_y)$ for $\boldsymbol{x} \perp \boldsymbol{y}$, then

$$\boldsymbol{a} + \boldsymbol{B}\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{a} + \boldsymbol{B}\boldsymbol{m}_x, \boldsymbol{B}\boldsymbol{V}_x\boldsymbol{B}^T)$$

The multivariate normal distribution

- The *multivariate normal distribution*

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\mu)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\mu)\right)$$

- Two parameters: $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{x}]$ and $\boldsymbol{\Sigma} = \text{cov}[\boldsymbol{x}]$

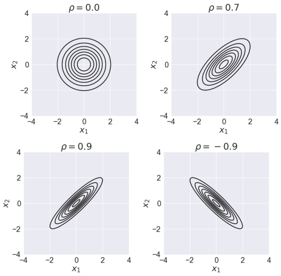
- Covariance matrix for $D = 2$

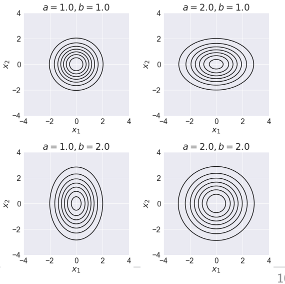
$$\boldsymbol{\Sigma} = \begin{bmatrix} a & c \\ c & b \end{bmatrix}$$

- Correlation coefficient $\rho = \frac{c}{\sqrt{ab}}$

- Let $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{m}_x, \boldsymbol{V}_x)$ and $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{m}_y, \boldsymbol{V}_y)$ for $\boldsymbol{x} \perp \boldsymbol{y}$, then

$$\boldsymbol{a} + \boldsymbol{B}\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{a} + \boldsymbol{B}\boldsymbol{m}_x, \boldsymbol{B}\boldsymbol{V}_x\boldsymbol{B}^T)$$
$$\boldsymbol{x} + \boldsymbol{y} \sim \mathcal{N}(\boldsymbol{m}_x + \boldsymbol{m}_y, \boldsymbol{V}_x + \boldsymbol{V}_y)$$





The functional form of multivariate Gaussians

- Consider now the log density, focussing only on terms dependent on $\boldsymbol{x}$.

$$\ln p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \ln \left[(2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp \left(-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right) \right]$$

The functional form of multivariate Gaussians

- Consider now the log density, focussing only on terms dependent on $\boldsymbol{x}$.

$$\begin{aligned}\ln p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \ln\left[(2\pi)^{-\frac{D}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)\right] \\ &= -\frac{D}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\end{aligned}$$

The functional form of multivariate Gaussians

- Consider now the log density, focussing only on terms dependent on $\boldsymbol{x}$.

$$
\begin{aligned}
\ln p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \ln \left[(2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \right) \right] \\
&= -\frac{D}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \\
&= -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) + \textit{constant}
\end{aligned}
$$

The functional form of multivariate Gaussians

- Consider now the log density, focussing only on terms dependent on $\boldsymbol{x}$.

$$
\begin{aligned}
\ln p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \ln \left[(2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \right) \right] \\
&= -\frac{D}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \\
&= -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) + \mathit{constant} \\
&= -\frac{1}{2}(\boldsymbol{x}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{x} - 2\boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}) + \mathit{constant}
\end{aligned}
$$

The functional form of multivariate Gaussians

- Consider now the log density, focussing only on terms dependent on $\boldsymbol{x}$.

$$
\begin{aligned}
\ln p(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \ln\left[(2\pi)^{-\frac{D}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)\right] \\
&= -\frac{D}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \\
&= -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) + \mathit{constant} \\
&= -\frac{1}{2}(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - 2\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}) + \mathit{constant} \\
&= -\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \mathit{constant}
\end{aligned}
$$

The functional form of multivariate Gaussians

- Consider now the log density, focussing only on terms dependent on $\boldsymbol{x}$.

$$
\begin{aligned}
\ln p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \ln\left[(2\pi)^{-\frac{D}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)\right] \\
&= -\frac{D}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \\
&= -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) + \textit{constant} \\
&= -\frac{1}{2}(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - 2\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}) + \textit{constant} \\
&= -\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \textit{constant}
\end{aligned}
$$

- Key take-aways

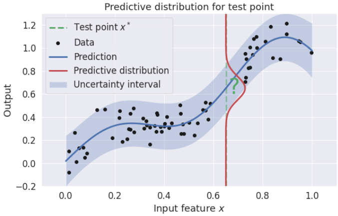
  1. Every time we encounter a distribution with a *quadratic log density*, it must be a Gaussian distribution (if $\boldsymbol{\Sigma}$ is a valid covariance matrix)

  2. We can *match coefficients* of first and second order term to *determine mean and covariance*

Recap: Linear regression and supervised learning

Supervised learning: linear regression

- Dataset $\mathcal{D} = \{(\boldsymbol{x_1}, y_1), (\boldsymbol{x_2}, y_2), \ldots, (\boldsymbol{x_N}, y_N)\}$
  - Input features: $\boldsymbol{x}_i \in \mathbb{R}^D$
  - Targets: $y_i \in \mathbb{R}$

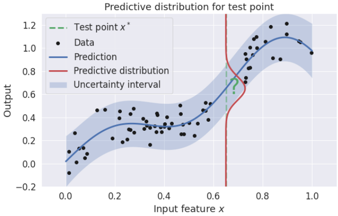


Supervised learning: linear regression

- Dataset $\mathcal{D} = \{(\boldsymbol{x_1}, y_1), (\boldsymbol{x_2}, y_2), \ldots, (\boldsymbol{x_N}, y_N)\}$
  - Input features: $\boldsymbol{x}_i \in \mathbb{R}^D$
  - Targets: $y_i \in \mathbb{R}$

- Additive noise models

$$y_i = f(\boldsymbol{x}_i | \boldsymbol{w}) + \epsilon_i$$



Supervised learning: linear regression

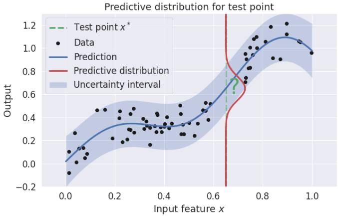
- Dataset $\mathcal{D} = \{(\boldsymbol{x_1}, y_1), (\boldsymbol{x_2}, y_2), \ldots, (\boldsymbol{x_N}, y_N)\}$
  - Input features: $\boldsymbol{x}_i \in \mathbb{R}^D$
  - Targets: $y_i \in \mathbb{R}$

- Additive noise models

$$y_i = f(\boldsymbol{x}_i | \boldsymbol{w}) + \epsilon_i$$

- Linear models are *linear wrt. parameters*, not data!

$$f(\boldsymbol{x}|\boldsymbol{w}) = w_0 + w_1 x_1 + \ldots w_M x_M = \boldsymbol{w}^T \boldsymbol{x}$$



Supervised learning: linear regression

- Dataset $\mathcal{D} = \{(\boldsymbol{x_1}, y_1), (\boldsymbol{x_2}, y_2), \ldots, (\boldsymbol{x_N}, y_N)\}$
  - Input features: $\boldsymbol{x}_i \in \mathbb{R}^D$
  - Targets: $y_i \in \mathbb{R}$

- Additive noise models

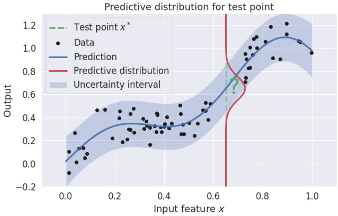
$$y_i = f(\boldsymbol{x}_i | \boldsymbol{w}) + \epsilon_i$$

- Linear models are *linear wrt. parameters*, not data!

$$f(\boldsymbol{x}|\boldsymbol{w}) = w_0 + w_1 x_1 + \ldots w_M x_M = \boldsymbol{w}^T \boldsymbol{x}$$

- *Non-linear* feature extractors $\phi(\cdot)$ (basis functions)

$$f(\boldsymbol{x}|\boldsymbol{w}) = \sum_{j=0}^{M} w_j \phi_j(\boldsymbol{x}) = \boldsymbol{w}^T \phi(\boldsymbol{x})$$



Linear regression: the probabilistic model

- The *predictive distribution* of y^* given $\boldsymbol{x}^*$ and the data $\mathcal{D}$ is our goal

$$p\left(y^*|\mathcal{D}, \boldsymbol{x}^*\right)$$

- Model for the "signal"

$$f(\boldsymbol{x}_i|\boldsymbol{w}) = \boldsymbol{w}^T\phi(\boldsymbol{x}_i)$$

- The gaussian noise ϵ_i is assumed to be *independent and identically distributed (i.i.d)*

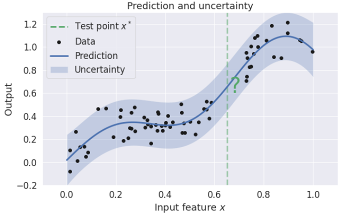
$$y_i = f(\boldsymbol{x}_i|\boldsymbol{w}) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

- The *likelihood* for the i'th data point

$$p(y_i|\boldsymbol{x}_i, \boldsymbol{w}, \sigma^2) = \mathcal{N}(y_i|\boldsymbol{w}^T\phi(\boldsymbol{x}_i), \sigma^2)$$

- Using the maximum likelihood solution as a *plug-in* estimator

$$p\left(y^*|\mathcal{D}, \boldsymbol{x}^*\right) = \mathcal{N}(y_i|\hat{\boldsymbol{w}}_{\text{MLE}}^T\phi(\boldsymbol{x}_*), \sigma^2)$$



Estimating the parameters using maximum likelihood

- Given a dataset $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \dots, (\boldsymbol{x}_N, y_N)\}$, the *likelihood* for dataset is

$$p(\boldsymbol{y}|\boldsymbol{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(y_n | f(\boldsymbol{x}_{\boldsymbol{n}}|\boldsymbol{w}), \sigma^2)$$

Estimating the parameters using maximum likelihood

- Given a dataset $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_N, y_N)\}$, the *likelihood* for dataset is

$$p(\boldsymbol{y}|\boldsymbol{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(y_n|f(\boldsymbol{x_n}|\boldsymbol{w}), \sigma^2)$$

- Taking the logarithm and using $f(\boldsymbol{x}_n, \boldsymbol{w}) = \boldsymbol{w}^T \phi(\boldsymbol{x}_n)$

$$\ln p(\boldsymbol{y}|\boldsymbol{w}, \sigma^2) = \sum_{n=1}^{N} \ln \mathcal{N}(y_n|\boldsymbol{w}^T \phi(\boldsymbol{x}_n), \sigma^2)$$

Estimating the parameters using maximum likelihood

- Given a dataset $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_N, y_N)\}$, the *likelihood* for dataset is

$$p(\boldsymbol{y}|\boldsymbol{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(y_n | f(\boldsymbol{x}_n|\boldsymbol{w}), \sigma^2)$$

- Taking the logarithm and using $f(\boldsymbol{x}_n, \boldsymbol{w}) = \boldsymbol{w}^T \phi(\boldsymbol{x}_n)$

$$\begin{aligned}
\ln p(\boldsymbol{y}|\boldsymbol{w}, \sigma^2) &= \sum_{n=1}^{N} \ln \mathcal{N}(y_n | \boldsymbol{w}^T \phi(\boldsymbol{x}_n), \sigma^2) \\
&= \sum_{n=1}^{N} \left[-\frac{1}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} (y_n - \boldsymbol{w}^T \phi(\boldsymbol{x}_n))^2 \right]
\end{aligned}$$

Estimating the parameters using maximum likelihood

- Given a dataset $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_N, y_N)\}$, the *likelihood* for dataset is

$$p(\boldsymbol{y}|\boldsymbol{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(y_n | f(\boldsymbol{x}_n|\boldsymbol{w}), \sigma^2)$$

- Taking the logarithm and using $f(\boldsymbol{x}_n, \boldsymbol{w}) = \boldsymbol{w}^T \phi(\boldsymbol{x}_n)$

$$\begin{aligned}
\ln p(\boldsymbol{y}|\boldsymbol{w}, \sigma^2) &= \sum_{n=1}^{N} \ln \mathcal{N}(y_n | \boldsymbol{w}^T \phi(\boldsymbol{x}_n), \sigma^2) \\
&= \sum_{n=1}^{N} \left[-\frac{1}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} (y_n - \boldsymbol{w}^T \phi(\boldsymbol{x}_n))^2 \right] \\
&= -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - \boldsymbol{w}^T \phi(\boldsymbol{x}_n))^2
\end{aligned}$$

Estimating the parameters using maximum likelihood

- Given a dataset $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_N, y_N)\}$, the *likelihood* for dataset is

$$p(\boldsymbol{y}|\boldsymbol{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(y_n|f(\boldsymbol{x_n}|\boldsymbol{w}), \sigma^2)$$

- Taking the logarithm and using $f(\boldsymbol{x}_n, \boldsymbol{w}) = \boldsymbol{w}^T \phi(\boldsymbol{x}_n)$

$$\begin{aligned}
\ln p(\boldsymbol{y}|\boldsymbol{w}, \sigma^2) &= \sum_{n=1}^{N} \ln \mathcal{N}(y_n|\boldsymbol{w}^T \phi(\boldsymbol{x}_n), \sigma^2) \\
&= \sum_{n=1}^{N} \left[-\frac{1}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y_n - \boldsymbol{w}^T \phi(\boldsymbol{x}_n))^2 \right] \\
&= -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - \boldsymbol{w}^T \phi(\boldsymbol{x}_n))^2
\end{aligned}$$

- Maximum likelihood estimator $\hat{\boldsymbol{w}}_{\text{MLE}}$ is equivalent to minimizing sum-of-squares error

$$\hat{\boldsymbol{w}}_{\text{MLE}} = \left(\boldsymbol{\Phi}^T \boldsymbol{\Phi}\right)^{-1} \boldsymbol{\Phi}^T \boldsymbol{y} \qquad \text{(Normal equations)}$$

Estimating the parameters using maximum likelihood

- Given a dataset $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_N, y_N)\}$, the *likelihood* for dataset is

$$p(\boldsymbol{y}|\boldsymbol{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(y_n | f(\boldsymbol{x_n}|\boldsymbol{w}), \sigma^2)$$

- Taking the logarithm and using $f(\boldsymbol{x}_n, \boldsymbol{w}) = \boldsymbol{w}^T \phi(\boldsymbol{x}_n)$

$$\begin{aligned}
\ln p(\boldsymbol{y}|\boldsymbol{w}, \sigma^2) &= \sum_{n=1}^{N} \ln \mathcal{N}(y_n | \boldsymbol{w}^T \phi(\boldsymbol{x}_n), \sigma^2) \\
&= \sum_{n=1}^{N} \left[-\frac{1}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} (y_n - \boldsymbol{w}^T \phi(\boldsymbol{x}_n))^2 \right] \\
&= -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - \boldsymbol{w}^T \phi(\boldsymbol{x}_n))^2
\end{aligned}$$

- Maximum likelihood estimator $\hat{\boldsymbol{w}}_{\text{MLE}}$ is equivalent to minimizing sum-of-squares error

$$\hat{\boldsymbol{w}}_{\text{MLE}} = \left(\boldsymbol{\Phi}^T \boldsymbol{\Phi}\right)^{-1} \boldsymbol{\Phi}^T \boldsymbol{y} \qquad \text{(Normal equations)}$$

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{N} \sum_{n=1}^{N} (y_n - \hat{\boldsymbol{w}}_{\text{MLE}}^T \phi(\boldsymbol{x_n}))^2$$

Example: Polynomial regression using maximum likelihood I

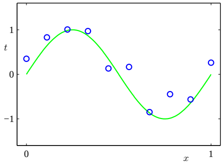
Example from Bishop

- Polynomial basis functions

$$f(x|\boldsymbol{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \boldsymbol{w}^T \phi(x)$$

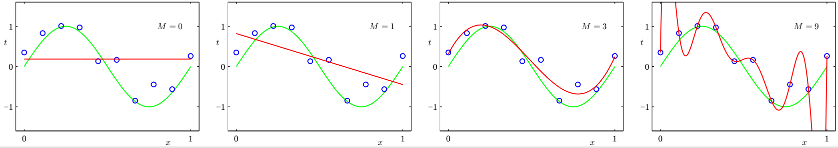
- Feature transformations

$$\phi(x) = \begin{bmatrix} 1 & x & x^2 & \ldots & x^M \end{bmatrix}^T$$



- M controls the *model complexity*: Underfitting vs overfitting

- *Model selection*: How to choose M?

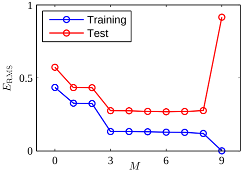


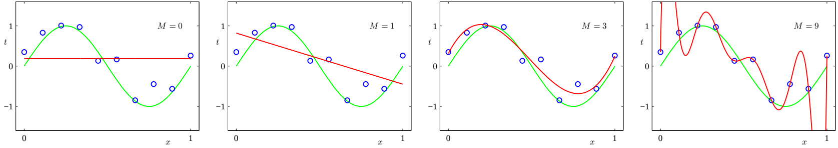
Example: Polynomial regression using maximum likelihood II

- *Cross-validation*
  Split data into training and test sets

- *Overfitting (low training error, high test error)*
  As the function become more flexible we start to fit the noise in the data

- *"Underfitting" (high training error, high test error)*
  When the function is not sufficiently flexible to fit the data





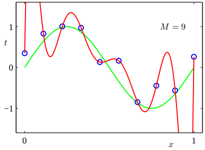
Example: Polynomial regression using maximum likelihood III

- *The optimal model complexity depends on the amount of data*
  The more data the more flexible model we can "afford" to fit

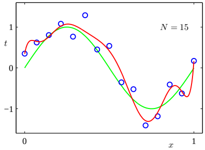
- *Regularization: Controlling the effective model complexity*
  We can use regularization to control the effect model complexity when we have limited data

	$M=0$	$M=1$	$M=6$	$M=9$
$w_0^\star$	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

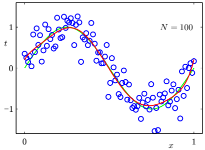
$N = 10$, $M = 9$



$N = 15$, $M = 9$



$N = 100$, $M = 9$



Regularized least squares

- Recall: *Maximum likelihood* is equivalent *minimizng sum-of-squares error*

$$E_D(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}(y_n - f(x_n|\boldsymbol{w}))^2$$

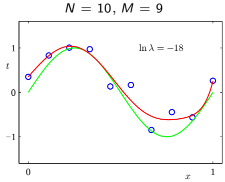
- Adding *penalty term* to prevent weights from becoming too large

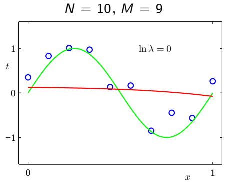
$$\tilde{E}_D(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}(y_n - f(x_n|\boldsymbol{w}))^2 + \frac{\lambda}{2}||\boldsymbol{w}||^2$$

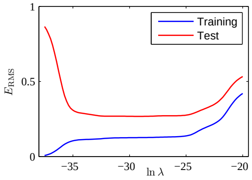
- What happens when $\lambda = 0$? $\lambda \to \infty$?

- Many names: Ridge regression, shrinkage, weight decay

- Regularization parameter λ controls the *effective complexity*



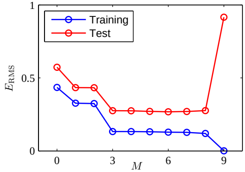


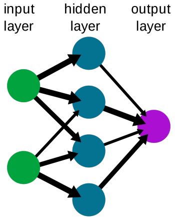


Bayesian linear regression

Bayesian Linear regression: motivation

- *Overfitting*
  Maximum likelihood can be problematic for flexible models
- *Controlling model complexity*
  Limitting number of basis functions and/or regularization?
- *Model selection*
  How to choose the optimal value of λ?
- *Cross-validation*
  Training + validation/development + test
- *Bayesian methods*
  - less prone to overfitting
  - can (often) adapt model complexity automatically
- *Applications in modern machine learning*
  1. Small datasets
  2. Transfer learning
  3. Component in more complex models
  4. Simple uncertainty quantification for neural networks





Bayesian Linear regression: prior and likelihood

- Simplified set-up: assuming σ^2 is fixed and known, then Bayes' rule states

$$p(\boldsymbol{w}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{w})p(\boldsymbol{w})}{p(\boldsymbol{y})}$$

Bayesian Linear regression: prior and likelihood

- Simplified set-up: assuming σ^2 is fixed and known, then Bayes' rule states

$$p(\boldsymbol{w}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{w})p(\boldsymbol{w})}{p(\boldsymbol{y})}$$

- We already know the *likelihood*

$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(y_n|\boldsymbol{w}^T\phi(\boldsymbol{x}_n), \sigma^2) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{\Phi}\boldsymbol{w}, \sigma^2\boldsymbol{I})$$

Bayesian Linear regression: prior and likelihood

- Simplified set-up: assuming σ^2 is fixed and known, then Bayes' rule states

$$p(\boldsymbol{w}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{w})p(\boldsymbol{w})}{p(\boldsymbol{y})}$$

- We already know the *likelihood*

$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(y_n|\boldsymbol{w}^T\phi(\boldsymbol{x}_n), \sigma^2) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{\Phi}\boldsymbol{w}, \sigma^2\boldsymbol{I})$$

- *The marginal likelihood* is the denominator in Bayes's theorem and is independent of $\boldsymbol{w}$

$$p(\boldsymbol{y}) = \int p(\boldsymbol{y}|\boldsymbol{w})p(\boldsymbol{w})\mathrm{d}\boldsymbol{w} = \mathbb{E}_{p(\boldsymbol{w})}\left[p(\boldsymbol{y}|\boldsymbol{w})\right]$$

Bayesian Linear regression: prior and likelihood

- Simplified set-up: assuming σ^2 is fixed and known, then Bayes' rule states

$$p(\boldsymbol{w}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{w})p(\boldsymbol{w})}{p(\boldsymbol{y})}$$

- We already know the *likelihood*

$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(y_n|\boldsymbol{w}^T\phi(\boldsymbol{x}_n), \sigma^2) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{\Phi}\boldsymbol{w}, \sigma^2\boldsymbol{I})$$

- *The marginal likelihood* is the denominator in Bayes's theorem and is independent of $\boldsymbol{w}$

$$p(\boldsymbol{y}) = \int p(\boldsymbol{y}|\boldsymbol{w})p(\boldsymbol{w})\mathrm{d}\boldsymbol{w} = \mathbb{E}_{p(\boldsymbol{w})}\left[p(\boldsymbol{y}|\boldsymbol{w})\right]$$

- The multivariate normal distribution is a *conjugate prior* for the $\boldsymbol{w}$

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

Bayesian Linear regression: prior and likelihood

- Simplified set-up: assuming σ^2 is fixed and known, then Bayes' rule states

$$p(\boldsymbol{w}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{w})p(\boldsymbol{w})}{p(\boldsymbol{y})}$$

- We already know the *likelihood*

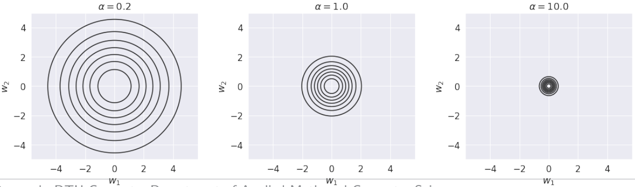
$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(y_n|\boldsymbol{w}^T\phi(\boldsymbol{x}_n), \sigma^2) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{\Phi}\boldsymbol{w}, \sigma^2\boldsymbol{I})$$

- *The marginal likelihood* is the denominator in Bayes's theorem and is independent of $\boldsymbol{w}$

$$p(\boldsymbol{y}) = \int p(\boldsymbol{y}|\boldsymbol{w})p(\boldsymbol{w})\mathrm{d}\boldsymbol{w} = \mathbb{E}_{p(\boldsymbol{w})}\left[p(\boldsymbol{y}|\boldsymbol{w})\right]$$

- The multivariate normal distribution is a *conjugate prior* for the $\boldsymbol{w}$

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



Bayesian Linear regression: The MAP estimator

- The *posterior distribution* of the weights $\boldsymbol{w}$ given the data $\boldsymbol{y}$ is given by

$$p(\boldsymbol{w}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{w})p(\boldsymbol{w})}{p(\boldsymbol{y})} = \frac{\mathcal{N}(\boldsymbol{y}|\boldsymbol{\Phi}\boldsymbol{w}, \sigma^2\boldsymbol{I})\mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1}\boldsymbol{I})}{p(\boldsymbol{y})}$$

Bayesian Linear regression: The MAP estimator

- The *posterior distribution* of the weights $\boldsymbol{w}$ given the data $\boldsymbol{y}$ is given by

$$p(\boldsymbol{w}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{w})p(\boldsymbol{w})}{p(\boldsymbol{y})} = \frac{\mathcal{N}(\boldsymbol{y}|\boldsymbol{\Phi w}, \sigma^2 \boldsymbol{I})\mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1}\boldsymbol{I})}{p(\boldsymbol{y})}$$

- Let's look at the *maximum a posteriori* (MAP) estimate: $\hat{\boldsymbol{w}}_{\text{MAP}} = \arg\max_{\boldsymbol{w}} p(\boldsymbol{w}|\boldsymbol{y})$

$$p(\boldsymbol{w}|\boldsymbol{y}) \propto \mathcal{N}(\boldsymbol{y}|\boldsymbol{\Phi w}, \sigma^2 \boldsymbol{I})\mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1}\boldsymbol{I})$$

Bayesian Linear regression: The MAP estimator

- The *posterior distribution* of the weights $\boldsymbol{w}$ given the data $\boldsymbol{y}$ is given by

$$p(\boldsymbol{w}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{w})p(\boldsymbol{w})}{p(\boldsymbol{y})} = \frac{\mathcal{N}(\boldsymbol{y}|\boldsymbol{\Phi}\boldsymbol{w}, \sigma^2\boldsymbol{I})\mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1}\boldsymbol{I})}{p(\boldsymbol{y})}$$

- Let's look at the *maximum a posteriori* (MAP) estimate: $\hat{\boldsymbol{w}}_{\text{MAP}} = \arg\max_{\boldsymbol{w}} p(\boldsymbol{w}|\boldsymbol{y})$

$$p(\boldsymbol{w}|\boldsymbol{y}) \propto \mathcal{N}(\boldsymbol{y}|\boldsymbol{\Phi}\boldsymbol{w}, \sigma^2\boldsymbol{I})\mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1}\boldsymbol{I})$$

- Taking the logarithm

$$\ln p(\boldsymbol{w}|\boldsymbol{y}) \propto \ln \mathcal{N}(\boldsymbol{y}|\boldsymbol{\Phi}\boldsymbol{w}, \sigma^2\boldsymbol{I}) + \ln \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1}\boldsymbol{I})$$

Bayesian Linear regression: The MAP estimator

- The *posterior distribution* of the weights $\boldsymbol{w}$ given the data $\boldsymbol{y}$ is given by

$$p(\boldsymbol{w}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{w})p(\boldsymbol{w})}{p(\boldsymbol{y})} = \frac{\mathcal{N}(\boldsymbol{y}|\boldsymbol{\Phi}\boldsymbol{w}, \sigma^2\boldsymbol{I})\mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1}\boldsymbol{I})}{p(\boldsymbol{y})}$$

- Let's look at the *maximum a posteriori* (MAP) estimate: $\hat{\boldsymbol{w}}_{\text{MAP}} = \arg\max_{\boldsymbol{w}} p(\boldsymbol{w}|\boldsymbol{y})$

$$p(\boldsymbol{w}|\boldsymbol{y}) \propto \mathcal{N}(\boldsymbol{y}|\boldsymbol{\Phi}\boldsymbol{w}, \sigma^2\boldsymbol{I})\mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1}\boldsymbol{I})$$

- Taking the logarithm

$$\begin{aligned}\ln p(\boldsymbol{w}|\boldsymbol{y}) &\propto \ln \mathcal{N}(\boldsymbol{y}|\boldsymbol{\Phi}\boldsymbol{w}, \sigma^2\boldsymbol{I}) + \ln \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1}\boldsymbol{I}) \\ &= -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{\beta}{2}\sum_{n=1}^{N}(y_n - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n))^2 - \frac{D}{2}\ln(2\pi\alpha^{-1}) - \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w}\end{aligned}$$

Bayesian Linear regression: The MAP estimator

- The *posterior distribution* of the weights $\boldsymbol{w}$ given the data $\boldsymbol{y}$ is given by

$$p(\boldsymbol{w}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{w})p(\boldsymbol{w})}{p(\boldsymbol{y})} = \frac{\mathcal{N}(\boldsymbol{y}|\boldsymbol{\Phi}\boldsymbol{w}, \sigma^2\boldsymbol{I})\mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1}\boldsymbol{I})}{p(\boldsymbol{y})}$$

- Let's look at the *maximum a posteriori* (MAP) estimate: $\hat{\boldsymbol{w}}_{\text{MAP}} = \arg\max_{\boldsymbol{w}} p(\boldsymbol{w}|\boldsymbol{y})$

$$p(\boldsymbol{w}|\boldsymbol{y}) \propto \mathcal{N}(\boldsymbol{y}|\boldsymbol{\Phi}\boldsymbol{w}, \sigma^2\boldsymbol{I})\mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1}\boldsymbol{I})$$

- Taking the logarithm

$$\begin{aligned}
\ln p(\boldsymbol{w}|\boldsymbol{y}) &\propto \ln \mathcal{N}(\boldsymbol{y}|\boldsymbol{\Phi}\boldsymbol{w}, \sigma^2\boldsymbol{I}) + \ln \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1}\boldsymbol{I}) \\
&= -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{\beta}{2}\sum_{n=1}^{N}(y_n - \boldsymbol{w}^T\phi(\boldsymbol{x}_n))^2 - \frac{D}{2}\ln(2\pi\alpha^{-1}) - \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} \\
&= -\frac{\beta}{2}\sum_{n=1}^{N}(y_n - \boldsymbol{w}^T\phi(\boldsymbol{x}_n))^2 - \frac{\alpha}{2}\sum_{i=1}^{D} w_i^2 + \text{constant}
\end{aligned}$$

Bayesian Linear regression: The MAP estimator

- The *posterior distribution* of the weights $\boldsymbol{w}$ given the data $\boldsymbol{y}$ is given by

$$p(\boldsymbol{w}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{w})p(\boldsymbol{w})}{p(\boldsymbol{y})} = \frac{\mathcal{N}(\boldsymbol{y}|\boldsymbol{\Phi}\boldsymbol{w}, \sigma^2\boldsymbol{I})\mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1}\boldsymbol{I})}{p(\boldsymbol{y})}$$

- Let's look at the *maximum a posteriori* (MAP) estimate: $\hat{\boldsymbol{w}}_{\text{MAP}} = \arg\max_{\boldsymbol{w}} p(\boldsymbol{w}|\boldsymbol{y})$

$$p(\boldsymbol{w}|\boldsymbol{y}) \propto \mathcal{N}(\boldsymbol{y}|\boldsymbol{\Phi}\boldsymbol{w}, \sigma^2\boldsymbol{I})\mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1}\boldsymbol{I})$$

- Taking the logarithm

$$\begin{aligned}
\ln p(\boldsymbol{w}|\boldsymbol{y}) &\propto \ln \mathcal{N}(\boldsymbol{y}|\boldsymbol{\Phi}\boldsymbol{w}, \sigma^2\boldsymbol{I}) + \ln \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1}\boldsymbol{I}) \\
&= -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{\beta}{2}\sum_{n=1}^{N}(y_n - \boldsymbol{w}^T\phi(\boldsymbol{x}_n))^2 - \frac{D}{2}\ln(2\pi\alpha^{-1}) - \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} \\
&= -\frac{\beta}{2}\sum_{n=1}^{N}(y_n - \boldsymbol{w}^T\phi(\boldsymbol{x}_n))^2 - \frac{\alpha}{2}\sum_{i=1}^{D} w_i^2 + \text{constant}
\end{aligned}$$

- The *mode of the posterior (MAP)* is equivalent to ridge regression with $\lambda = \frac{\alpha}{\beta}$ and to maximum likelihood when $\alpha \to 0$

Deriving the posterior distribution of the weights

Recall the functional form of a generic multivariate Gaussian $\mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{S})$

$$\ln \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{S}) = -\frac{1}{2}\boldsymbol{w}^T\boldsymbol{S}^{-1}\boldsymbol{w} + \boldsymbol{m}^T\boldsymbol{S}^{-1}\boldsymbol{w} + \textit{constant}$$

Deriving the posterior distribution of the weights

Recall the functional form of a generic multivariate Gaussian $\mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{S})$

$$\ln \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{S}) = -\frac{1}{2}\boldsymbol{w}^T\boldsymbol{S}^{-1}\boldsymbol{w} + \boldsymbol{m}^T\boldsymbol{S}^{-1}\boldsymbol{w} + \textit{constant}$$

- We focus on term that depends on $\boldsymbol{w}$. From the previous slide, we have

$$\ln p(\boldsymbol{w}|\boldsymbol{y}) = -\frac{\beta}{2}\sum_{n=1}^{N}(y_n - \boldsymbol{w}^T\phi(\boldsymbol{x}_n))^2 - \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} + \text{constant}$$

Deriving the posterior distribution of the weights

Recall the functional form of a generic multivariate Gaussian $\mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{S})$

$$\ln \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{S}) = -\frac{1}{2}\boldsymbol{w}^T\boldsymbol{S}^{-1}\boldsymbol{w} + \boldsymbol{m}^T\boldsymbol{S}^{-1}\boldsymbol{w} + \textit{constant}$$

- We focus on term that depends on $\boldsymbol{w}$. From the previous slide, we have

$$\begin{aligned}\ln p(\boldsymbol{w}|\boldsymbol{y}) &= -\frac{\beta}{2}\sum_{n=1}^{N}(y_n - \boldsymbol{w}^T\phi(\boldsymbol{x}_n))^2 - \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} + \text{constant} \\ &= -\frac{\beta}{2}(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{w}) - \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} + \text{constant}\end{aligned}$$

Deriving the posterior distribution of the weights

Recall the functional form of a generic multivariate Gaussian $\mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{S})$

$$\ln \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{S}) = -\frac{1}{2}\boldsymbol{w}^T\boldsymbol{S}^{-1}\boldsymbol{w} + \boldsymbol{m}^T\boldsymbol{S}^{-1}\boldsymbol{w} + \textit{constant}$$

- We focus on term that depends on $\boldsymbol{w}$. From the previous slide, we have

$$\begin{aligned}
\ln p(\boldsymbol{w}|\boldsymbol{y}) &= -\frac{\beta}{2}\sum_{n=1}^{N}(y_n - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n))^2 - \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} + \text{constant} \\
&= -\frac{\beta}{2}(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{w}) - \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} + \text{constant} \\
&= -\frac{\beta}{2}\left(\boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\boldsymbol{\Phi}\boldsymbol{w} + \boldsymbol{w}\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{w}\right) - \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} + \text{constant}
\end{aligned}$$

Deriving the posterior distribution of the weights

Recall the functional form of a generic multivariate Gaussian $\mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{S})$

$$\ln \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{S}) = -\frac{1}{2}\boldsymbol{w}^T\boldsymbol{S}^{-1}\boldsymbol{w} + \boldsymbol{m}^T\boldsymbol{S}^{-1}\boldsymbol{w} + \textit{constant}$$

- We focus on term that depends on $\boldsymbol{w}$. From the previous slide, we have

$$\begin{aligned}
\ln p(\boldsymbol{w}|\boldsymbol{y}) &= -\frac{\beta}{2}\sum_{n=1}^{N}(y_n - \boldsymbol{w}^T\phi(\boldsymbol{x}_n))^2 - \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} + \text{constant} \\
&= -\frac{\beta}{2}(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{w}) - \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} + \text{constant} \\
&= -\frac{\beta}{2}\left(\boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\boldsymbol{\Phi}\boldsymbol{w} + \boldsymbol{w}\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{w}\right) - \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} + \text{constant} \\
&= -\frac{\beta}{2}\left(-2\boldsymbol{y}^T\boldsymbol{\Phi}\boldsymbol{w} + \boldsymbol{w}\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{w}\right) - \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} + \text{constant}
\end{aligned}$$

Deriving the posterior distribution of the weights

Recall the functional form of a generic multivariate Gaussian $\mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{S})$

$$\ln \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{S}) = -\frac{1}{2}\boldsymbol{w}^T\boldsymbol{S}^{-1}\boldsymbol{w} + \boldsymbol{m}^T\boldsymbol{S}^{-1}\boldsymbol{w} + \textit{constant}$$

- We focus on term that depends on $\boldsymbol{w}$. From the previous slide, we have

$$\begin{aligned}
\ln p(\boldsymbol{w}|\boldsymbol{y}) &= -\frac{\beta}{2}\sum_{n=1}^{N}(y_n - \boldsymbol{w}^T\phi(\boldsymbol{x}_n))^2 - \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} + \text{constant} \\
&= -\frac{\beta}{2}(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{w}) - \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} + \text{constant} \\
&= -\frac{\beta}{2}\left(\boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\boldsymbol{\Phi}\boldsymbol{w} + \boldsymbol{w}\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{w}\right) - \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} + \text{constant} \\
&= -\frac{\beta}{2}\left(-2\boldsymbol{y}^T\boldsymbol{\Phi}\boldsymbol{w} + \boldsymbol{w}\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{w}\right) - \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} + \text{constant} \\
&= -\frac{1}{2}\boldsymbol{w}^T\left(\beta\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \alpha\boldsymbol{I}\right)\boldsymbol{w} + \beta\boldsymbol{y}^T\boldsymbol{\Phi}\boldsymbol{w} + \text{constant}
\end{aligned}$$

Deriving the posterior distribution of the weights

Recall the functional form of a generic multivariate Gaussian $\mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{S})$

$$\ln \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{S}) = -\frac{1}{2}\boldsymbol{w}^T\boldsymbol{S}^{-1}\boldsymbol{w} + \boldsymbol{m}^T\boldsymbol{S}^{-1}\boldsymbol{w} + \textit{constant}$$

- We focus on term that depends on $\boldsymbol{w}$. From the previous slide, we have

$$\begin{aligned}
\ln p(\boldsymbol{w}|\boldsymbol{y}) &= -\frac{\beta}{2}\sum_{n=1}^{N}(y_n - \boldsymbol{w}^T\phi(\boldsymbol{x}_n))^2 - \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} + \text{constant} \\
&= -\frac{\beta}{2}(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{w}) - \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} + \text{constant} \\
&= -\frac{\beta}{2}\left(\boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\boldsymbol{\Phi}\boldsymbol{w} + \boldsymbol{w}\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{w}\right) - \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} + \text{constant} \\
&= -\frac{\beta}{2}\left(-2\boldsymbol{y}^T\boldsymbol{\Phi}\boldsymbol{w} + \boldsymbol{w}\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{w}\right) - \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} + \text{constant} \\
&= -\frac{1}{2}\boldsymbol{w}^T\left(\beta\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \alpha\boldsymbol{I}\right)\boldsymbol{w} + \beta\boldsymbol{y}^T\boldsymbol{\Phi}\boldsymbol{w} + \text{constant}
\end{aligned}$$

- Equating coefficients for the second order term

$$\boldsymbol{S}^{-1} = \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \alpha\boldsymbol{I} \quad \iff \quad \boldsymbol{S} = \left(\beta\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \alpha\boldsymbol{I}\right)^{-1}$$

Deriving the posterior distribution of the weights

Recall the functional form of a generic multivariate Gaussian $\mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{S})$

$$\ln \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{S}) = -\frac{1}{2}\boldsymbol{w}^T\boldsymbol{S}^{-1}\boldsymbol{w} + \boldsymbol{m}^T\boldsymbol{S}^{-1}\boldsymbol{w} + \textit{constant}$$

- We focus on term that depends on $\boldsymbol{w}$. From the previous slide, we have

$$\begin{aligned}
\ln p(\boldsymbol{w}|\boldsymbol{y}) &= -\frac{\beta}{2}\sum_{n=1}^{N}(y_n - \boldsymbol{w}^T\phi(\boldsymbol{x}_n))^2 - \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} + \text{constant} \\
&= -\frac{\beta}{2}(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{w}) - \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} + \text{constant} \\
&= -\frac{\beta}{2}\left(\boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\boldsymbol{\Phi}\boldsymbol{w} + \boldsymbol{w}\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{w}\right) - \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} + \text{constant} \\
&= -\frac{\beta}{2}\left(-2\boldsymbol{y}^T\boldsymbol{\Phi}\boldsymbol{w} + \boldsymbol{w}\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{w}\right) - \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} + \text{constant} \\
&= -\frac{1}{2}\boldsymbol{w}^T\left(\beta\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \alpha\boldsymbol{I}\right)\boldsymbol{w} + \beta\boldsymbol{y}^T\boldsymbol{\Phi}\boldsymbol{w} + \text{constant}
\end{aligned}$$

- Equating coefficients for the second order term

$$\boldsymbol{S}^{-1} = \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \alpha\boldsymbol{I} \quad \iff \quad \boldsymbol{S} = \left(\beta\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \alpha\boldsymbol{I}\right)^{-1}$$

- Equating coefficients for the first order term

$$\boldsymbol{m}^T\boldsymbol{S}^{-1} = \beta\boldsymbol{y}^T\boldsymbol{\Phi} \quad \iff \quad \boldsymbol{m} = \beta\boldsymbol{S}\boldsymbol{\Phi}^T\boldsymbol{y}$$

Bayesian linear regression model: the key equations

- Given design matrix $\mathbf{\Phi} \in \mathbb{R}^{N \times D}$ and observations $\boldsymbol{y} \in \mathbb{R}^N$:

$$
\begin{aligned}
p(\boldsymbol{w}) &= \mathcal{N}(\boldsymbol{w}|0, \alpha^{-1}\boldsymbol{I}) && \text{(}\textit{prior}\text{)}\\
p(\boldsymbol{y}|\boldsymbol{w}) &= \mathcal{N}(\boldsymbol{y}|\mathbf{\Phi}\boldsymbol{w}, \sigma^2\boldsymbol{I}) && \text{(}\textit{likelihood}\text{)}\\
p(\boldsymbol{w}|\boldsymbol{y}) &= \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{S}) && \text{(}\textit{posterior}\text{)}\\
p(\boldsymbol{y}) &= \mathcal{N}(\boldsymbol{y}|\mathbf{0}, \sigma^2\boldsymbol{I} + \alpha^{-1}\mathbf{\Phi}\mathbf{\Phi}^T) && \text{(}\textit{marginal likelihood}\text{)}
\end{aligned}
$$

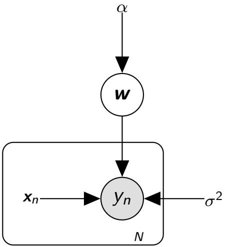
- The *posterior parameters* are given by (using $\beta \equiv \frac{1}{\sigma^2}$)

$$
\boldsymbol{m} = \beta \boldsymbol{S}\mathbf{\Phi}^T\boldsymbol{y}
$$

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

- *Two hyperparameters*

  α: prior precision of the regression weights
  β: precision of the measurements



Linear Gaussian-systems in general (see Section 3.3 in Murphy1)

- For *linear* systems: the Gaussian distribution is *conjugate* to itself
- The *posterior* for a *linear* Gaussian model with Gaussian prior is also *Gaussian*

$$p(\boldsymbol{y}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{W}\boldsymbol{z} + \boldsymbol{b}, \boldsymbol{\Sigma}_y) \qquad\qquad p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$$

Linear Gaussian-systems in general (see Section 3.3 in Murphy1)

- For *linear* systems: the Gaussian distribution is *conjugate* to itself
- The *posterior* for a *linear* Gaussian model with Gaussian prior is also *Gaussian*

$$p(\boldsymbol{y}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{W}\boldsymbol{z} + \boldsymbol{b}, \boldsymbol{\Sigma}_y) \qquad\qquad p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$$

- The *joint* distribution $p(\boldsymbol{z}, \boldsymbol{y}) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{z}\\ \boldsymbol{y}\end{bmatrix} |\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$

$$\boldsymbol{\mu} = \begin{bmatrix}\boldsymbol{\mu}_z \\ \boldsymbol{W}\boldsymbol{\mu}_z + \boldsymbol{b}\end{bmatrix} \qquad\qquad \boldsymbol{\Sigma} = \begin{bmatrix}\boldsymbol{\Sigma}_z & \boldsymbol{\Sigma}_z\boldsymbol{W}^T \\ \boldsymbol{W}\boldsymbol{\Sigma}_Z & \boldsymbol{\Sigma}_y + \boldsymbol{W}\boldsymbol{\Sigma}_z\boldsymbol{W}^T\end{bmatrix}$$

Linear Gaussian-systems in general (see Section 3.3 in Murphy1)

- For *linear* systems: the Gaussian distribution is *conjugate* to itself
- The *posterior* for a *linear* Gaussian model with Gaussian prior is also *Gaussian*

$$p(\boldsymbol{y}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{W}\boldsymbol{z} + \boldsymbol{b}, \boldsymbol{\Sigma}_y) \qquad\qquad p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$$

- The *joint* distribution $p(\boldsymbol{z}, \boldsymbol{y}) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{z}\\ \boldsymbol{y}\end{bmatrix} | \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$

$$\boldsymbol{\mu} = \begin{bmatrix}\boldsymbol{\mu}_z \\ \boldsymbol{W}\boldsymbol{\mu}_z + \boldsymbol{b}\end{bmatrix} \qquad\qquad \boldsymbol{\Sigma} = \begin{bmatrix}\boldsymbol{\Sigma}_z & \boldsymbol{\Sigma}_z\boldsymbol{W}^T \\ \boldsymbol{W}\boldsymbol{\Sigma}_Z & \boldsymbol{\Sigma}_y + \boldsymbol{W}\boldsymbol{\Sigma}_z\boldsymbol{W}^T\end{bmatrix}$$

- The *posterior* distribution of $\boldsymbol{z}$ given $\boldsymbol{y}$

$$p(\boldsymbol{y}|\boldsymbol{z}) = \mathcal{N}\left(\boldsymbol{z}|\boldsymbol{\mu}_{z|y}, \boldsymbol{\Sigma}_{z|y}\right)$$

$$\boldsymbol{\Sigma}_{z|y}^{-1} = \boldsymbol{\Sigma}_z^{-1} + \boldsymbol{W}^T\boldsymbol{\Sigma}_y\boldsymbol{W}$$

$$\boldsymbol{\mu}_{z|y} = \boldsymbol{\Sigma}_{z|y}\left[\boldsymbol{W}^T\boldsymbol{\Sigma}_y^{-1}(\boldsymbol{y} - \boldsymbol{b}) + \boldsymbol{\Sigma}_z^{-1}\boldsymbol{\mu}_z\right]$$

Linear Gaussian-systems in general (see Section 3.3 in Murphy1)

- For *linear* systems: the Gaussian distribution is *conjugate* to itself
- The *posterior* for a *linear* Gaussian model with Gaussian prior is also *Gaussian*

$$p(\boldsymbol{y}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{W}\boldsymbol{z} + \boldsymbol{b}, \boldsymbol{\Sigma}_y) \qquad\qquad p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$$

- The *joint* distribution $p(\boldsymbol{z}, \boldsymbol{y}) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{z}\\ \boldsymbol{y}\end{bmatrix} | \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$

$$\boldsymbol{\mu} = \begin{bmatrix}\boldsymbol{\mu}_z \\ \boldsymbol{W}\boldsymbol{\mu}_z + \boldsymbol{b}\end{bmatrix} \qquad\qquad \boldsymbol{\Sigma} = \begin{bmatrix}\boldsymbol{\Sigma}_z & \boldsymbol{\Sigma}_z \boldsymbol{W}^T \\ \boldsymbol{W}\boldsymbol{\Sigma}_Z & \boldsymbol{\Sigma}_y + \boldsymbol{W}\boldsymbol{\Sigma}_z\boldsymbol{W}^T\end{bmatrix}$$

- The *posterior* distribution of $\boldsymbol{z}$ given $\boldsymbol{y}$

$$p(\boldsymbol{y}|\boldsymbol{z}) = \mathcal{N}\left(\boldsymbol{z}|\boldsymbol{\mu}_{z|y}, \boldsymbol{\Sigma}_{z|y}\right)$$
$$\boldsymbol{\Sigma}_{z|y}^{-1} = \boldsymbol{\Sigma}_z^{-1} + \boldsymbol{W}^T\boldsymbol{\Sigma}_y\boldsymbol{W}$$
$$\boldsymbol{\mu}_{z|y} = \boldsymbol{\Sigma}_{z|y}\left[\boldsymbol{W}^T\boldsymbol{\Sigma}_y^{-1}(\boldsymbol{y} - \boldsymbol{b}) + \boldsymbol{\Sigma}_z^{-1}\boldsymbol{\mu}_z\right]$$

- The *marginal* distribution $\boldsymbol{y}$

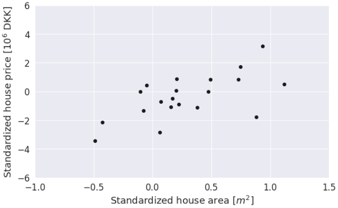
$$p(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{W}\boldsymbol{\mu}_z + \boldsymbol{b}, \boldsymbol{\Sigma}_y + \boldsymbol{W}\boldsymbol{\Sigma}_z\boldsymbol{W}^T)$$

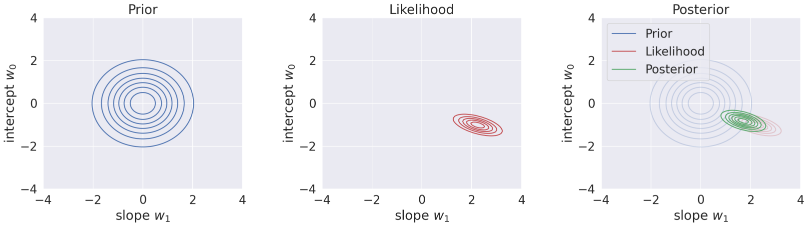
Example

- Simple linear model for fictive house prices

$$f(x|\boldsymbol{w}) = w_0 + w_1 x$$

- The posterior summarizes our beliefs about the parameters after seeing the data





$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|0, \alpha^{-1}\boldsymbol{I}) \qquad p(\boldsymbol{y}|\boldsymbol{w}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{\Phi}\boldsymbol{w}, \sigma^2\boldsymbol{I}) \qquad p(\boldsymbol{w}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{S})$$

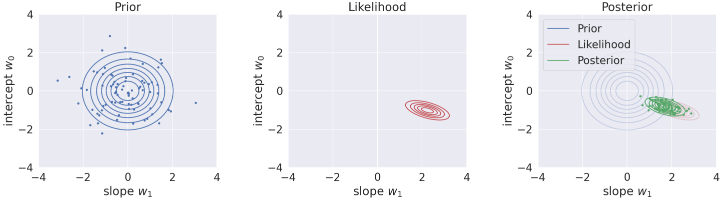
Posterior inference

$$\text{house price} = w_0 + w_1 \cdot \text{area}$$

- Question: Are larger areas associated with bigger house prices?
- We can calculate various probabilities of interest directly from the posterior (analytically or via sampling), e.g.

$$p(w_1 > 0|y) \approx 0.99$$

- No need to remember whether to should use t-tests, F-tests, χ^2-tests etc



The posterior predictive distribution

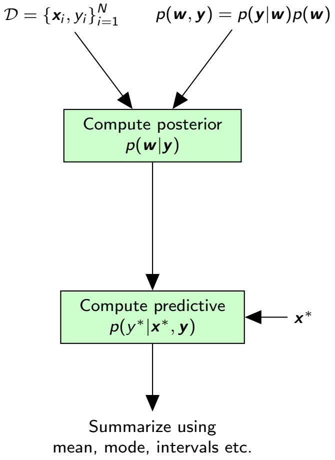
But how about making predictions?

- We defined a *linear* regression model

$$y_i = \boldsymbol{w}^T \phi(\boldsymbol{x}_i) + \epsilon_i$$

- We derived the *posterior distribution*

$$p(\boldsymbol{w}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{S})$$



But how about making predictions?

- We defined a *linear* regression model

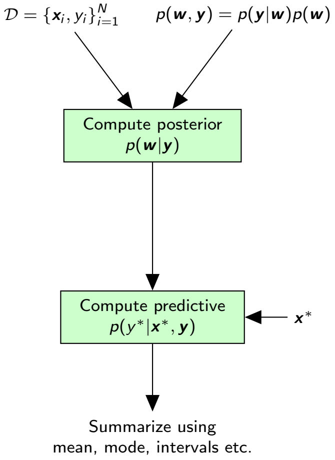
$$y_i = \boldsymbol{w}^T \phi(\boldsymbol{x}_i) + \epsilon_i$$

- We derived the *posterior distribution*

$$p(\boldsymbol{w}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{S})$$

- Next goal: *making predictions* for new $\boldsymbol{x}^*$

$$p(y^*|\boldsymbol{x}^*, \boldsymbol{w}) = \mathcal{N}(y^*|\boldsymbol{w}^T \underbrace{\phi(\boldsymbol{x}^*)}_{\boldsymbol{\phi}_*}, \sigma^2)$$



But how about making predictions?

- We defined a *linear* regression model

$$y_i = \boldsymbol{w}^T \phi(\boldsymbol{x}_i) + \epsilon_i$$

- We derived the *posterior distribution*

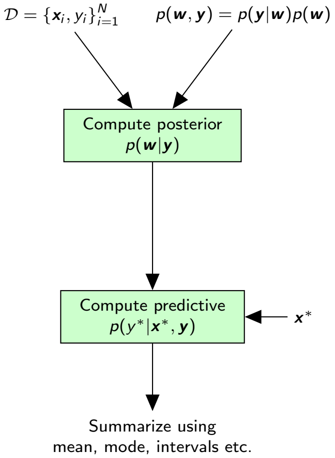
$$p(\boldsymbol{w}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{S})$$

- Next goal: *making predictions* for new $\boldsymbol{x}^*$

$$p(y^*|\boldsymbol{x}^*, \boldsymbol{w}) = \mathcal{N}(y^*|\boldsymbol{w}^T \underbrace{\phi(\boldsymbol{x}^*)}_{\phi_*}, \sigma^2)$$

- The *posterior predictive distribution* is given by

$$p(y^*|\boldsymbol{y}, \boldsymbol{x}^*) = \int \underbrace{p(y^*|\boldsymbol{x}^*, \boldsymbol{w})}_{\text{Pred. likelihood}} \underbrace{p(\boldsymbol{w}|\boldsymbol{y})}_{\text{Posterior}} \mathrm{d}\boldsymbol{w}$$



But how about making predictions?

- We defined a *linear* regression model

$$y_i = \boldsymbol{w}^T \phi(\boldsymbol{x}_i) + \epsilon_i$$

- We derived the *posterior distribution*

$$p(\boldsymbol{w}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{S})$$

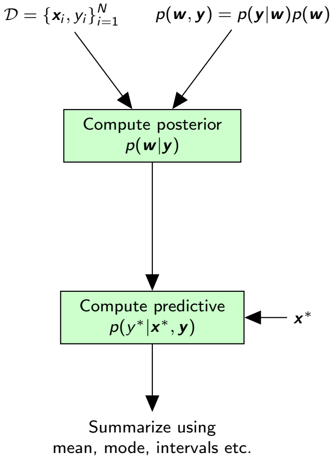
- Next goal: *making predictions* for new $\boldsymbol{x}^*$

$$p(y^*|\boldsymbol{x}^*, \boldsymbol{w}) = \mathcal{N}(y^*|\boldsymbol{w}^T \underbrace{\phi(\boldsymbol{x}^*)}_{\phi_*}, \sigma^2)$$

- The *posterior predictive distribution* is given by

$$p(y^*|\boldsymbol{y}, \boldsymbol{x}^*) = \int \underbrace{p(y^*|\boldsymbol{x}^*, \boldsymbol{w})}_{\text{Pred. likelihood}} \underbrace{p(\boldsymbol{w}|\boldsymbol{y})}_{\text{Posterior}} \mathrm{d}\boldsymbol{w}$$

$$= \int \mathcal{N}(y^*|\boldsymbol{w}^T \phi_*, \sigma^2) \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{S}) \mathrm{d}\boldsymbol{w}$$



But how about making predictions?

- We defined a *linear* regression model

$$y_i = \boldsymbol{w}^T \phi(\boldsymbol{x}_i) + \epsilon_i$$

- We derived the *posterior distribution*

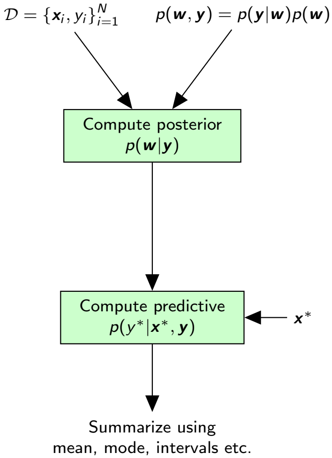
$$p(\boldsymbol{w}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{S})$$

- Next goal: *making predictions* for new $\boldsymbol{x}^*$

$$p(y^*|\boldsymbol{x}^*, \boldsymbol{w}) = \mathcal{N}(y^*|\boldsymbol{w}^T \underbrace{\phi(\boldsymbol{x}^*)}_{\phi_*}, \sigma^2)$$

- The *posterior predictive distribution* is given by

$$\begin{aligned} p(y^*|\boldsymbol{y}, \boldsymbol{x}^*) &= \int \underbrace{p(y^*|\boldsymbol{x}^*, \boldsymbol{w})}_{\text{Pred. likelihood}} \underbrace{p(\boldsymbol{w}|\boldsymbol{y})}_{\text{Posterior}} \mathrm{d}\boldsymbol{w} \\ &= \int \mathcal{N}(y^*|\boldsymbol{w}^T\phi_*, \sigma^2)\mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{S})\mathrm{d}\boldsymbol{w} \\ &= \mathcal{N}(y^*|\mathbf{m}^T\phi_*, \phi_*^T\boldsymbol{S}\phi_* + \sigma^2) \end{aligned}$$



But how about making predictions?

- We defined a *linear* regression model

$$y_i = \mathbf{w}^T \phi(\mathbf{x}_i) + \epsilon_i$$

- We derived the *posterior distribution*

$$p(\mathbf{w}|\mathbf{y}) = \mathcal{N}(\mathbf{w}|\mathbf{m}, \mathbf{S})$$

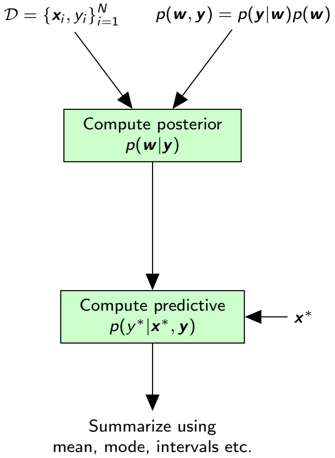
- Next goal: *making predictions* for new $\mathbf{x}^*$

$$p(y^*|\mathbf{x}^*, \mathbf{w}) = \mathcal{N}(y^*|\mathbf{w}^T \underbrace{\phi(\mathbf{x}^*)}_{\phi_*}, \sigma^2)$$

- The *posterior predictive distribution* is given by

$$\begin{aligned} p(y^*|\mathbf{y}, \mathbf{x}^*) &= \int \underbrace{p(y^*|\mathbf{x}^*, \mathbf{w})}_{\text{Pred. likelihood}} \underbrace{p(\mathbf{w}|\mathbf{y})}_{\text{Posterior}} \mathrm{d}\mathbf{w} \\ &= \int \mathcal{N}(y^*|\mathbf{w}^T \phi_*, \sigma^2) \mathcal{N}(\mathbf{w}|\mathbf{m}, \mathbf{S}) \mathrm{d}\mathbf{w} \\ &= \mathcal{N}(y^*|\mathbf{m}^T \phi_*, \phi_*^T \mathbf{S} \phi_* + \sigma^2) \end{aligned}$$

- Can be derived using eq. (3.38) in Murphy1



Quiz time!

$$p(\boldsymbol{w}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{w})p(\boldsymbol{w})}{p(\boldsymbol{y})} \qquad (\textit{Bayes' rule})$$

$$p(\boldsymbol{y}) = \int p(\boldsymbol{y}|\boldsymbol{w})p(\boldsymbol{w})\mathrm{d}\boldsymbol{w} \qquad (\textit{marginal likelihood})$$

$$p(y^*|\boldsymbol{y}, \boldsymbol{x}^*) = \int p(y^*|\boldsymbol{x}^*, \boldsymbol{w})p(\boldsymbol{w}|\boldsymbol{y})\mathrm{d}\boldsymbol{w} \qquad (\textit{Posterior predictive dist.})$$

- Spend 5 minutes DTU Learn quiz: "Lecture 3: Bayesian inference"

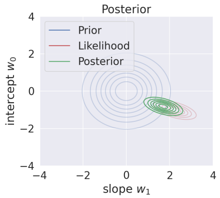
Posterior Predictive distributions

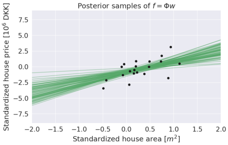
Building intuition

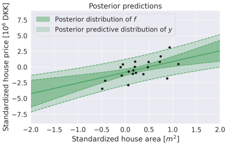
- The *posterior distribution* is $p(\boldsymbol{w}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{S})$

- Two *model quantities* of interest

$$f^* = f(\boldsymbol{x}_*) = \boldsymbol{w}^T \boldsymbol{\phi}_*$$
$$y^* = y(\boldsymbol{x}_*) = \boldsymbol{w}^T \boldsymbol{\phi}_* + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$







Posterior Predictive distributions

Building intuition

- The *posterior distribution* is $p(\boldsymbol{w}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{S})$

- Two *model quantities* of interest

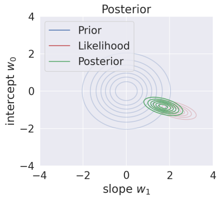
$$f^* = f(\boldsymbol{x}_*) = \boldsymbol{w}^T \boldsymbol{\phi}_*$$

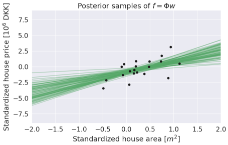
$$y^* = y(\boldsymbol{x}_*) = \boldsymbol{w}^T \boldsymbol{\phi}_* + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

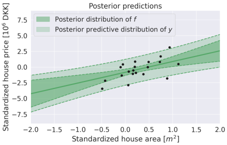
- Computing the *posterior* mean and variance

$$\mathbb{E}\left[f^*\right] =$$

$$\mathbb{V}\left[f^*\right] =$$







Posterior Predictive distributions

Building intuition

- The *posterior distribution* is $p(\boldsymbol{w}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{S})$

- Two *model quantities* of interest

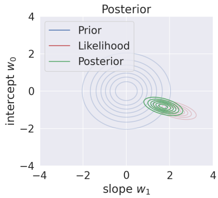
$$f^* = f(\boldsymbol{x}_*) = \boldsymbol{w}^T \boldsymbol{\phi}_*$$

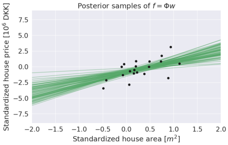
$$y^* = y(\boldsymbol{x}_*) = \boldsymbol{w}^T \boldsymbol{\phi}_* + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

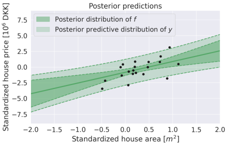
- Computing the *posterior* mean and variance

$$\mathbb{E}\left[f^*\right] = \mathbb{E}\left[\boldsymbol{w}^T \boldsymbol{\phi}_*\right] =$$

$$\mathbb{V}\left[f^*\right] =$$







Posterior Predictive distributions

Building intuition

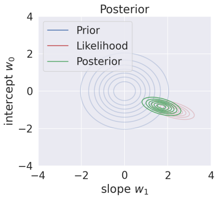
- The *posterior distribution* is $p(\boldsymbol{w}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{S})$

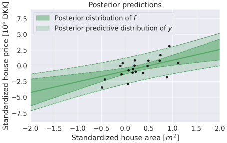
- Two *model quantities* of interest

$$f^* = f(\boldsymbol{x}_*) = \boldsymbol{w}^T \boldsymbol{\phi}_*$$
$$y^* = y(\boldsymbol{x}_*) = \boldsymbol{w}^T \boldsymbol{\phi}_* + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

- Computing the *posterior* mean and variance

$$\mathbb{E}\left[f^*\right] = \mathbb{E}\left[\boldsymbol{w}^T \boldsymbol{\phi}_*\right] = \mathbb{E}\left[\boldsymbol{w}^T\right] \boldsymbol{\phi}_* =$$
$$\mathbb{V}\left[f^*\right] =$$





Posterior Predictive distributions

Building intuition

- The *posterior distribution* is $p(\boldsymbol{w}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{S})$

- Two *model quantities* of interest

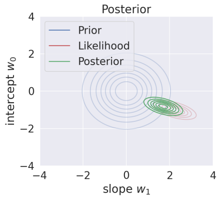
$$f^* = f(\boldsymbol{x}_*) = \boldsymbol{w}^T \boldsymbol{\phi}_*$$

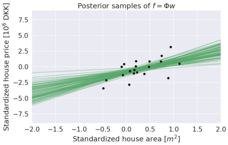
$$y^* = y(\boldsymbol{x}_*) = \boldsymbol{w}^T \boldsymbol{\phi}_* + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

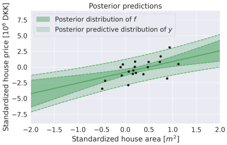
- Computing the *posterior* mean and variance

$$\mathbb{E}\left[f^*\right] = \mathbb{E}\left[\boldsymbol{w}^T \boldsymbol{\phi}_*\right] = \mathbb{E}\left[\boldsymbol{w}^T\right] \boldsymbol{\phi}_* = \boldsymbol{m}^T \boldsymbol{\phi}_*$$

$$\mathbb{V}\left[f^*\right] =$$







Posterior Predictive distributions

Building intuition

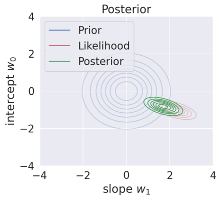
- The *posterior distribution* is $p(\boldsymbol{w}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{S})$

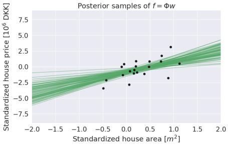
- Two *model quantities* of interest

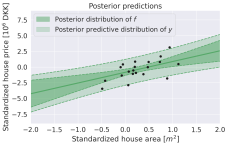
$$f^* = f(\boldsymbol{x}_*) = \boldsymbol{w}^T \boldsymbol{\phi}_*$$
$$y^* = y(\boldsymbol{x}_*) = \boldsymbol{w}^T \boldsymbol{\phi}_* + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

- Computing the *posterior* mean and variance

$$\mathbb{E}\left[f^*\right] = \mathbb{E}\left[\boldsymbol{w}^T \boldsymbol{\phi}_*\right] = \mathbb{E}\left[\boldsymbol{w}^T\right] \boldsymbol{\phi}_* = \boldsymbol{m}^T \boldsymbol{\phi}_*$$
$$\mathbb{V}\left[f^*\right] = \mathbb{V}\left[\boldsymbol{w}^T \boldsymbol{\phi}_*\right] =$$







Posterior Predictive distributions

Building intuition

- The *posterior distribution* is $p(\boldsymbol{w}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{S})$

- Two *model quantities* of interest

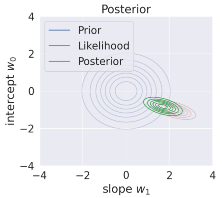
$$f^* = f(\boldsymbol{x}_*) = \boldsymbol{w}^T \boldsymbol{\phi}_*$$

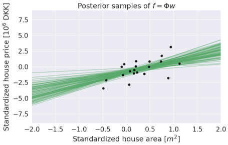
$$y^* = y(\boldsymbol{x}_*) = \boldsymbol{w}^T \boldsymbol{\phi}_* + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

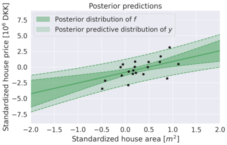
- Computing the *posterior* mean and variance

$$\mathbb{E}\left[f^*\right] = \mathbb{E}\left[\boldsymbol{w}^T \boldsymbol{\phi}_*\right] = \mathbb{E}\left[\boldsymbol{w}^T\right] \boldsymbol{\phi}_* = \boldsymbol{m}^T \boldsymbol{\phi}_*$$

$$\mathbb{V}\left[f^*\right] = \mathbb{V}\left[\boldsymbol{w}^T \boldsymbol{\phi}_*\right] = \boldsymbol{\phi}_*^T \mathbb{V}\left[\boldsymbol{w}^T\right] \boldsymbol{\phi}_* =$$







Posterior Predictive distributions

Building intuition

- The *posterior distribution* is $p(\boldsymbol{w}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{S})$

- Two *model quantities* of interest

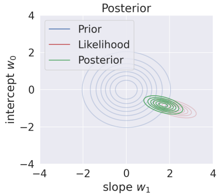
$$f^* = f(\boldsymbol{x}_*) = \boldsymbol{w}^T \boldsymbol{\phi}_*$$

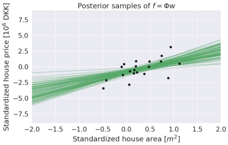
$$y^* = y(\boldsymbol{x}_*) = \boldsymbol{w}^T \boldsymbol{\phi}_* + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

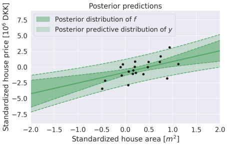
- Computing the *posterior* mean and variance

$$\mathbb{E}\left[f^*\right] = \mathbb{E}\left[\boldsymbol{w}^T \boldsymbol{\phi}_*\right] = \mathbb{E}\left[\boldsymbol{w}^T\right] \boldsymbol{\phi}_* = \boldsymbol{m}^T \boldsymbol{\phi}_*$$

$$\mathbb{V}\left[f^*\right] = \mathbb{V}\left[\boldsymbol{w}^T \boldsymbol{\phi}_*\right] = \boldsymbol{\phi}_*^T \mathbb{V}\left[\boldsymbol{w}^T\right] \boldsymbol{\phi}_* = \boldsymbol{\phi}_*^T \boldsymbol{S} \boldsymbol{\phi}_*$$







Posterior Predictive distributions

Building intuition

- The *posterior distribution* is $p(\boldsymbol{w}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{S})$

- Two *model quantities* of interest

$$f^* = f(\boldsymbol{x}_*) = \boldsymbol{w}^T \boldsymbol{\phi}_*$$

$$y^* = y(\boldsymbol{x}_*) = \boldsymbol{w}^T \boldsymbol{\phi}_* + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

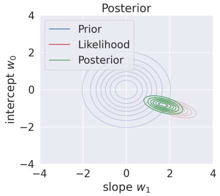
- Computing the *posterior* mean and variance

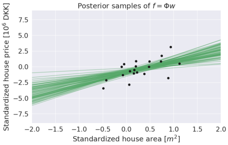
$$\mathbb{E}\left[f^*\right] = \mathbb{E}\left[\boldsymbol{w}^T \boldsymbol{\phi}_*\right] = \mathbb{E}\left[\boldsymbol{w}^T\right] \boldsymbol{\phi}_* = \boldsymbol{m}^T \boldsymbol{\phi}_*$$

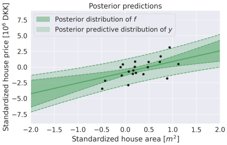
$$\mathbb{V}\left[f^*\right] = \mathbb{V}\left[\boldsymbol{w}^T \boldsymbol{\phi}_*\right] = \boldsymbol{\phi}_*^T \mathbb{V}\left[\boldsymbol{w}^T\right] \boldsymbol{\phi}_* = \boldsymbol{\phi}_*^T \boldsymbol{S} \boldsymbol{\phi}_*$$

- Hence, the *posterior distribution* of f^* given the data $\boldsymbol{y}$ is

$$p(f^*|\boldsymbol{y}, \boldsymbol{x}_*) = \mathcal{N}(f^*|\mathbf{m}^T \boldsymbol{\phi}_*, \boldsymbol{\phi}_*^T \boldsymbol{S} \boldsymbol{\phi}_*)$$







Posterior Predictive distributions

Building intuition

- The *posterior distribution* is $p(\boldsymbol{w}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{S})$

- Two *model quantities* of interest

$$f^* = f(\boldsymbol{x}_*) = \boldsymbol{w}^T\boldsymbol{\phi}_*$$
$$y^* = y(\boldsymbol{x}_*) = \boldsymbol{w}^T\boldsymbol{\phi}_* + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

- Computing the *posterior* mean and variance

$$\mathbb{E}\left[f^*\right] = \mathbb{E}\left[\boldsymbol{w}^T\boldsymbol{\phi}_*\right] = \mathbb{E}\left[\boldsymbol{w}^T\right]\boldsymbol{\phi}_* = \boldsymbol{m}^T\boldsymbol{\phi}_*$$
$$\mathbb{V}\left[f^*\right] = \mathbb{V}\left[\boldsymbol{w}^T\boldsymbol{\phi}_*\right] = \boldsymbol{\phi}_*^T\mathbb{V}\left[\boldsymbol{w}^T\right]\boldsymbol{\phi}_* = \boldsymbol{\phi}_*^T\boldsymbol{S}\boldsymbol{\phi}_*$$

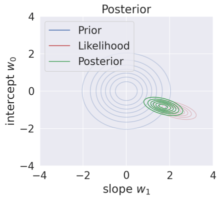
- Hence, the *posterior distribution* of f^* given the data $\boldsymbol{y}$ is

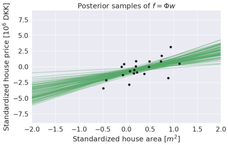
$$p(f^*|\boldsymbol{y}, \boldsymbol{x}_*) = \mathcal{N}(f^*|\mathbf{m}^T\boldsymbol{\phi}_*, \boldsymbol{\phi}_*^T\boldsymbol{S}\boldsymbol{\phi}_*)$$

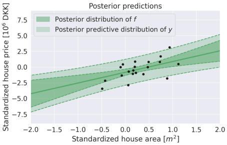
- and the *posterior predictive distribution* for y^* becomes

$$p(y^*|\boldsymbol{y}, \boldsymbol{x}_*) = \mathcal{N}(y^*|\mathbf{m}^T\boldsymbol{\phi}_*, \boldsymbol{\phi}_*^T\boldsymbol{S}\boldsymbol{\phi}_* + \sigma^2)$$

since $\mathbb{V}[\epsilon] = \sigma^2$







Posterior Predictive distributions

Building intuition

- The *posterior distribution* is $p(\boldsymbol{w}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{S})$

- Two *model quantities* of interest

$$f^* = f(\boldsymbol{x}_*) = \boldsymbol{w}^T \boldsymbol{\phi}_*$$

$$y^* = y(\boldsymbol{x}_*) = \boldsymbol{w}^T \boldsymbol{\phi}_* + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

- Computing the *posterior* mean and variance

$$\mathbb{E}\left[f^*\right] = \mathbb{E}\left[\boldsymbol{w}^T \boldsymbol{\phi}_*\right] = \mathbb{E}\left[\boldsymbol{w}^T\right] \boldsymbol{\phi}_* = \boldsymbol{m}^T \boldsymbol{\phi}_*$$

$$\mathbb{V}\left[f^*\right] = \mathbb{V}\left[\boldsymbol{w}^T \boldsymbol{\phi}_*\right] = \boldsymbol{\phi}_*^T \mathbb{V}\left[\boldsymbol{w}^T\right] \boldsymbol{\phi}_* = \boldsymbol{\phi}_*^T \boldsymbol{S} \boldsymbol{\phi}_*$$

- Hence, the *posterior distribution* of f^* given the data $\boldsymbol{y}$ is

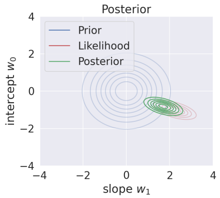
$$p(f^*|\boldsymbol{y}, \boldsymbol{x}_*) = \mathcal{N}(f^*|\mathbf{m}^T \boldsymbol{\phi}_*, \boldsymbol{\phi}_*^T \boldsymbol{S} \boldsymbol{\phi}_*)$$

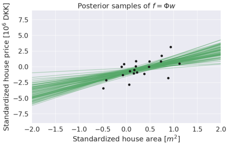
- and the *posterior predictive distribution* for y^* becomes

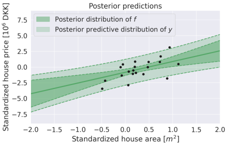
$$p(y^*|\boldsymbol{y}, \boldsymbol{x}_*) = \mathcal{N}(y^*|\mathbf{m}^T \boldsymbol{\phi}_*, \boldsymbol{\phi}_*^T \boldsymbol{S} \boldsymbol{\phi}_* + \sigma^2)$$

  since $\mathbb{V}\left[\epsilon\right] = \sigma^2$

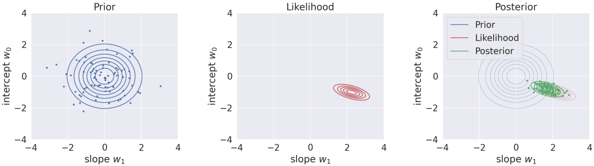
- We can also compute the *prior* predictive distribution.



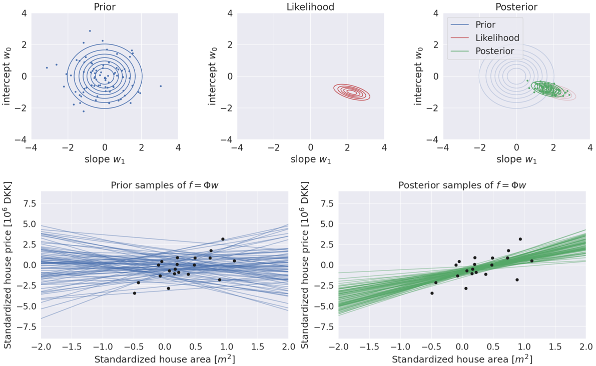




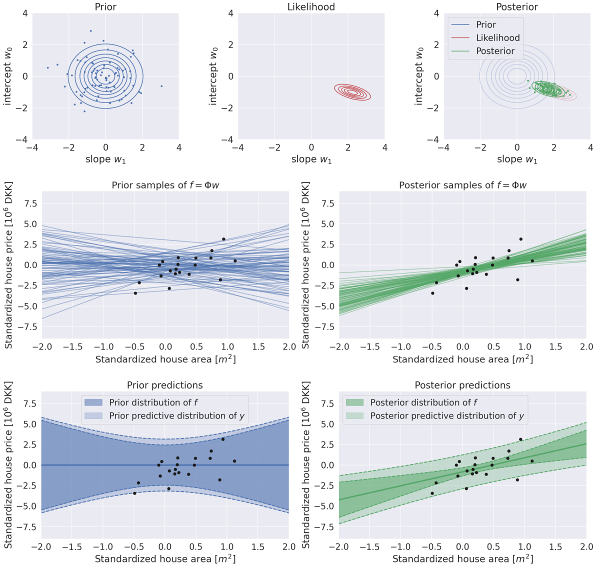
Prior and posterior predictive distributions



Prior and posterior predictive distributions

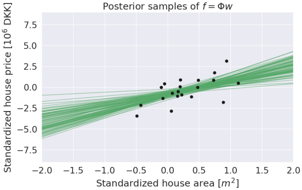


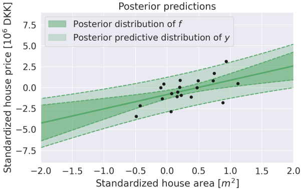
Prior and posterior predictive distributions



Posterior Predictive distributions

Epistemic and aleatoric uncertainty



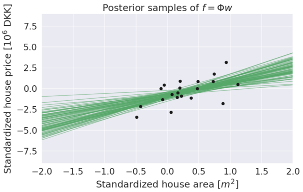


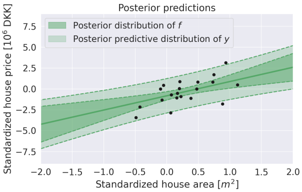
Posterior Predictive distributions

Epistemic and aleatoric uncertainty

- The *posterior predictive distribution* of $y^* = \boldsymbol{w}^T\phi(\boldsymbol{x}^*) + \epsilon$ is

$$p(y^*|\boldsymbol{y}, \boldsymbol{x}_*) = \mathcal{N}(y^*|\mathbf{m}^T\phi_*, \phi_*^T\boldsymbol{S}\phi_* + \sigma^2)$$





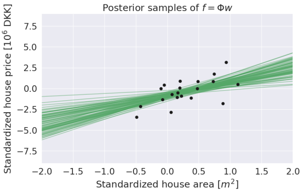
Posterior Predictive distributions

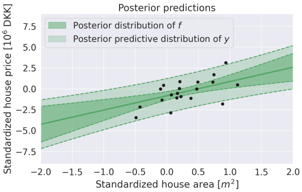
Epistemic and aleatoric uncertainty

- The *posterior predictive distribution* of $y^* = \boldsymbol{w}^T\phi(\boldsymbol{x}^*) + \epsilon$ is

$$p(y^*|\boldsymbol{y}, \boldsymbol{x}_*) = \mathcal{N}(y^*|\mathbf{m}^T\boldsymbol{\phi}_*, \boldsymbol{\phi}_*^T\boldsymbol{S}\boldsymbol{\phi}_* + \sigma^2)$$

- Variance: The first term is due to parameter uncertainty (*epistemic/reducible*) and the second term is due to measurement noise (*aleatoric/irreducible*)





Posterior Predictive distributions

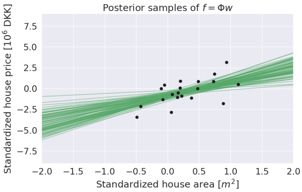
Epistemic and aleatoric uncertainty

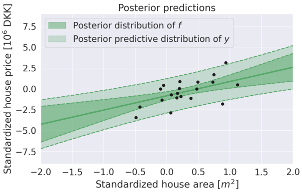
- The *posterior predictive distribution* of $y^* = \boldsymbol{w}^T\phi(\boldsymbol{x}^*) + \epsilon$ is

$$p(y^*|\boldsymbol{y}, \boldsymbol{x}_*) = \mathcal{N}(y^*|\mathbf{m}^T\phi_*, \phi_*^T\boldsymbol{S}\phi_* + \sigma^2)$$

- Variance: The first term is due to parameter uncertainty (*epistemic/reducible*) and the second term is due to measurement noise (*aleatoric/irreducible*)

- The term $\phi(\boldsymbol{x}_*)^T\boldsymbol{S}\phi(\boldsymbol{x}_*)$ is the *posterior uncertainty projected to data space*





Posterior Predictive distributions

Epistemic and aleatoric uncertainty

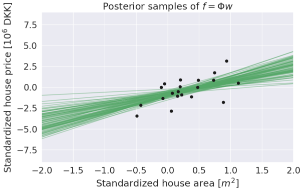
- The *posterior predictive distribution* of $y^* = \boldsymbol{w}^T\phi(\boldsymbol{x}^*) + \epsilon$ is

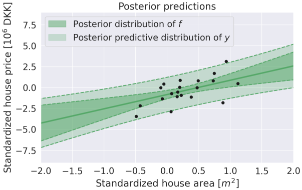
$$p(y^*|\boldsymbol{y}, \boldsymbol{x}_*) = \mathcal{N}(y^*|\mathbf{m}^T\phi_*, \phi_*^T\boldsymbol{S}\phi_* + \sigma^2)$$

- Variance: The first term is due to parameter uncertainty (*epistemic/reducible*) and the second term is due to measurement noise (*aleatoric/irreducible*)

- The term $\phi(\boldsymbol{x}_*)^T\boldsymbol{S}\phi(\boldsymbol{x}_*)$ is the *posterior uncertainty projected to data space*

- A couple of questions for you:
  1. What happens to the predictive variance when $N \to \infty$?
  2. Why is the uncertainty "U-shaped"?
  3. Why does the aleatoric uncertainty appear to dominate the total uncertainty near the data?



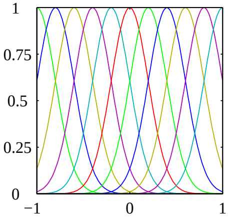


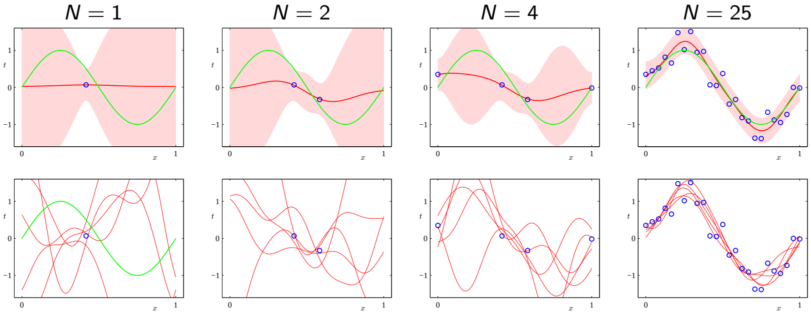
Example: posterior predictive distributions

- *Predictive distributions* for simple sinoidal toy dataset using Gaussian Basis functions

- Samples from the posterior

Feature extractors $\phi_i(x)$





Dealing with hyperparameters

But what about the hyperparameters?

- The Bayesian linear regression model

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|0, \alpha^{-1}\boldsymbol{I}) \qquad \textit{(prior)}$$

$$p(\boldsymbol{y}|\boldsymbol{w}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{\Phi}\boldsymbol{w}, \sigma^2\boldsymbol{I}) \qquad \textit{(likelihood)}$$

$$p(\boldsymbol{w}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{S}) \qquad \textit{(posterior)}$$

$$p(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{0}, \sigma^2\boldsymbol{I} + \alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^T) \qquad \textit{(marginal likelihood)}$$

But what about the hyperparameters?

- The Bayesian linear regression model

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|0, \alpha^{-1}\boldsymbol{I}) \qquad (\textit{prior})$$

$$p(\boldsymbol{y}|\boldsymbol{w}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{\Phi}\boldsymbol{w}, \sigma^2\boldsymbol{I}) \qquad (\textit{likelihood})$$

$$p(\boldsymbol{w}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{S}) \qquad (\textit{posterior})$$

$$p(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{0}, \sigma^2\boldsymbol{I} + \alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^T) \qquad (\textit{marginal likelihood})$$

- A fully Bayesian solution would require us to impose priors on α, β (recall $\beta = \frac{1}{\sigma^2}$)

$$p(\alpha, \beta|\boldsymbol{y}) \propto p(\boldsymbol{y}|\alpha, \beta)p(\alpha, \beta)$$

But what about the hyperparameters?

- The Bayesian linear regression model

$$
\begin{aligned}
p(\boldsymbol{w}) &= \mathcal{N}(\boldsymbol{w}|0, \alpha^{-1}\boldsymbol{I}) && (\textit{prior}) \\
p(\boldsymbol{y}|\boldsymbol{w}) &= \mathcal{N}(\boldsymbol{y}|\boldsymbol{\Phi}\boldsymbol{w}, \sigma^2\boldsymbol{I}) && (\textit{likelihood}) \\
p(\boldsymbol{w}|\boldsymbol{y}) &= \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{S}) && (\textit{posterior}) \\
p(\boldsymbol{y}) &= \mathcal{N}(\boldsymbol{y}|\boldsymbol{0}, \sigma^2\boldsymbol{I} + \alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^T) && (\textit{marginal likelihood})
\end{aligned}
$$

- A fully Bayesian solution would require us to impose priors on α, β (recall $\beta = \frac{1}{\sigma^2}$)

$$p(\alpha, \beta|\boldsymbol{y}) \propto p(\boldsymbol{y}|\alpha, \beta)p(\alpha, \beta)$$

- ... and integrate them out

$$p(y^*|\boldsymbol{y}) = \iiint p(y^*|\boldsymbol{w}, \beta)p(\boldsymbol{w}|\boldsymbol{y}, \alpha, \beta)p(\alpha, \beta|\boldsymbol{y}) \,\mathrm{d}\boldsymbol{w}\mathrm{d}\alpha\mathrm{d}\beta$$

But what about the hyperparameters?

- The Bayesian linear regression model

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|0, \alpha^{-1}\boldsymbol{I}) \qquad (\textit{prior})$$

$$p(\boldsymbol{y}|\boldsymbol{w}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{\Phi w}, \sigma^2\boldsymbol{I}) \qquad (\textit{likelihood})$$

$$p(\boldsymbol{w}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{S}) \qquad (\textit{posterior})$$

$$p(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{0}, \sigma^2\boldsymbol{I} + \alpha^{-1}\boldsymbol{\Phi\Phi}^T) \qquad (\textit{marginal likelihood})$$

- A fully Bayesian solution would require us to impose priors on α, β (recall $\beta = \frac{1}{\sigma^2}$)

$$p(\alpha, \beta|\boldsymbol{y}) \propto p(\boldsymbol{y}|\alpha, \beta)p(\alpha, \beta)$$

- ... and integrate them out

$$p(y^*|\boldsymbol{y}) = \iiint p(y^*|\boldsymbol{w}, \beta)p(\boldsymbol{w}|\boldsymbol{y}, \alpha, \beta)p(\alpha, \beta|\boldsymbol{y})\,\mathrm{d}\boldsymbol{w}\mathrm{d}\alpha\mathrm{d}\beta$$

$$= \mathbb{E}_{p(\alpha,\beta|\boldsymbol{y})}\left[\quad \mathbb{E}_{p(\boldsymbol{w}|\boldsymbol{y},\alpha,\beta)}\left[p(y^*|\boldsymbol{w}, \beta)\right] \quad\right]$$

But what about the hyperparameters?

- The Bayesian linear regression model

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|0, \alpha^{-1}\boldsymbol{I}) \qquad (\textit{prior})$$

$$p(\boldsymbol{y}|\boldsymbol{w}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{\Phi}\boldsymbol{w}, \sigma^2\boldsymbol{I}) \qquad (\textit{likelihood})$$

$$p(\boldsymbol{w}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{S}) \qquad (\textit{posterior})$$

$$p(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{0}, \sigma^2\boldsymbol{I} + \alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^T) \qquad (\textit{marginal likelihood})$$

- A fully Bayesian solution would require us to impose priors on α, β (recall $\beta = \frac{1}{\sigma^2}$)

$$p(\alpha, \beta|\boldsymbol{y}) \propto p(\boldsymbol{y}|\alpha, \beta)p(\alpha, \beta)$$

- ... and integrate them out

$$p(y^*|\boldsymbol{y}) = \iiint p(y^*|\boldsymbol{w}, \beta)p(\boldsymbol{w}|\boldsymbol{y}, \alpha, \beta)p(\alpha, \beta|\boldsymbol{y})\, \mathrm{d}\boldsymbol{w}\mathrm{d}\alpha\mathrm{d}\beta$$

$$= \mathbb{E}_{p(\alpha,\beta|\boldsymbol{y})}\left[\quad \mathbb{E}_{p(\boldsymbol{w}|\boldsymbol{y},\alpha,\beta)}\left[p(y^*|\boldsymbol{w}, \beta)\right] \quad\right]$$

- ... but this is not analytically tractable. Later in the course we will learn tools to deal with this in general

The evidence approximation

- How to deal with this bastard?

$$p(y^*|\boldsymbol{y}) = \int\int\int p(y^*|\boldsymbol{w},\beta)p(\boldsymbol{w}|\boldsymbol{y},\alpha,\beta)p(\alpha,\beta|\boldsymbol{y})\mathrm{d}\boldsymbol{w}\mathrm{d}\alpha\mathrm{d}\beta$$

$$= \mathbb{E}_{p(\alpha,\beta|\boldsymbol{y})}\left[\quad \mathbb{E}_{p(\boldsymbol{w}|\boldsymbol{y},\alpha,\beta)}\left[p(y^*|\boldsymbol{w},\beta)\right] \quad\right]$$

The evidence approximation

- How to deal with this bastard?

$$p(y^*|\boldsymbol{y}) = \iiint p(y^*|\boldsymbol{w},\beta)p(\boldsymbol{w}|\boldsymbol{y},\alpha,\beta)p(\alpha,\beta|\boldsymbol{y})\mathrm{d}\boldsymbol{w}\mathrm{d}\alpha\mathrm{d}\beta$$
$$= \mathbb{E}_{p(\alpha,\beta|\boldsymbol{y})}\left[\quad \mathbb{E}_{p(\boldsymbol{w}|\boldsymbol{y},\alpha,\beta)}\left[p(y^*|\boldsymbol{w},\beta)\right]\quad\right]$$

- *The evidence approximation*, *maximum likelihood type II*, *Empirical Bayes*

The evidence approximation

- How to deal with this bastard?

$$p(y^*|\boldsymbol{y}) = \iiint p(y^*|\boldsymbol{w},\beta)p(\boldsymbol{w}|\boldsymbol{y},\alpha,\beta)p(\alpha,\beta|\boldsymbol{y})\mathrm{d}\boldsymbol{w}\mathrm{d}\alpha\mathrm{d}\beta$$
$$= \mathbb{E}_{p(\alpha,\beta|\boldsymbol{y})}\left[\quad \mathbb{E}_{p(\boldsymbol{w}|\boldsymbol{y},\alpha,\beta)}\left[p(y^*|\boldsymbol{w},\beta)\right] \quad\right]$$

- *The evidence approximation*, *maximum likelihood type II*, *Empirical Bayes*

- If the posterior is sharply peaked around $\hat{\alpha}$ and $\hat{\beta}$, then $p(\alpha,\beta|\boldsymbol{y})$ can be approximated by a Dirac's delta distribution

$$p(y^*|\boldsymbol{y}) \approx \mathbb{E}_{p(\boldsymbol{w}|\boldsymbol{y},\hat{\alpha},\hat{\beta})}\left[p(y^*|\boldsymbol{w},\beta)\right]$$

The evidence approximation

- How to deal with this bastard?

$$p(y^*|\boldsymbol{y}) = \int\!\!\int\!\!\int p(y^*|\boldsymbol{w},\beta)p(\boldsymbol{w}|\boldsymbol{y},\alpha,\beta)p(\alpha,\beta|\boldsymbol{y})\mathrm{d}\boldsymbol{w}\mathrm{d}\alpha\mathrm{d}\beta$$
$$= \mathbb{E}_{p(\alpha,\beta|\boldsymbol{y})}\left[\ \mathbb{E}_{p(\boldsymbol{w}|\boldsymbol{y},\alpha,\beta)}\left[p(y^*|\boldsymbol{w},\beta)\right]\ \right]$$

- *The evidence approximation*, *maximum likelihood type II*, *Empirical Bayes*

- If the posterior is sharply peaked around $\hat{\alpha}$ and $\hat{\beta}$, then $p(\alpha,\beta|\boldsymbol{y})$ can be approximated by a Dirac's delta distribution

$$p(y^*|\boldsymbol{y}) \approx \mathbb{E}_{p(\boldsymbol{w}|\boldsymbol{y},\hat{\alpha},\hat{\beta})}\left[p(y^*|\boldsymbol{w},\beta)\right]$$

- Assume we impose a *flat prior* on α and β, then

$$p(\alpha,\beta|\boldsymbol{y}) \propto p(\boldsymbol{y}|\alpha,\beta)p(\alpha,\beta) \propto p(\boldsymbol{y}|\alpha,\beta)$$

The evidence approximation

- How to deal with this bastard?

$$p(y^*|\boldsymbol{y}) = \iiint p(y^*|\boldsymbol{w},\beta)p(\boldsymbol{w}|\boldsymbol{y},\alpha,\beta)p(\alpha,\beta|\boldsymbol{y})\mathrm{d}\boldsymbol{w}\mathrm{d}\alpha\mathrm{d}\beta$$
$$= \mathbb{E}_{p(\alpha,\beta|\boldsymbol{y})}\left[\quad \mathbb{E}_{p(\boldsymbol{w}|\boldsymbol{y},\alpha,\beta)}\left[p(y^*|\boldsymbol{w},\beta)\right] \quad\right]$$

- *The evidence approximation*, *maximum likelihood type II*, *Empirical Bayes*

- If the posterior is sharply peaked around $\hat{\alpha}$ and $\hat{\beta}$, then $p(\alpha,\beta|\boldsymbol{y})$ can be approximated by a Dirac's delta distribution

$$p(y^*|\boldsymbol{y}) \approx \mathbb{E}_{p(\boldsymbol{w}|\boldsymbol{y},\hat{\alpha},\hat{\beta})}\left[p(y^*|\boldsymbol{w},\beta)\right]$$

- Assume we impose a *flat prior* on α and β, then

$$p(\alpha,\beta|\boldsymbol{y}) \propto p(\boldsymbol{y}|\alpha,\beta)p(\alpha,\beta) \propto p(\boldsymbol{y}|\alpha,\beta)$$

- We can estimate $\hat{\alpha},\hat{\beta}$ by *optimizing the marginal likelihood* $p(\boldsymbol{y}|\alpha,\beta)$

$$\hat{\alpha},\hat{\beta} = \arg\max_{\alpha,\beta} \log p(\boldsymbol{y}|\alpha,\beta)$$

Sinoidal example revisited using the evidence approximation