UPPSALA UNIVERSITET

Advanced Probabilistic Machine Learning

Lecture 2 – Bayesian linear regression

UPPSALA UNIVERSITET

Niklas Wahlström
Division of Systems and Control
Department of Information Technology
Uppsala University

niklas.wahlstrom@it.uu.se
www.it.uu.se/katalog/nikwa778

Summary of lecture 1 (I/IV)

Conditional probability is defined as

$$p(x|y) = \frac{p(x,y)}{p(y)} \quad \text{where} \quad p(y) \neq 0$$

Marginalization is defined as

$$p(x) = \sum_y p(x,y) \quad \text{or} \quad p(x) = \int_y p(x,y)dy$$

Much of the probability theory can be derived from these two rules.

Bayes' theorem is derived by using the def. of conditional probability twice

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

Summary of lecture 1 (II/IV)

In this course we solve problems using Bayes' theorem

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}$$

- $\mathcal{D}$: observed data
- θ : parameters of some model explaining the data
- $p(\theta)$: **prior** belief of parameters before we collected any data
- $p(\theta|\mathcal{D})$: **posterior** belief of parameters after inferring data
- $p(\mathcal{D}|\theta)$: **likelihood** of the data in view of the parameters
- $p(\mathcal{D})$: The **marginal likelihood**

UPPSALA UNIVERSITET

Summary of lecture 1 (III/IV)

If we view the quantities as functions of θ, we can disregard the normalization constant $p(y)$.

$$\underbrace{p(\theta|\mathcal{D})}_{\text{posterior}} \propto \underbrace{p(\mathcal{D}|\theta)}_{\text{likelihood}}\, \underbrace{p(\theta)}_{\text{prior}}$$

Conjugate prior A prior ensuring that the posterior and the prior belong to the same probability distribution family.

Example: Beta-Binomial

$$\underbrace{\text{Beta}\,(\mu;\, a^*,\, b^*)}_{\text{posterior}} \propto \underbrace{\text{Bin}\,(m;\, N,\, \mu)}_{\text{likelihood}}\, \underbrace{\text{Beta}\,(\mu;\, a,\, b)}_{\text{prior}} \qquad a^* = a + m$$

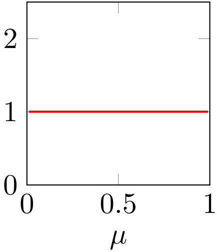
$$b^* = b + N - m$$

Beta distribution is a conjugate prior to the binomial likelihood.

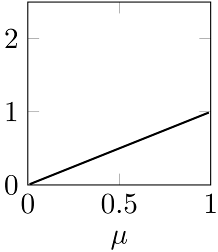
UPPSALA UNIVERSITET

Summary of lecture 1 (IV/IV)

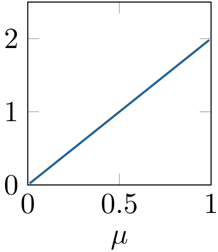
Prior	Likelihood function	Posterior
$p(\mu)$	$p(m\|\mu)$	$p(\mu\|m)$
Beta $(\mu;\, a,\, b)$	Bin $(m;\, N,\, \mu)$	Beta $(\mu;\, a^*,\, b^*)$
$a = 1, b = 1$	$m = 1, N = 1$	$a^* = 2, b^* = 1$



$\times$



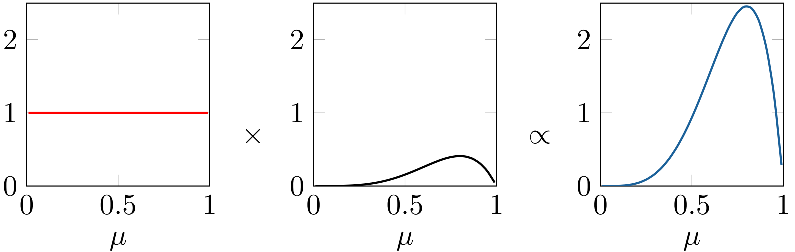
$\propto$



Assume you get $N = 1$ data point, of which $m = 1$ is head, $\mathcal{D} = \{1\}$.

Summary of lecture 1 (IV/IV)

Prior	Likelihood function	Posterior
$p(\mu)$	$p(m\|\mu)$	$p(\mu\|m)$
Beta $(\mu;\ a,\ b)$	Bin $(m;\ N,\ \mu)$	Beta $(\mu;\ a^*,\ b^*)$
$a = 1, b = 1$	$m = 4, N = 5$	$a^* = 5, b^* = 2$

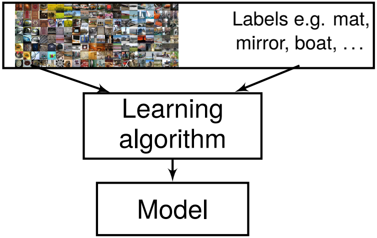


Assume you get $N = 5$ data points, of which $m = 4$ are heads, $\mathcal{D} = \{1, 0, 1, 1, 1\}$.

Supervised machine learning

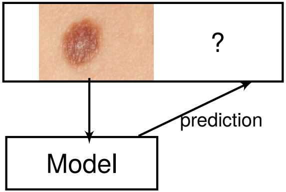
Learning a model from labeled data.

Training data



Predicting output of new data based on this model.

Unseen data



How do we rephrase supervised machine learning as a within the probabilistic methodology?

Supervised machine learning – Probabilistic perspective

Given: Data of inputs & outputs $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$.

Task: Predict the output y_* for a new unseen input $\mathbf{x}_*$.

Solution:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\mathsf{T} \\ \vdots \\ \mathbf{x}_N^\mathsf{T} \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$$

1. **Likelihood** Define the likelihood
$$p(\mathbf{y}|\theta, \mathbf{X})$$
2. **Prior** Define the prior $p(\theta)$
3. **Learning** Do inference by applying Bayes' theorem
$$p(\theta|\mathbf{y}, \mathbf{X}) \propto p(\mathbf{y}|\theta, \mathbf{X})p(\theta)$$
4. **Prediction** Compute **predictive distribution** by marginalizing
$$p(y_*|x_*, \mathbf{y}, \mathbf{X}) = \int p(y_*|\theta, x_*)p(\theta|\mathbf{y}, \mathbf{X})d\theta$$

Example: Linear regression model

- Recall the linear regression from lecture 2 in the SML course
- Now we introduce a prior over the parameter $\mathbf{w}$

Linear regression model

$$y_n = \mathbf{w}^\mathsf{T}\mathbf{x}_n + \varepsilon_n, \qquad \varepsilon_n \sim \mathcal{N}(0, \sigma^2), \qquad n = 1, ..., N$$
$$\mathbf{w} \sim p(\mathbf{w}).$$

Present assumptions:

1. y_n – observed **random** variable.
2. $\mathbf{w}$ – unknown **deterministic**
3. $\mathbf{x}_n$ – known **deterministic** variable.
4. ε_n – unknown **random** variable.
5. σ – known **deterministic** variable.

Linear regression: Maximum likelihood

Two equivalent ways of expressing the linear regression model:

1. $y_n = \mathbf{w}^\mathsf{T}\mathbf{x}_n + \varepsilon_n, \quad \varepsilon_n \sim \mathcal{N}(0, \sigma^2)$
2. $p(y_n \,|\, \mathbf{w}) = \mathcal{N}\left(y_n; \mathbf{w}^\mathsf{T}\mathbf{x}_n,\, \sigma^2\right)$.

The **likelihood** $p(\mathbf{y} \,|\, \mathbf{w})$ is given by

$$p(\mathbf{y} \,|\, \mathbf{w}) = \prod_{n=1}^{N} p(y_n \,|\, \mathbf{w}) = \prod_{n=1}^{N} \mathcal{N}\left(y_n; \mathbf{w}^\mathsf{T}\mathbf{x}_n,\, \sigma^2\right)$$
$$= \mathcal{N}\left(\mathbf{y}; \mathbf{X}\mathbf{w},\, \sigma^2\mathbf{I}_N\right).$$

$$\boxed{\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\mathsf{T} \\ \vdots \\ \mathbf{x}_N^\mathsf{T} \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}}$$

The solution if found by maximizing the likelihood

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} p(\mathbf{y} \,|\, \mathbf{w})$$

Example: Linear regression model

- Recall the linear regression from lecture 2 in the SML course
- Now we introduce a prior over the parameter $\mathbf{w}$

Bayesian linear regression model

$$y_n = \mathbf{w}^\mathsf{T}\mathbf{x}_n + \varepsilon_n, \qquad \varepsilon_n \sim \mathcal{N}(0, \sigma^2), \qquad n = 1, ..., N$$
$$\mathbf{w} \sim p(\mathbf{w}).$$

Present assumptions:

1. y_n – observed **random** variable.
2. $\mathbf{w}$ – unknown **random** variable. **(difference from SML)**
3. $\mathbf{x}_n$ – known **deterministic** variable.
4. ε_n – unknown **random** variable.
5. σ – known **deterministic** variable.

Bayesian linear regression model

Remember Bayes' theorem

$$p(\mathbf{w} \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \mathbf{w})p(\mathbf{w})}{p(\mathbf{y})}$$

- **Prior distribution:** $p(\mathbf{w})$ describes the knowledge we have about $\mathbf{w}$ before observing any data.
- **Likelihood:** $p(\mathbf{y} \mid \mathbf{w})$ described how "likely" the observed data is for a particular parameter value.
- **Posterior distribution:** $p(\mathbf{w} \mid \mathbf{y})$ summarize all our knowledge about $\mathbf{w}$ from the observed data and the model.

In Bayesian linear regression we use a Gaussian distribution as prior

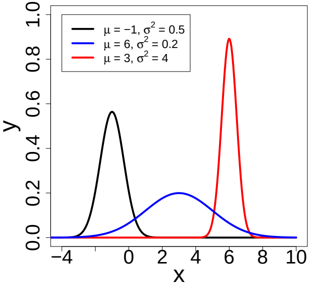
$$p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w}; \mathbf{m}_0, \, \boldsymbol{\Sigma}_0\right)$$

Scalar Gaussian (Normal) distribution

For a scalar variable x, the Gaussian distribution can be written on the form

$$\mathcal{N}\left(x; \mu, \sigma^2\right) = \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}}}_{Z} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- μ is the mean (expected value of the distribution)
- σ is the standard deviation
- σ^2 is the variance
- Z is the normalization constant



What if $\mathbf{x}$ is a vector $\mathbf{x} = \begin{pmatrix} x_1 & x_2 & \cdots & x_D \end{pmatrix}^\mathsf{T}$?

Multivariate Gaussian

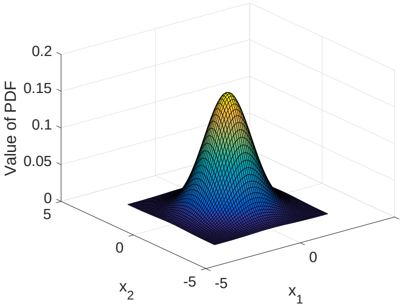
For a D-dimensional vector $\mathbf{x}$, the **multivariate** Gaussian distribution can be written on the form

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\underbrace{(2\pi)^{D/2}\sqrt{\det \boldsymbol{\Sigma}}}_{Z}} \exp\left(-\frac{1}{2} \underbrace{(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})}_{\text{quadratic form}} \right).$$

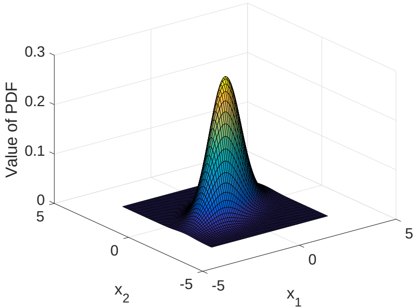
- $\boldsymbol{\mu}$ is the mean vector
- $\boldsymbol{\Sigma}$ is the covariance matrix
- Z is the normalization constant

Gaussian $\propto e^{\text{quadratic form}}$

Multivariate Gaussian



$$\mathbf{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$



$$\mathbf{\Sigma} = \begin{pmatrix} 1 & 0.4 \\ 0.4 & 0.5 \end{pmatrix}$$

UPPSALA UNIVERSITET

Partitioned Gaussian – marginalization

Partition the Gaussian random vector $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\mathbf{x} \in \mathbb{R}^n$ into two sets of random variables $\mathbf{x}_a \in \mathbb{R}^{n_a}$ and $\mathbf{x}_b \in \mathbb{R}^{n_b}$,

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

Task: Compute the marginal distribution $p(\mathbf{x}_a)$,

$$p(\mathbf{x}_a) = \int p(\mathbf{x}_a, \mathbf{x}_b) \mathrm{d}\mathbf{x}_b.$$

Partitioned Gaussian – marginalization

Theorem 1 (Marginalization)

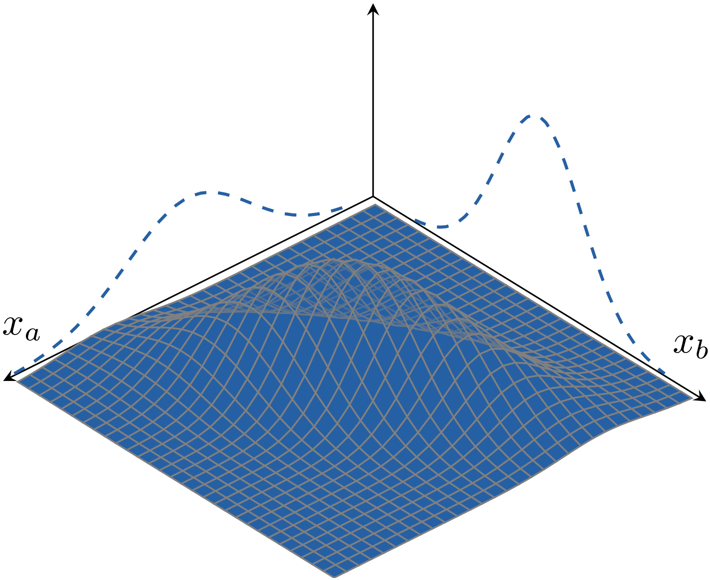
Partition the Gaussian random vector $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *according to*

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

The marginal distribution $p(\mathbf{x}_a)$ *is then given by*

$$p(\mathbf{x}_a) = \mathcal{N}\left(\mathbf{x}_a; \boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa}\right).$$

Partitioned Gaussian – marginalization



Partitioned Gaussian – Theorems

$$
\begin{tikzcd}
 & p(\mathbf{x}_a, \mathbf{x}_b) \arrow[dl, "\text{Thm 1}"'] \arrow[ddl, bend right] \\
p(\mathbf{x}_a) & \\
p(\mathbf{x}_b) &
\end{tikzcd}
$$

UPPSALA UNIVERSITET

Partitioned Gaussian – conditioning

Theorem 2 (Conditioning)

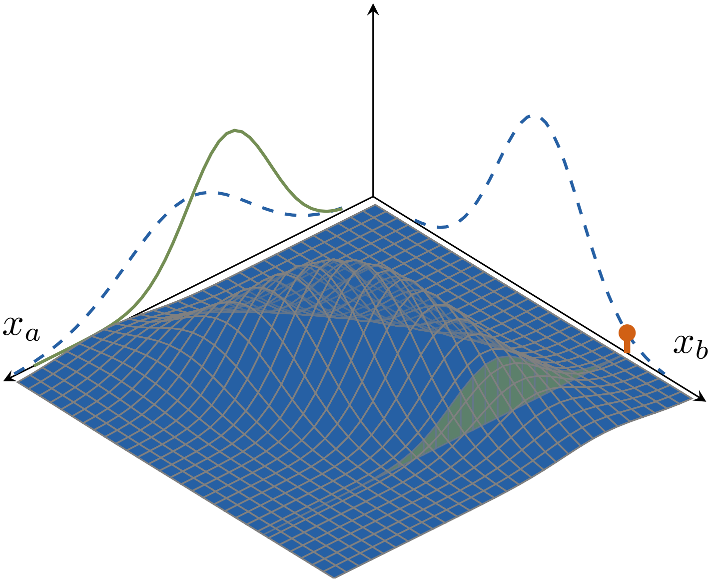
Partition the Gaussian random vector $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *according to*

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

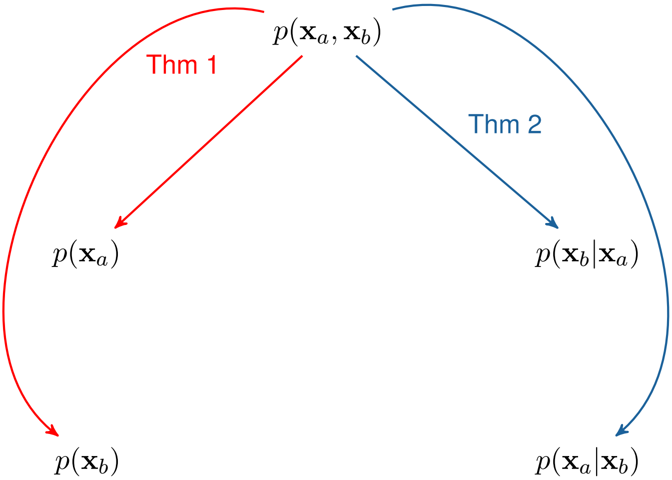
The conditional distribution $p(\mathbf{x}_a \,|\, \mathbf{x}_b)$ *is then given by*

$$\begin{aligned} p(\mathbf{x}_a \,|\, \mathbf{x}_b) &= \mathcal{N}\left(\mathbf{x}_a; \boldsymbol{\mu}_{a\,|\,b}, \boldsymbol{\Sigma}_{a\,|\,b}\right), \\ \boldsymbol{\mu}_{a\,|\,b} &= \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b), \\ \boldsymbol{\Sigma}_{a\,|\,b} &= \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}. \end{aligned}$$

Partitioned Gaussian – conditioning



Partitioned Gaussian – Theorems



Affine transformation of multivar. Gauss

We can also do the opposite:
compute $p(\mathbf{x}_a, \mathbf{x}_b)$ based on $p(\mathbf{x}_b \,|\, \mathbf{x}_a)$ and $p(\mathbf{x}_a)$

Theorem 3 (Affine transformation)

Assume that $\mathbf{x}_a$, as well as $\mathbf{x}_b$ conditioned on $\mathbf{x}_a$, are Gaussian distributed according to

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

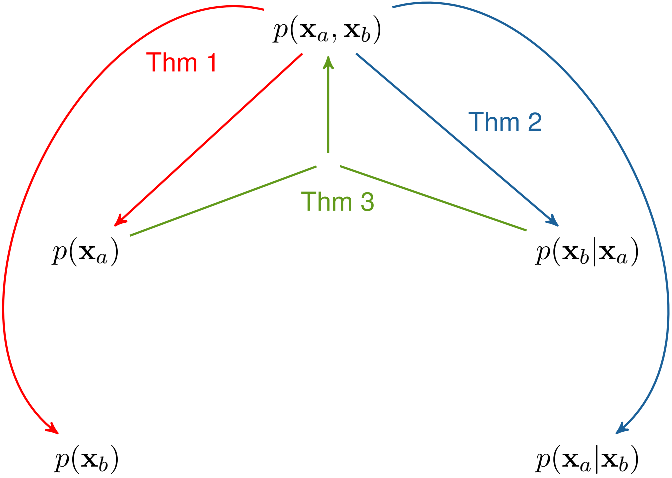
Then the joint distribution of $\mathbf{x}_a$ and $\mathbf{x}_b$ is

$$p(\mathbf{x}_a, \mathbf{x}_b) = \mathcal{N}\left(\begin{bmatrix}\mathbf{x}_a \\ \mathbf{x}_b\end{bmatrix}; \begin{bmatrix}\boldsymbol{\mu}_a \\ \mathbf{M}\boldsymbol{\mu}_a\end{bmatrix}, \mathbf{R}\right)$$

with

$$\mathbf{R} = \begin{bmatrix}\boldsymbol{\Sigma}_a & \boldsymbol{\Sigma}_a\mathbf{M}^\mathsf{T} \\ \mathbf{M}\boldsymbol{\Sigma}_a & \boldsymbol{\Sigma}_{b\,|\,a} + \mathbf{M}\boldsymbol{\Sigma}_a\mathbf{M}^\mathsf{T}\end{bmatrix}$$

Partitioned Gaussian – Theorems



Bayesian linear regression model

Bayesian linear regression model:

$$y_n = \mathbf{w}^\mathsf{T}\mathbf{x}_n + \varepsilon_n, \qquad \varepsilon_n \sim \mathcal{N}(0, \beta^{-1}),$$
$$\mathbf{w} \sim p(\mathbf{w}).$$

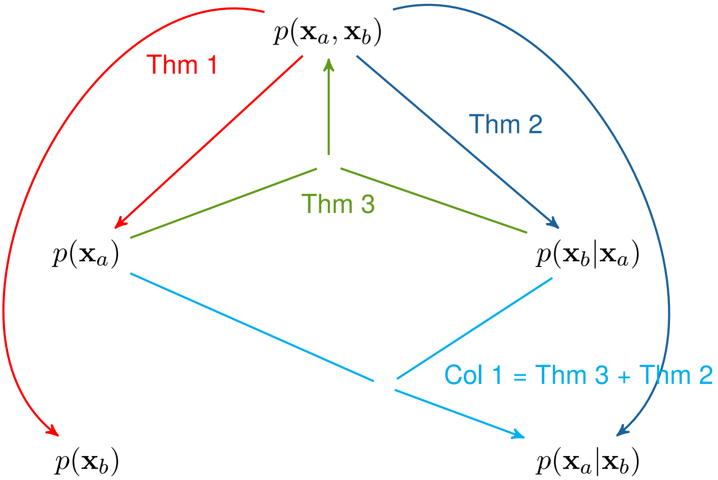
$\beta = \sigma^{-2}$ is called the precision

The probabilistic model is given by:

$$p(\mathbf{y} \mid \mathbf{w}) = \mathcal{N}\left(\mathbf{y}; \mathbf{X}\mathbf{w},\, \beta^{-1}\mathbf{I}_N\right), \quad \text{likelihood}$$
$$p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w}; \mathbf{m}_0,\, \mathbf{S}_0\right), \quad \text{prior distribution}$$

Task: Compute the posterior distribution: $p(\mathbf{w} \mid \mathbf{y})$.

Partitioned Gaussian – Theorems



Affine transformation of multivar. Gauss

By combining **Theorem 3** and **Theorem 2** we get

Corollary 1 (Affine transformation – conditional)

Assume that $\mathbf{x}_a$*, as well as* $\mathbf{x}_b$ *conditioned on* $\mathbf{x}_a$*, are Gaussian distributed according to*

$$\begin{aligned} p(\mathbf{x}_a) &= \mathcal{N}\left(\mathbf{x}_a; \boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a\right), \\ p(\mathbf{x}_b \,|\, \mathbf{x}_a) &= \mathcal{N}\left(\mathbf{x}_b; \mathbf{M}\mathbf{x}_a + \mathbf{b}, \boldsymbol{\Sigma}_{b\,|\,a}\right). \end{aligned}$$

Then the conditional distribution of $\mathbf{x}_a$ *given* $\mathbf{x}_b$ *is*

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

with

$$\begin{aligned} \boldsymbol{\mu}_{a\,|\,b} &= \boldsymbol{\Sigma}_{a\,|\,b}\left(\boldsymbol{\Sigma}_a^{-1}\boldsymbol{\mu}_a + \mathbf{M}^\mathsf{T}\boldsymbol{\Sigma}_{b\,|\,a}^{-1}(\mathbf{x}_b - \mathbf{b})\right), \\ \boldsymbol{\Sigma}_{a\,|\,b} &= \left(\boldsymbol{\Sigma}_a^{-1} + \mathbf{M}^\mathsf{T}\boldsymbol{\Sigma}_{b\,|\,a}^{-1}\mathbf{M}\right)^{-1}. \end{aligned}$$

Bayesian linear regression

The probabilistic model is given by:

$$p(\mathbf{y} \mid \mathbf{w}) = \mathcal{N}\left(\mathbf{y}; \mathbf{X}\mathbf{w}, \beta^{-1}\mathbf{I}_N\right), \quad \text{likelihood}$$

$$p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w}; \mathbf{m}_0, \mathbf{S}_0\right), \quad \text{prior distribution}$$

Task: Compute the posterior distribution: $p(\mathbf{w} \mid \mathbf{y})$.

Solution: Identify

$$\mathbf{x}_a = \mathbf{w}, \qquad \mathbf{x}_b = \mathbf{y},$$

With **Corollary 1** we get the posterior distribution

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

where

$$\mathbf{m}_N = \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\mathbf{X}^\mathsf{T}\mathbf{y}),$$

$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta\mathbf{X}^\mathsf{T}\mathbf{X},$$

UPPSALA UNIVERSITET

ex) Bayesian linear regression

Consider the problem of fitting a straight line to noisy measurements.

Let the model be ($y_n \in \mathbb{R}, x_n \in \mathbb{R}$)

$$y_n = \underbrace{w_0 + w_1 x_n}_{\mathbf{w}^\mathsf{T}\mathbf{x}_n} + \varepsilon_n, \qquad n = 1, \dots, N.$$

where

$$\mathbf{x}_n = \begin{bmatrix} 1 \\ x_n \end{bmatrix}, \qquad \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

$$\varepsilon_n \sim \mathcal{N}(0, \beta^{-1}), \qquad \beta = 5^2.$$

Furthermore, let the prior be

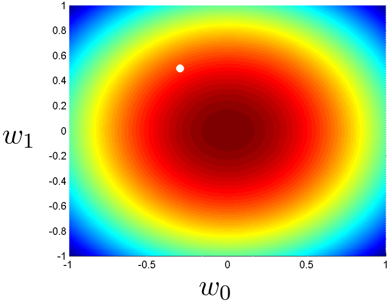
$$p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w} \,|\, \begin{pmatrix} 0 & 0 \end{pmatrix}^\mathsf{T}, \alpha^{-1}\mathbf{I}_2\right),$$

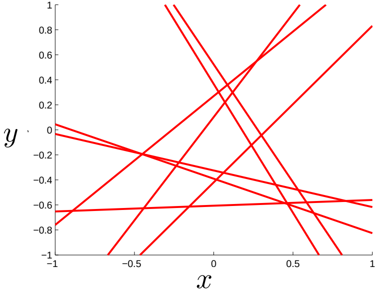
where

$$\alpha = 2.$$

ex) Bayesian linear regression

Plot of the situation before any data arrives.





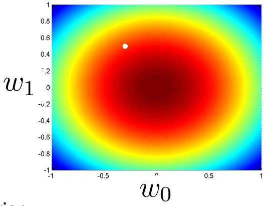
Prior,

$$p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w} \mid \begin{pmatrix}0 & 0\end{pmatrix}^\mathsf{T}, \frac{1}{2}\mathbf{I}_2\right)$$

Example of a few realizations from the prior.

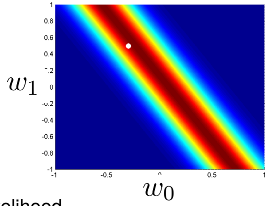
ex) Bayesian linear regression

Plot of the situation after **one** measurement has arrived.



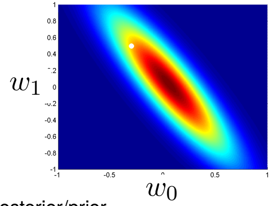
Prior

$$p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w} \,|\, \mathbf{m}_0, \mathbf{S}_0\right)$$



Likelihood

$$p(y_1 \,|\, \mathbf{w}) = \mathcal{N}(y_1 \,|\, w_0 + w_1 x_1, \beta^{-1})$$

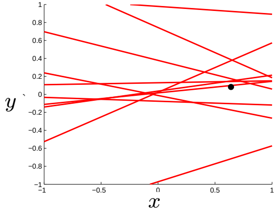


Posterior/prior,

$$p(\mathbf{w} \,|\, y_1) = \mathcal{N}\left(\mathbf{w} \,|\, \mathbf{m}_1, \mathbf{S}_1\right),$$

$$\mathbf{m}_1 = \beta \mathbf{S}_1 \mathbf{X}^{\mathsf{T}} y_1,$$

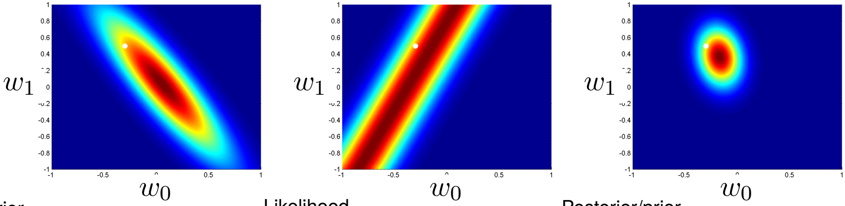
$$\mathbf{S}_1 = (\alpha \mathbf{I}_2 + \beta \mathbf{X}^{\mathsf{T}} \mathbf{X})^{-1}.$$



Example of a few realizations from the posterior and the first measurement (black circle).

ex) Bayesian linear regression

Plot of the situation after **two** measurements have arrived.



Prior

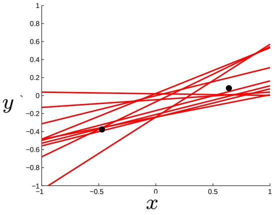
Likelihood

Posterior/prior,

$$p(\color{red}{\mathbf{w}}\,|\,y_1) = \mathcal{N}\left(\color{red}{\mathbf{w}}\,|\,\mathbf{m}_1, \mathbf{S}_1\right)$$

$$p(y_2\,|\,\color{red}{\mathbf{w}}) = \mathcal{N}(y_2\,|\,\color{red}{w_0} + \color{red}{w_1}x_2, \beta^{-1})$$

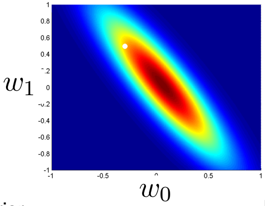
$$p(\color{red}{\mathbf{w}}\,|\,y_2) = \mathcal{N}\left(\color{red}{\mathbf{w}}\,|\,\mathbf{m}_2, \mathbf{S}_2\right),$$
$$\mathbf{m}_2 = \beta \mathbf{S}_2 \mathbf{X}^{\mathsf{T}} \mathbf{y},$$
$$\mathbf{S}_2 = (\alpha \mathbf{I}_2 + \beta \mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}.$$



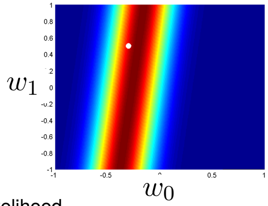
Example of a few realizations from the posterior and the first measurement (black circle).

ex) Bayesian linear regression

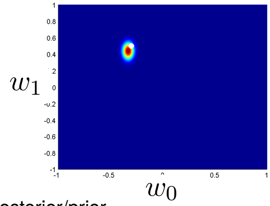
Plot of the situation after **30** measurements have arrived.



Prior



Likelihood



Posterior/prior,

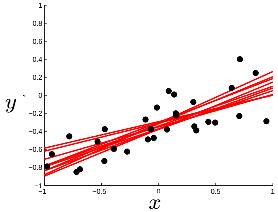
$$p(\mathbf{w} \mid y_2) = \mathcal{N}\left(\mathbf{w} \mid \mathbf{m}_2, \mathbf{S}_2\right)$$

$$p(y_3 \mid \mathbf{w}) = \mathcal{N}(y_3 \mid w_0 + w_1 x_3, \beta^{-1})$$

$$p(\mathbf{w} \mid y_3) = \mathcal{N}\left(\mathbf{w} \mid \mathbf{m}_3, \mathbf{S}_3\right),$$

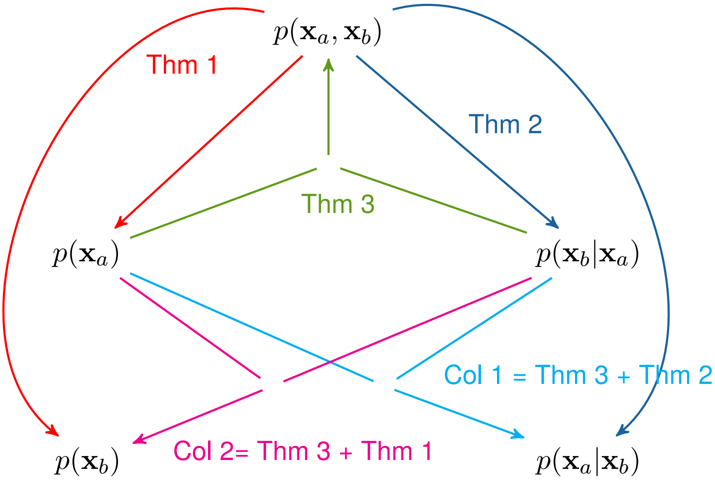
$$\mathbf{m}_3 = \beta \mathbf{S}_3 \mathbf{X}^\mathsf{T} \mathbf{y},$$

$$\mathbf{S}_3 = (\alpha \mathbf{I}_2 + \beta \mathbf{X}^\mathsf{T} \mathbf{X})^{-1}.$$



Example of a few realizations from the posterior and the first measurement (black circle).

Partitioned Gaussian – Theorems



Affine transformation of multivar. Gauss

By combining **Theorem 3** and **Theorem 1** we get

Corollary 2 (Affine transformation – Marginalization)

Assume that $\mathbf{x}_a$, as well as $\mathbf{x}_b$ conditioned on $\mathbf{x}_a$, are Gaussian distributed according to

$$p(\mathbf{x}_a) = \mathcal{N}\left(\mathbf{x}_a; \boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a\right),$$
$$p(\mathbf{x}_b \,|\, \mathbf{x}_a) = \mathcal{N}\left(\mathbf{x}_b; \mathbf{M}\mathbf{x}_a + \mathbf{b}, \boldsymbol{\Sigma}_{b\,|\,a}\right).$$

Then the marginal distribution of $\mathbf{x}_b$ is then given by

$$p(\mathbf{x}_b) = \mathcal{N}\left(\mathbf{x}_b; \boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b\right),$$

where

$$\boldsymbol{\mu}_b = \mathbf{M}\boldsymbol{\mu}_a + \mathbf{b},$$
$$\boldsymbol{\Sigma}_b = \boldsymbol{\Sigma}_{b\,|\,a} + \mathbf{M}\boldsymbol{\Sigma}_a\mathbf{M}^\mathsf{T}.$$

Predictive distribution

For a new data point $(y_*, \mathbf{x}_*)$, we have:

$$p(y_*|\mathbf{w}) = \mathcal{N}\left(y_*; \mathbf{x}_*^\mathsf{T}\mathbf{w}, \beta^{-1}\right), \quad \text{likelihood}$$

$$p(\mathbf{w}|\mathbf{y}) = \mathcal{N}\left(\mathbf{w}; \mathbf{m}_N, \mathbf{S}_N\right) \quad \text{posterior}$$

Identify

$$\mathbf{x}_a = \mathbf{w}, \qquad \mathbf{x}_b = y_*,$$

With **Corollary 2** we get the predictive distribution

$$p(y_*|\mathbf{y}) = \mathcal{N}\left(y_*; m_*, s_*\right)$$

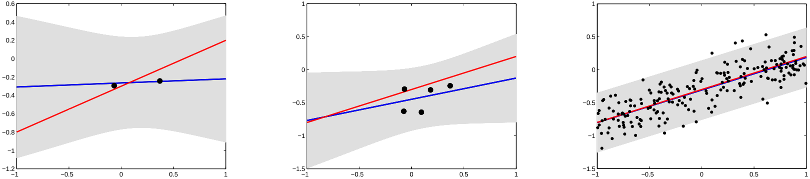
where

$$m_* = \mathbf{x}_*^\mathsf{T}\mathbf{m}_N$$

$$s_* = \beta^{-1} + \mathbf{x}_*^\mathsf{T}\mathbf{S}_N\mathbf{x}_*$$

ex) Predictive distribution

Investigating the predictive distribution for the example above



$N = 2$ observations $N = 5$ observations $N = 200$ observations

- Gray shaded area: One standard deviation of the predictive distribution as function of x^*

$$p(y_*|\mathbf{y}) = \mathcal{N}\left(y_*; \mathbf{x}_*^\mathsf{T}\mathbf{m}_N,\ \beta^{-1} + \mathbf{x}_*^\mathsf{T}\mathbf{S}_N\mathbf{x}_*\right) \quad \text{where} \quad \mathbf{x}_* = \begin{bmatrix} 1 \\ x_* \end{bmatrix}$$

- Blue line: Mean of predictive distribution
- Black circles: Observations
- Red line: true model

Conjugate priors (I/II)

The probabilistic model with unknown $\mathbf{w}$ is given by:

$$p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w}; \mathbf{m}_0, \mathbf{S}_0\right) \qquad \text{prior distribution}$$

$$p(\mathbf{y} \mid \mathbf{w}) = \mathcal{N}\left(\mathbf{y}; \mathbf{X}\mathbf{w}, \beta^{-1}\mathbf{I}_N\right) \qquad \text{likelihood}$$

which gives the posterior

$$p(\mathbf{w}|\mathbf{y}) = \mathcal{N}\left(\mathbf{w}; \mathbf{m}_N, \mathbf{S}_N\right) \qquad \text{posterior}$$

Note that, using a Gaussian prior gives a Gaussian posterior

$$\overbrace{\underbrace{p(\mathbf{w}|\mathbf{y})}_{\text{Gaussian}}}^{\text{posterior}} \propto \overbrace{\underbrace{p(\mathbf{y} \mid \mathbf{w})}_{\text{Gaussian}}}^{\text{likelihood}} \overbrace{\underbrace{p(\mathbf{w})}_{\text{Gaussian}}}^{\text{prior}}$$

Hence, the Gaussian prior is a **conjugate prior** for the Gaussian likelihood unknown $\mathbf{w}$.

***Q**: What if also precision β is unknown?*

Conjugate prior (II/II)

The probabilistic model with unknown $\mathbf{w}$ and β is given by:

$$p(\mathbf{w}, \beta) = \mathcal{N}\left(\mathbf{w}; \mathbf{m}_0,\ \beta^{-1}\mathbf{S}_0\right) \text{Gam}\left(\beta; a_0,\ b_0\right) \qquad \text{prior}$$

$$p(\mathbf{y} \mid \mathbf{w}) = \mathcal{N}\left(\mathbf{y}; \mathbf{X}\mathbf{w},\ \beta^{-1}\mathbf{I}_N\right) \qquad \text{likelihood}$$

which gives the posterior

$$p(\mathbf{w}, \beta|\mathbf{y}) = \mathcal{N}\left(\mathbf{w}; \mathbf{m}_N,\ \beta^{-1}\mathbf{S}_N\right) \text{Gam}\left(\beta; a_N,\ b_N\right) \qquad \text{posterior}$$

Using a Gauss-Gamma prior gives a Gauss-Gamma posterior

$$\underbrace{\overbrace{p(\mathbf{w}, \beta|\mathbf{y})}^{\text{posterior}}}_{\text{Gauss-Gamma}} \propto \underbrace{\overbrace{p(\mathbf{y} \mid \mathbf{w}, \beta)}^{\text{likelihood}}}_{\text{Gauss}} \underbrace{\overbrace{p(\mathbf{w}, \beta)}^{\text{prior}}}_{\text{Gauss-Gamma}}$$

Hence, the Gauss-Gamma prior is a **conjugate prior** for the Gaussian likelihood with unknown $\mathbf{w}$ *and* unknown precision β.

See further in Exercise 2.11

Non-conjugate priors

In the first two lectures we could solve Bayes' theorem analytically since we used **conjugate priors**

$$p(\mathbf{w} \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \mathbf{w})p(\mathbf{w})}{p(\mathbf{y})}$$

However, often you have a **personal belief incompatible with the conjugacy**.

For example:

- Likelihood with heavy tails
- Multi modal distributions

We have to use **approximative inference** methods. In this course will discuss two methods

- **Monte carlo** (lecture 4)
- **Variational inference** (lecture 6)

A few concepts to summarize lecture 2

Prior distribution: $p(\mathbf{w})$ The representation we have about the unknown parameters $\mathbf{w}$ before we have considered any data.

Likelihood distribution: $p(\mathbf{y} \mid \mathbf{w})$ describes how likely the measurements are for a particular parameter value.

Posterior distribution: $p(\mathbf{w} \mid \mathbf{y})$ summarize our knowledge about the parameters $\mathbf{w}$ based on the information we have from the measurements $\mathbf{y}$ and the model.

Predictive distribution: $p(y_\star \mid \mathbf{y})$ the distribution of unobserved observations $y_\star$ conditional on the observed data $\mathbf{y}$.