Summary: MAP Estimation

To summarize, we have the following optimization problem:

$$J(w) = \frac{1}{2} \sum_{n=1}^{N} (w^T \phi(x_n) - t_n)^2 + \frac{\lambda}{2} w^T w$$

$$\phi(x_n) \in \mathbb{R}^M$$

The same in vector notation:

$$J(w) = \frac{1}{2} w^T \Phi^T \Phi w - w \Phi^T t + \frac{1}{2} t^T t + \frac{\lambda}{2} w^T w \quad t \in \mathbb{R}^N$$

$$\Phi = \begin{pmatrix}
\phi_0(x_1) & \phi_1(x_1) & \ldots & \phi_{M-1}(x_1) \\
\phi_0(x_2) & \phi_1(x_2) & \ldots & \phi_{M-1}(x_2) \\
\vdots & \vdots & \ddots & \vdots \\
\phi_0(x_N) & \phi_1(x_N) & \ldots & \phi_{M-1}(x_N)
\end{pmatrix} \in \mathbb{R}^{N \times M}$$

“Feature Matrix”
Summary: MAP Estimation

To summarize, we have the following optimization problem:

$$J(w) = \frac{1}{2} \sum_{n=1}^{N} (w^T \phi(x_n) - t_n)^2 + \frac{\lambda}{2} w^T w$$

$$\phi(x_n) \in \mathbb{R}^M$$

The same in vector notation:

$$J(w) = \frac{1}{2} w^T \Phi^T \Phi w - w \Phi^T t + \frac{1}{2} t^T t + \frac{\lambda}{2} w^T w$$

t \in \mathbb{R}^N$$

And the solution is

$$w^* = (\lambda I_M + \Phi^T \Phi)^{-1} \Phi^T t$$

Identity matrix of size $M$ by $M$
**MLE And MAP**

- The benefit of MAP over MLE is that prediction is less sensitive to **overfitting**, i.e. even if there is only little data the model predicts well.
- This is achieved by using **prior information**, i.e. model assumptions that are not based on any observations (= data)
- But: both methods only give the **most likely** model, there is no notion of **uncertainty** yet

Idea 1: Find a **distribution** over model parameters ("parameter posterior")
MLE And MAP

• The benefit of MAP over MLE is that prediction is less sensitive to **overfitting**, i.e. even if there is only little data the model predicts well.

• This is achieved by using **prior information**, i.e. model assumptions that are not based on any observations (= data)

• But: both methods only give the **most likely** model, there is no notion of **uncertainty** yet

Idea 1: Find a **distribution** over model parameters

Idea 2: Use that distribution to estimate **prediction uncertainty** (“predictive distribution”)
When Bayes Meets Gauß

**Theorem:** If we are given this:

1. \( p(x) = \mathcal{N}(x \mid \mu, \Sigma_1) \)

2. \( p(y \mid x) = \mathcal{N}(y \mid Ax + b, \Sigma_2) \)

Then it follows (properties of Gaussians):

3. \( p(y) = \mathcal{N}(y \mid A\mu + b, \Sigma_2 + A\Sigma_1 A^T) \)

4. \( p(x \mid y) = \mathcal{N}(x \mid \Sigma(A^T\Sigma_2^{-1}(y - b) + \Sigma_1^{-1}\mu), \Sigma) \)

where

\[
\Sigma = (\Sigma_1^{-1} + A^T\Sigma_2^{-1}A)^{-1}
\]

“Linear Gaussian Model”

See Bishop’s book for the proof!
When Bayes Meets Gauß

Thus: When using the Bayesian approach, we can do even more than MLE and MAP by using these formulae.

This means:

If the prior and the likelihood are Gaussian then the posterior and the normalizer are also Gaussian and we can compute them in closed form.

This gives us a natural way to compute uncertainty!
The Posterior Distribution

Remember Bayes Rule:

\[ p(w \mid x, t) \propto p(t \mid w, x) p(w) \]

Posterior \quad Likelihood \quad Prior

With our theorem, we can compute the posterior in **closed form** (and not just its maximum)!

The posterior is also a Gaussian and its **mean** is the MAP solution.
The Posterior Distribution

We have
\[ p(w) = \mathcal{N}(w; 0, \sigma_2^2 I_M) \]
and
\[ p(t \mid w, x) = \mathcal{N}(t; \Phi w, \sigma_1^2 I_N) \]

From this and IV. we get the posterior covariance:
\[ \Sigma = \left( \sigma_2^{-2} I_M + \sigma_1^{-2} \Phi^T \Phi \right)^{-1} \]
\[ = \sigma_1^2 \left( \frac{\sigma_1^2}{\sigma_2^2} I_M + \Phi^T \Phi \right)^{-1} \]

and the mean:
\[ \mu = \sigma_1^{-2} \Sigma \Phi^T t \]

So the entire posterior distribution is
\[ p(w \mid t, x) = \mathcal{N}(w; \mu, \Sigma) \]

Note: So far we only used the training data! (x, t)
The Predictive Distribution

We obtain the **predictive distribution** by integrating over all possible model parameters ("inference"): 

\[
p(t^* | x^*, t, x) = \int p(t^* | x^*, w)p(w | x, t)dw
\]

This distribution can be computed in closed form, because both terms on the RHS are Gaussian. From above we have 

\[
p(w | t, x) = \mathcal{N}(w; \mu, \Sigma)
\]

where 

\[
\mu = \sigma_1^{-2} \sum \Phi^T t
\]

and 

\[
\Sigma = \sigma_1^2 \left( \frac{\sigma_1^2}{\sigma_2^2} I_M + \Phi^T \Phi \right)^{-1}
\]
The Predictive Distribution

We obtain the **predictive distribution** by integrating over all possible model parameters ("inference"):

\[
p(t^* | x^*, t, x) = \int p(t^* | x^*, w)p(w | x, t)dw
\]

This distribution can be computed in closed form, because both terms on the RHS are Gaussian.

From above we have:

\[
p(w | t, x) = \mathcal{N}(w; \mu, \Sigma)
\]

where

\[
\mu = \sigma_1^{-2}\Sigma\Phi^T t
\]

and

\[
\Sigma = \sigma_1^2\left(\frac{\sigma_1^2}{\sigma_2^2}I_M + \Phi^T\Phi\right)^{-1}
\]

\[
\Rightarrow \mu = (\lambda I_M + \Phi^T\Phi)^{-1}\Phi^T t
\]

**MAP solution**
The Predictive Distribution

Using formula III. from above (linear Gaussian),

\[ p(t^* \mid x^*, t, x) = \int p(t^* \mid x^*, w)p(w \mid x, t)dw \]

\[ = \int \mathcal{N}(t^*; \phi(x)^T w, \sigma)\mathcal{N}(w; \mu, \Sigma)dw \]

\[ = \mathcal{N}(t^*; \phi(x)^T \mu, \sigma_N^2(x^*)) \]

where

\[ \sigma_N^2(x) = \sigma^2 + \phi(x)^T \Sigma \phi(x) \]
The Predictive Distribution (2)

- Example: Sinusoidal data, 9 Gaussian basis functions, 1 data point

From: C.M. Bishop
Predictive Distribution (3)

- Example: Sinusoidal data, 9 Gaussian basis functions, 2 data points

The predictive distribution

Some samples from the posterior

From: C.M. Bishop
Predictive Distribution (4)

- Example: Sinusoidal data, 9 Gaussian basis functions, 4 data points

From: C.M. Bishop
Predictive Distribution (5)

- Example: Sinusoidal data, 9 Gaussian basis functions, 25 data points

From: C.M. Bishop

The predictive distribution

Some samples from the posterior
Summary

• Regression can be expressed as a **least-squares** problem
• To avoid overfitting, we need to introduce a **regularisation term** with an additional parameter $\lambda$
• Regression **without** regularisation is equivalent to Maximum Likelihood Estimation
• Regression **with** regularisation is Maximum A-Posteriori
• When using Gaussian priors (and Gaussian noise), all computations can be done **analytically**
• This gives a closed form of the **parameter posterior** and the **predictive distribution**
3. Probabilistic Graphical Models
Directed Models
The Bayes Filter (Rep.)

\[
\text{Bel}(x_t) = p(x_t \mid u_1, z_1, \ldots, u_t, z_t)
\]

(Bayes)
\[
= \eta \ p(z_t \mid x_t, u_1, z_1, \ldots, u_t) p(x_t \mid u_1, z_1, \ldots, u_t)
\]

(Markov)
\[
= \eta \ p(z_t \mid x_t) p(x_t \mid u_1, z_1, \ldots, u_t)
\]

(Tot. prob.)
\[
= \eta \ p(z_t \mid x_t) \int p(x_t \mid u_1, z_1, \ldots, u_t, x_{t-1}) \ p(x_{t-1} \mid u_1, z_1, \ldots, u_t) \, dx_{t-1}
\]

(Markov)
\[
= \eta \ p(z_t \mid x_t) \int p(x_t \mid u_t, x_{t-1}) p(x_{t-1} \mid u_1, z_1, \ldots, u_t) \, dx_{t-1}
\]

(Markov)
\[
= \eta \ p(z_t \mid x_t) \int p(x_t \mid u_t, x_{t-1}) p(x_{t-1} \mid u_1, z_1, \ldots, z_{t-1}) \, dx_{t-1}
\]

\[
= \eta \ p(z_t \mid x_t) \int p(x_t \mid u_t, x_{t-1}) \text{Bel}(x_{t-1}) \, dx_{t-1}
\]
Graphical Representation (Rep.)

We can describe the overall process using a *Dynamic Bayes Network*:

- This incorporates the following Markov assumptions:
  
  $p(z_t \mid x_{0:t}, u_{1:t}, z_{1:t}) = p(z_t \mid x_t)$ (measurement)
  
  $p(x_t \mid x_{0:t-1}, u_{1:t}, z_{1:t}) = p(x_t \mid x_{t-1}, u_t)$ (state)
Definition

A Probabilistic Graphical Model is a diagrammatic representation of a probability distribution.

- In a Graphical Model, random variables are represented as **nodes**, and statistical dependencies are represented using **edges** between the nodes.
- The resulting graph can have the following properties:
  - Cyclic / acyclic
  - Directed / undirected
- The simplest graphs are Directed Acyclic Graphs (DAG).
Simple Example

- Given: 3 random variables $a$, $b$, and $c$
- Joint prob: $p(a, b, c) = p(c \mid a, b)p(a, b) = p(c \mid a, b)p(b \mid a)p(a)$

A Graphical Model based on a DAG is called a **Bayesian Network**
Simple Example

- In general: \( K \) random variables \( x_1, x_2, \ldots, x_K \)
- Joint prob:
  \[
p(x_1, \ldots, x_K) = p(x_K|x_1, \ldots, x_{K-1}) \ldots p(x_2|x_1)p(x_1)
\]
- This leads to a fully connected graph.
- Note: The ordering of the nodes in such a fully connected graph is **arbitrary**. They all represent the joint probability distribution:

\[
p(a, b, c) = p(a|b, c)p(b|c)p(c)
\]
\[
p(a, b, c) = p(b|a, c)p(a|c)p(c)
\]
\[
\vdots
\]
Bayesian Networks

Statistical independence can be represented by the absence of edges. This makes the computation efficient.

\[
p(x_1, \ldots, x_7) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3) \\
p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5)
\]

**Intuitively:** only \( x_1 \) and \( x_3 \) have an influence on \( x_5 \)
Bayesian Networks

We can now define a one-to-one mapping from graphical models to probabilistic formulations:

General Factorization:

\[ p(x) = \prod_{k=1}^{K} p(x_k | p_{a_k}) \]

where

\[ p_{a_k} \triangleq \text{ancestors of } p_k \]

and

\[ p(x) = p(x_1, \ldots, x_K) \]
Elements of Graphical Models

In case of a series of random variables with equal dependencies, we can subsume them using a **plate**:

\[
p(t, w) = p(w) \prod_{n=1}^{N} p(t_n | w)
\]
Elements of Graphical Models (2)

We distinguish between input variables and explicit hyper-parameters:

\[
p(t, w | x, \alpha, \sigma^2) = p(w | \alpha) \prod_{n=1}^{N} p(t_n | w, x_n, \sigma^2).
\]
Elements of Graphical Models (3)

We distinguish between **observed** variables and **hidden** variables:

\[
p(w|t) \propto p(w) \prod_{n=1}^{N} p(t_n|w)
\]

(deterministic parameters omitted in formula)
Example: Regression as a Graphical Model

Aim: Find a general expression to compute the predictive distribution: \( p(\hat{t} \mid \hat{x}, x, t) \)

This expression should

- model all conditional independencies
- explicitly incorporate all parameters (also the deterministic ones)
Example: Regression as a Graphical Model

Aim: Find a general expression to compute the predictive distribution: \( p(\hat{t} \mid \hat{x}, x, t) \)

This expression should

- model all conditional independencies
- explicitly incorporate all parameters (also the deterministic ones)

\[
p(\hat{t} \mid \hat{x}, x, t, \alpha, \sigma^2) = \int p(\hat{t}, w \mid \hat{x}, x, t, \alpha, \sigma^2) dw
\]

\[
= \int \frac{p(\hat{t}, w, t \mid \hat{x}, x, \alpha, \sigma^2)}{p(t \mid \hat{x}, x, \alpha, \sigma^2)} dw \propto \int p(\hat{t}, w, t \mid \hat{x}, x, \alpha, \sigma^2) dw
\]
Regression as a Graphical Model

Regression: Prediction of a new target value $\hat{t}$

$$p(\hat{t}, t, w \mid \hat{x}, x, \alpha, \sigma^2) = \prod_{n=1}^{N} p(t_n \mid x_n, w, \sigma^2) p(w \mid \alpha) p(\hat{t} \mid \hat{x}, w, \sigma^2)$$

Here: conditioning on all deterministic parameters

Using this, we can obtain the predictive distribution:

$$p(t \mid \hat{x}, x, t, \alpha, \sigma^2) \propto \int p(\hat{t}, t, w \mid \hat{x}, x, \alpha, \sigma^2) dw$$

Notation:

$\hat{t} = t^*$
Example: Discrete Variables

- Two dependent variables: $K^2 - 1$ parameters

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>$p(x_1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td>2</td>
<td>0.8</td>
</tr>
</tbody>
</table>

| $x_1$ | $x_2$ | $p(x_2 | x_1)$ |
|-------|-------|---------------|
| 1     | 1     | 0.25          |
| 1     | 2     | 0.75          |
| 2     | 1     | 0.1           |
| 2     | 2     | 0.9           |

Here: $K = 2$

$K - 1 + K(K - 1) = K^2 - 1$

- Independent joint distribution: $2(K - 1)$ parameters

$K - 1 + K - 1 = 2(K - 1)$
Discrete Variables: General Case

In a general joint distribution with $M$ variables we need to store $K^M - 1$ parameters.

If the distribution can be described by this graph:

```
  x_1  x_2  ...  x_M
```

then we have only $K - 1 + (M - 1) K (K - 1)$ parameters.

This graph is called a *Markov chain* with $M$ nodes. The number of parameters grows only linearly with the number of variables.
Independence (Rep.)

**Definition 1.4:** Two random variables $X$ and $Y$ are **independent** iff:

\[ p(x, y) = p(x)p(y) \]

For independent random variables $X$ and $Y$ we have:

\[
p(x \mid y) = \frac{p(x, y)}{p(y)} = \frac{p(x)p(y)}{p(y)} = p(x)
\]

**Notation:** $x \perp y \mid \emptyset$

Independence does **not** imply conditional independence!
The same is true for the opposite case.
Conditional Independence (Rep.)

**Definition 1.5:** Two random variables $X$ and $Y$ are *conditional independent* given a third random variable $Z$ iff:

$$p(x, y \mid z) = p(x \mid z)p(y \mid z)$$

This is equivalent to:

$$p(x \mid z) = p(x \mid y, z) \quad \text{and} \quad p(y \mid z) = p(y \mid x, z)$$

Notation: $x \perp y \mid z$
Conditional Independence: Example 1

This graph represents the probability distribution:

\[ p(a, b, c) = p(a|c)p(b|c)p(c) \]

Marginalizing out \( c \) on both sides gives

\[ p(a, b) = \sum_c p(a|c)p(b|c)p(c) \]

This is in general not equal to \( p(a)p(b) \).

Thus: \( a \) and \( b \) are not independent: \( a \not\perp\!\!\!\perp b \mid \emptyset \)
Conditional Independence: Example 1

Now, we condition on \( c \) (it is assumed to be known):

\[
p(a, b|c) = \frac{p(a, b, c)}{p(c)} = p(a|c)p(b|c)
\]

**Thus:** \( a \) and \( b \) are conditionally independent given \( c \): \( a \perp b | c \)

We say that the node at \( c \) is a **tail-to-tail node** on the path between \( a \) and \( b \)
Conditional Independence: Example 2

This graph represents the distribution:

\[ p(a, b, c) = p(a)p(c|a)p(b|c) \]

Again, we marginalize over \( c \):

\[
p(a, b) = p(a) \sum_c p(c|a)p(b|c) = p(a) \sum_c p(c|a)p(b|c, a)
\]

\[
= p(a) \sum_c \frac{p(c, a)p(b, c, a)}{p(a)p(c, a)} = p(a) \sum_c p(b, c | a)
\]

\[
= p(a)p(b|a)
\]

And we obtain:

\[ a \notin b \mid \emptyset \]
Conditional Independence: Example 2

As before, now we condition on $c$:

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

And we obtain: $a \perp b | c$

We say that the node at $c$ is a **head-to-tail node** on the path between $a$ and $b$. 
Conditional Independence: Example 3

Now consider this graph:

\[ p(a, b, c) = p(a)p(b)p(c|a, b) \]

using:

\[ \sum_c p(a, b, c) = p(a)p(b) \sum_c p(c | a, b) \]

we obtain:

\[ p(a, b) = p(a)p(b) \]

And the result is: \( a \perp b \mid \emptyset \)
Conditional Independence: Example 3

Again, we condition on $c$

This results in:

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

We say that the node at $c$ is a head-to-head node on the path between $a$ and $b$. 
To Summarize

- When does the graph represent (conditional) independence?

**Tail-to-tail case:** if we condition on the tail-to-tail node

**Head-to-tail case:** if we cond. on the head-to-tail node

**Head-to-head case:** if we do **not** condition on the head-to-head node (and neither on any of its descendants)

In general, this leads to the notion of **D-separation** for directed graphical models.
D-Separation

Say: A, B, and C are non-intersecting subsets of nodes in a directed graph.
A path from A to B is blocked by C if it contains a node such that either

a) the arrows on the path meet either head-to-tail or tail-to-tail at the node, and the node is in the set C, or

b) the arrows meet head-to-head at the node, and neither the node, nor any of its descendants, are in the set C.

If all paths from A to B are blocked, A is said to be d-separated from B by C.

Notation: $dsep(A, B|C)$
D-Separation

Say: A, B, and C are non-intersecting subsets of nodes in a directed graph.

- A path contains a node such that either
  a) the arrows meet head-to-tail or tail-to-tail at the node, and the node is in the set C, or
  b) the arrows meet head-to-head, and neither the node nor any of its descendants are in the set C.

- If all paths from A to B are blocked, A is said to be d-separated from B by C.

Notation: \( \text{dsep}(A, B|C) \)
D-Separation: Example

We condition on a descendant of e, i.e. it does not block the path from a to b.

\[ \neg \text{dsep}(a, b | c) \]

We condition on a tail-to-tail node on the only path from a to b, i.e. f blocks the path.

\[ \text{dsep}(a, b | f) \]