Bayesian Inference
(I: fitting parameterized functions to data)
Task

Given: $N$ measurements.

Goal: develop and compare models to account for the measured data.

This is a central task in science!!!
Example: Curve fitting

- Find interpolant through data.
- Criteria: explain data and predict interpolate well to new data (predict/generalize well).
- Solution: 1) Assume a parameterized model is true, find the best fit parameters. 2) Compare different models.
- Questions: Best fit in what sense? How do we compare models?
Two levels of inference

1. Fix model (=function class). Fit parameters. Find the most probable parameters, given the data! Do this for each model.

2. Rank different models by the evidence we have for the model from the data.
1. Function fitting

Given a function class $F$ with parameterized functions $f(x,w)$, where $w$ are the parameters, find the function $f(x,w^*)$ with the most likely parameters $w^*$, given the data $D$.

“Most likely”? That’s a probabilistic notion.
Probability distributions

- Random variable: $X$ assumes values in some set $S$.
- Probability of an outcome having the value $x$: $p(X=x)$. Bounded by $0 \leq p(X = x) \leq 1$.
- Normalization: All probabilities have to sum up to 1:
  $$\sum_{x \in S} p(X = x) = 1$$
- Alternative (shorthand) notation: $p(x)$; $\sum_x p(x) = 1$.
- Example: Fair coin.
  $x = \{0,1\}$; $p(x) = 0.5$; $p(x=0) + p(x=1) = 0.5 + 0.5 = 1$. 
Joint and Conditional

Joint distribution of two random variables \(x\) and \(y\): 
\[p(x,y)\].

Conditional distribution of \(x\), conditioned on the knowledge of the value of \(y\): \(p(x|y)\).
(Functions of \(x\). Labeled by the value of \(y\).)

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

Example: \(x = \{\text{not going surfing, going surfing}\} = \{0,1\}\); 
\(y = \{\text{sunny, rainy}\} = \{s,r\}\).
\[p(\text{surfing} ; \text{sun} ) = p( \text{surfing} | \text{sun} ) p( \text{sun} )\]
\[p(x=1,y=s) = p(x=1|y=s)p(y=s)\]
Bayes' Rule

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

\[ \Rightarrow \quad p(y|x) = p(x|y) \frac{p(y)}{p(x)} \]
Finding the best parameters

Data: $D = \{x_i, y_i\}$

Model class $M$. Parameterized function: $f(x,w)$

Probability of the parameters, given the data (and given that we assume the model is true):
$p(w|D,M) \leftarrow$ This is hard to compute!

Probability (Likelihood) of the data, given the model and the parameters:
$p(D|w,M) \leftarrow$ Easy to compute
Use Bayes’ rule to calculate the posterior $p(w|D,M)$ probability of the parameters (after we have seen the data) from the likelihood $p(D|w,M)$ of the data given the parameters:

\[ p(w|D,M) = \frac{p(D|w,M)p(w|M)}{p(D|M)} \]

- $p(w|M)$: prior prob. of parameters; $p(D|M)$: Evidence for the model $M$.

Posterior = Likelihood * Prior / Evidence

Shorthand notation (M dropped, normalizing constant $p(D)$ ignored):

$p(w|D) \sim p(D|w)p(w)$
Example: Fitting function to noisy data

Assume gaussian noise: \( y = f(x,w) + \text{noise} \)

\[
p(y_i | x_i, w) = \frac{1}{\sqrt{2\pi \sigma}} e^{-\frac{1}{2\sigma^2} (y_i - f(x_i, w))^2}
\]

Likelihood:

\[
p(\{y_i\}|\{x_i\}, w) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi \sigma}} e^{-\frac{1}{2\sigma^2} (y_i - f(x_i, w))^2}
\]

\[
= \frac{1}{Z_D} e^{-\beta E_D}
\]

with: \( E_D = \sum_{i=1}^{N} (y_i - f(x_i, w))^2 \) (mean squared error)
Need a regularizer to control smoothness.

\[ P(f|R, \alpha) = \frac{1}{Z_f(\alpha)} e^{-\alpha E_f(f|R)} \]

Example: Cubic spline interpolation.

\[ E_f(f|R) = \int dx (f''(x))^2 \]

Can be expressed as prior on weights:

\[ P(w|M) = \frac{1}{Z_w(\alpha)} e^{-\alpha E_w} \]
 Altogether: Posterior $\sim$ Likelihood $\ast$ Prior

$$P(w|D) = \frac{1}{Z_M(\alpha, \beta)} e^{-(\alpha E_w + \beta E_D)}$$

Maximizing the (log of the) posterior is the same as minimizing

$$E_M = (\alpha E_w + \beta E_D)$$

Under gaussian noise model, minimizing MSE finds the maximum likelihood parameters. (No regularization)
Homework

- Write a program that lets you fit a polynomial of order $n$ to data
- (1) without and (2) with regularization.