### Contents – lecture 2

1. Summary of lecture 1
2. Linear basis function models
3. Maximum likelihood and least squares
4. Bias variance trade-off
5. Shrinkage methods
   - Ridge regression
   - LASSO
6. Bayesian linear regression
7. Motivation of kernel methods

---

### Summary of lecture 1 (I/III)

Machine learning develop methods allowing computers to improve their performance at certain tasks based on observed data.

The three cornerstones:

1. **Data:** The observed data becomes useful when we have extracted knowledge from it.
2. **Model:** A mathematical model is a compact representation (set of assumptions) of the data that in precise mathematical form captures the key properties of the underlying system.
3. **Inference algorithm:** The inference algorithms allows us to learn, or infer, model specific variables from data for any given model.

The ability to represent and manipulate uncertainty is central.

---

### Summary of lecture 1 (II/III)

The idea underlying **maximum likelihood** is that the parameters $w$ should be chosen such that the available measurements $\{x_i\}_{i=1}^N$ are as likely as possible,

$$\hat{w} = \arg\max_w p(x_{1:N} \mid w).$$

The goal in Bayesian modelling is to compute the posterior distribution $p(w \mid x_{1:N})$. Three steps:

1. Assign priors $p(w)$ to all unknown parameters $w$.
2. Decide on the likelihood $p(x_{1:N} \mid w)$.
3. Determine the posterior distribution $p(w \mid x_{1:N})$,

$$p(w \mid x_{1:N}) = \frac{p(x_{1:N} \mid w)p(w)}{p(x_{1:N})} \propto p(x_{1:N} \mid w)p(w)$$

If the posterior $p(\theta \mid x_{1}, \ldots, x_{N})$ and the prior $p(\theta)$ distributions are of the same functional form they are **conjugate distributions** and the prior is said to be a **conjugate prior** for the likelihood.

*Linear regression is the “working horse” of statistics and (supervised) machine learning.*
Summary of lecture 1 (III/III)

Modeling “heavy tails” using the Student’s t-distribution

\[ \text{St}(x \mid \mu, \lambda, \nu) = \int \mathcal{N}(x \mid \mu, (\eta \lambda)^{-1}) \text{Gam}(\eta \mid \nu/2, \nu/2) \, d\eta \]

\[ = \frac{\Gamma(\nu/2 + 1/2)}{\Gamma(\nu/2)} \left( \frac{\lambda}{\pi \nu} \right)^{1/2} \left( 1 + \frac{\lambda(x - \mu)^2}{\nu} \right)^{-\frac{\nu}{2} - \frac{1}{2}} \]

which according to the first expression can be interpreted as an infinite mixture of Gaussians with the same mean, but different variance.

Commonly used basis functions

In using nonlinear basis functions, \( y(x, w) \) can be a nonlinear function in the input variable \( x \) (still linear in \( w \)).

- Global (in the sense that a small change in \( x \) affects all basis functions) basis function
  1. Polynomial (see illustrative example in Section 1.1) (ex. identity \( \phi(x) = x \))
- Local (in the sense that a small change in \( x \) only affects the nearby basis functions) basis function
  1. Gaussian
  2. Sigmoidal

Linear regression model on matrix form

It is commonly convenient to write the linear regression model

\[ t_n = w^T \phi(x_n) + \varepsilon_n, \quad n = 1, \ldots, N, \]

with

\[ w = (w_0 \ w_1 \ \ldots \ w_{M-1})^T, \]

\[ \phi(x_n) = (\phi_0(x_n) \ \phi_1(x_n) \ \ldots \ \phi_{M-1}(x_n))^T, \quad \phi_0(x_n) = 1, \]

on matrix form

\[ T = \Phi w + E, \]

where

\[ T = \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{pmatrix}, \quad \Phi = \begin{pmatrix} \phi_0(x_1) & \phi_1(x_1) & \ldots & \phi_{M-1}(x_1) \\ \phi_0(x_2) & \phi_1(x_2) & \ldots & \phi_{M-1}(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(x_N) & \phi_1(x_N) & \ldots & \phi_{M-1}(x_N) \end{pmatrix}, \quad E = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{pmatrix} \]

Maximum likelihood and least squares (I/IV)

In our linear regression model,

\[ t_n = w^T \phi(x_n) + \varepsilon_n, \]

assume that \( \varepsilon_n \sim \mathcal{N}(0, \beta^{-1}) \) (i.i.d.). This results in the following likelihood function

\[ p(t_n \mid w, \beta) = \mathcal{N}(w^T \phi(x_n), \beta^{-1}) \]

Note that this is a slight abuse of notation, \( p_{w,\beta}(t_n) \) or \( p(t_n; w, \beta) \) would have been better, since \( w \) and \( \beta \) are both modelled as deterministic parameters in ML.
Maximum likelihood and LS (II/IV)

The available training data consist of \( N \) input variables \( X = \{ x_i \}_{i=1}^{N} \) and the corresponding target variables \( T = \{ t_i \}_{i=1}^{N} \). According to our assumptions, the likelihood function is given by

\[
p(T \mid w, \beta) = \prod_{n=1}^{N} p(t_n \mid w, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n \mid w^T \phi(x_n), \beta^{-1})
\]

which results in the following log-likelihood function

\[
L(w, \beta) \triangleq \ln p(t_1, \ldots, t_N \mid w, \beta) = \sum_{n=1}^{N} \ln \mathcal{N}(t_n \mid w^T \phi(x_n), \beta^{-1})
\]

\[
= \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) - \beta \sum_{n=1}^{N} (t_n - w^T \phi(x_n))^2
\]

Maximizing the log-likelihood function \( L(w, \beta) \) w.r.t. \( \beta \) results in the following estimate for \( \beta \)

\[
\frac{1}{\beta} = \frac{1}{N} \sum_{n=1}^{N} (t_n - \hat{w}^T \phi(x_n))^2
\]

Finally, note that if we are only interested in \( w \), the log-likelihood function is proportional to

\[
\sum_{n=1}^{N} (t_n - w^T \phi(x_n))^2,
\]

which clearly shows that assuming a Gaussian noise model and making use of Maximum Likelihood (ML) corresponds to a Least Squares (LS) problem.

Maximum likelihood and LS (III/IV)

The maximum likelihood problem now amounts to solving

\[
\arg \max_{w, \beta} L(w, \beta)
\]

Setting the derivative \( \partial L \partial w = 2\beta \sum_{n=1}^{N} (t_n - w^T \phi(x_n)) \phi(x_n)^T \) equal to 0 gives the following ML estimate for \( w \)

\[
\hat{w}_{ML} = (\Phi^T \Phi)^{-1} \Phi^T T,
\]

\( \Phi = \begin{pmatrix} \phi_0(x_1) & \phi_1(x_1) & \ldots & \phi_{M-1}(x_1) \\ \phi_0(x_2) & \phi_1(x_2) & \ldots & \phi_{M-1}(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(x_N) & \phi_1(x_N) & \ldots & \phi_{M-1}(x_N) \end{pmatrix} \)

Note that if \( \Phi^T \Phi \) is singular (or close to) we can fix this by adding \( \lambda I \),

\[
\hat{w}_{RR} = (\Phi^T \Phi + \lambda I)^{-1} \Phi^T T.
\]

Interpretation of the Gauss-Markov theorem

The least squares estimator has the smallest mean square error (MSE) of all linear estimators with no bias, BUT there may exist biased estimators with lower MSE.

“the restriction to unbiased estimates is not necessarily a wise one.”

[HTF, page 51]

Two interesting classes of potentially biased estimators, 1. Subset selection methods and 2. Shrinkage methods.

This is intimately connected to the bias-variance trade-off

- We will give a system identification example related to ridge regression to illustrate the bias-variance trade-off.
- See Section 3.2 for a slightly more abstract (but very informative) account of the bias-variance trade-off. (this is a perfect topic for discussion during the exercise sessions!)
Interpretation of RR using the SVD of $\Phi$

By studying the SVD of $\Phi$ it can be shown that ridge regression projects the measurements onto the principal components of $\Phi$ and then shrinks the coefficients of low-variance components more than the coefficients of high-variance components. (See Section 3.4.1 in HTF for details.)

Bias-variance tradeoff – example (I/IV)

(Ex. 2.3 in Henrik Ohlsson’s PhD thesis) Consider a SISO system

$$y_t = \sum_{k=1}^{n} g_k^0 u_{t-k} + e_t,$$

where $u_t$ denotes the input, $y_t$ denotes the output, $e_t$ denotes white noise ($E[e] = 0$ and $E[e_te_s] = \sigma^2 \delta(t-s)$) and $\{g_k^0\}_{k=1}^{n}$ denote the impulse response of the system.

Recall that the impulse response is the output $y_t$ when $u_t = \delta(t)$ is used in (1), which results in

$$y_t = \begin{cases} g_0^0 + e_t & t = 1, \ldots, n, \\ e_t & t > n. \end{cases}$$

Bias-variance tradeoff – example (II/IV)

The task is now to estimate the impulse response using an $n$th order FIR model,

$$y_t = w^T \phi_t + e_t,$$

where

$$\phi_t = (u_{t-1} \ldots u_{t-n})^T, \quad w \in \mathbb{R}^n$$

Let us use Ridge Regression (RR),

$$\hat{w}^{RR} = \arg \min_w \|Y - \Phi w\|_2^2 + \lambda w^T w$$

to find the parameters $w$.

Bias-variance tradeoff – example (III/IV)

Squared bias (gray line)

$$\left( E_{\hat{w}} \left[ \hat{w}^T \phi_s \right] - w_0^T \phi_s \right)^2$$

Variance (dashed line)

$$E_{\hat{w}} \left[ \left( E_{\hat{w}} \left[ \left( \hat{w}^T \phi_s \right) \right] - \hat{w}^T \phi_s \right)^2 \right]$$

MSE (black line)

$$\text{MSE} = (\text{bias})^2 + \text{variance} + \text{noise}$$
Bias-variance tradeoff – example (IV/IV)

“Flexible” models will have a low bias and high variance and more “restricted” models will have high bias and low variance.

The model with the best predictive capabilities is the one which strikes the best tradeoff between bias and variance. Regularization is used to automatically tune the model complexity in a data-driven fashion.

Recent contributions on linear impulse response identification using regularization, see


Regularizing nonlinear dynamical models

Regularization allows us to tune the model complexity.

Nonlinear state space model

\[ x_{t+1} = f_\theta(x_t) + w_t, \]
\[ y_t = g_\theta(x_t) + e_t. \]

We employ a nonparametric alternative to regression based on a tailored Gaussian process (GP).

Lasso

The Lasso was introduced during lecture 1 as the MAP estimate when a Laplacian prior is assigned to the parameters. Alternatively we can motivate the Lasso as the solution to

\[
\min_{\mathbf{w}} \sum_{n=1}^{N} (t_n - \mathbf{w}^T \phi(x_n))^2 \\
\text{s.t.} \quad \sum_{j=0}^{M-1} |w_j| \leq \eta
\]

which using a Lagrange multiplier \( \lambda \) can be formulated

\[
\min_{\mathbf{w}} \sum_{n=1}^{N} (t_n - \mathbf{w}^T \phi(x_n))^2 + \lambda \sum_{j=0}^{M-1} |w_j|
\]

The difference to ridge regression is simply that Lasso make use of the \( \ell_1 \)-norm \( \sum_{j=0}^{M-1} |w_j| \), rather than the \( \ell_2 \)-norm \( \sum_{j=0}^{M-1} w_j^2 \) in shrinking the parameters.

Graphical illustration of Lasso and RR

The circles are contours of the least squares cost function (LS estimate in the middle). The constraint regions are shown in gray \( |w_0| + |w_1| \leq \eta \) (Lasso) and \( w_0^2 + w_1^2 \leq \eta \) (RR). The shape of the constraints motivates why Lasso often leads to sparseness.
Implementing Lasso

The ℓ₁-regularized least squares problem (lasso)

$$\min_w \| T - \Phi w \|^2_2 + \lambda \| w \|_1$$  \hspace{1cm} (2)

YALMIP code solving (2). Download: [http://users.isy.liu.se/johanl/yalmip/](http://users.isy.liu.se/johanl/yalmip/)

```matlab
w=sdpvar(M,1);
opssdpsettings('verbose',0);
solvesdp([],(T-Phi*w)'*(T-Phi*w) + lambda*norm(w,1),ops)
```


```matlab
cvx_begin
variable w(M)
minimize((T-Phi*w)'*(y-Phi*w) + lambda*norm(w,1))
cvx_end
```

A MATLAB package dedicated to ℓ₁-regularized least squares problems is l1_ls. Download: [http://www.stanford.edu/~boyd/l1_ls/](http://www.stanford.edu/~boyd/l1_ls/)

---

Bayesian linear regression – example (I/VI)

Consider the problem of fitting a straight line to noisy measurements. Let the model be \((t_n \in \mathbb{R}, x_n \in \mathbb{R})\)

$$t_n = w_0 + w_1 x_n + \varepsilon_n, \quad n = 1, \ldots, N.$$  \hspace{1cm} (3)

where

$$\varepsilon_n \sim \mathcal{N}(0, 0.2^2), \quad \beta = \frac{1}{0.2^2} = 25.$$  

According to (3), the following identity basis function is used

$$\phi_0(x_n) = 1, \quad \phi_1(x_n) = x_n.$$  

The example lives in two dimensions, allowing us to plot the distributions in illustrating the inference.

---

Bayesian linear regression – example (II/VI)

Let the true values for \(w\) be

$$w^* = (-0.3 \ 0.5)^T,$$

plotted using a filled white circle below.

Generate synthetic measurements by

$$t_n = w_0 + w_1 x_n + \varepsilon_n, \quad n = 1, \ldots, N,$$

where \(x_n \sim \mathcal{U}(-1, 1)\).

Furthermore, let the prior be

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

where

$$\alpha = 2.$$  

---

Bayesian linear regression – example (III/VI)

Plot of the situation before any data arrives.

Example of a few realizations from the prior.
Bayesian linear regression – example (IV/VI)

Plot of the situation after one measurement has arrived.

Likelihood (plotted as a function of $w$)

$p(t_1 \mid w) = \mathcal{N}(t_1 \mid w_0 + w_1 x_1, \beta^{-1})$

Posterior/prior,

$p(w \mid t_1) = \mathcal{N}(w \mid m_1, S_1)$,

$m_1 = \beta S_1 \Phi^T t_1$,

$S_1 = (\alpha I + \beta \Phi^T \Phi)^{-1}$.

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

Bayesian linear regression – example (V/VI)

Plot of the situation after two measurements have arrived.

Likelihood (plotted as a function of $w$)

$p(t_2 \mid w) = \mathcal{N}(t_2 \mid w_0 + w_1 x_2, \beta^{-1})$

Posterior/prior,

$p(w \mid T) = \mathcal{N}(w \mid m_2, S_2)$,

$m_2 = \beta S_2 \Phi^T T$,

$S_2 = (\alpha I + \beta \Phi^T \Phi)^{-1}$.

Example of a few realizations from the posterior and the measurements (black circles).

Empirical Bayes (EB)

Important question: How do we decide on the suitable values for hyperparameters $\eta$?

Idea: Estimate the hyperparameters from the data by selecting them such that they maximize the marginal likelihood function,

$$p(T \mid \eta) = \int p(T \mid w, \eta) p(w \mid \eta) dw,$$

where $\eta$ denotes the hyperparameters to be estimated.

Travels under many names and besides empirical Bayes this is also referred to as type 2 maximum likelihood, generalized maximum likelihood, and evidence approximation.

Empirical Bayes combines the two statistical philosophies; frequentistic ideas are used to estimate the hyperparameters that are then used within the Bayesian inference.
Predictive distribution – example

Investigating the predictive distribution for the example above

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

- True system \(y(x) = -0.3 + 0.5x\) generating the data (red line)
- Mean of the predictive distribution (blue line)
- One standard deviation of the predictive distribution (gray shaded area)
  Note that this is the point-wise predictive standard deviation as a function of \(x\).
- Observations (black circles)

Posterior distribution

Recall that the posterior distribution is given by

\[
p(w | T) = \mathcal{N}(w | m_N, S_N),
\]

where

\[
m_N = \beta S_N \Phi^T T,
\]
\[
S_N = (\alpha I + \beta \Phi^T \Phi)^{-1}.
\]

Let us now investigate the posterior mean solution \(m_N\), which has an interpretation that directly leads to the kernel methods (lecture 5), including popular Gaussian process (GP).

A non-parametric alternative to regression

Let us study the predictive mean (recall our linear regression model \(t = y(x, w) + \varepsilon\) with \(w = m_N\),

\[
y(x, m_N) = m_N^T \phi(x) = \phi(x)^T m_N = \beta \phi(x)^T S_N \Phi^T T
= \sum_{n=1}^{N} \beta \phi(x)^T S_N \phi(x_n) t_n.
\]

Hence, the predictive mean can be written

\[
y(x, m_N) = \sum_{n=1}^{N} k(x, x_n) t_n
\]

where \(k(x, x') = \beta \phi(x)^T S_N \phi(x')\) is called the equivalent kernel.

A non-parametric alternative to regression

The exercise on the previous slide suggests an alternative approach to regression, where we instead of postulating a parametric model \(t = w^T \phi(x) + \varepsilon\) (using a set of basis functions) directly make use of a localized kernel.

General property of kernels

\[
k(x, z) = \psi(x)^T \psi(z).
\]

Teaser: The Gaussian process is one particular construction that provides a non-parametric alternative to regression via the direct use of a kernel.
A few concepts to summarize lecture 2

**Linear regression:** Models the relationship between a continuous target variable $t$ and a possibly nonlinear function $\phi(x)$ of the input variables.

**Hyperparameter:** A parameter of the prior distribution that controls the distribution of the parameters of the model.

**Maximum a Posteriori (MAP):** A point estimate obtained by maximizing the posterior distribution. Corresponds to a mode of the posterior distribution.

**Gauss Markov theorem:** States that in a linear regression model, the best (in the sense of minimum MSE) linear unbiased estimate (BLUE) is given by the least squares estimate.

**Ridge regression:** An $\ell_2$-regularized least squares problem used to solve the linear regression problem resulting in potentially biased estimates. A.k.a. Tikhonov regularization.

**Lasso:** An $\ell_1$-regularized least squares problem used to solve the linear regression problem resulting in potentially biased estimates. The Lasso typically produce sparse estimates.