Machine Learning 1 — Voluntary Midterm Exam

Preliminaries

- How to hand in:
  - write your answers on these exam sheets only;
  - write your immat **but not your name** on *every* page that you hand in. If you fill in your name you may lose 0.3 points;
  - hand in all exam sheets.
- The exam is open book. You may use all the material you want, while obeying the following rules:
  - you are not allowed to consult or communicate with other people, be it in the room or anywhere outside, except for with the examinators;
  - you must always place the screens of your computers and other used digital devices so that the examiners can see what you are doing;
  - failure to comply with these simple rules may lead to 0 points.

  In short, we will be as fair as we can, but expect the same from you in return.
- The exam is limited to 70 minutes.
- This exam consists of 8 pages, 7 sections, 7 problems.

1 Linear Regression

In the Linear Regression setting, we assumed that an observed target z is generated by a noisy observation of the true underlying function $y(\boldsymbol{w}, \boldsymbol{x})$, linear in $\boldsymbol{w}$. The employed noise model was a simple one: Additive Gaussian noise, with mean 0 and fixed variance σ^2.

Let's change the above assumptions in the following way: Instead of additive Gaussian noise we assume *additive Laplacian noise*, with mean 0 and some arbitrary (positive) scale parameter b. Furthermore, assume that you have N observations of the form $(\boldsymbol{x}, z)$. **Write** down the negative log-likelihood for $\boldsymbol{w}$ in this case. (Hint: The density function p of the Laplace distribution is given by $p(x) \propto \exp(-|x - \mu|/b)$ with mean μ and positive scale parameter b).

From the lecture on linear regression (page 10) we copy the case for the Gaussian case:

$$z = y(\boldsymbol{w}, \boldsymbol{x}) + \varepsilon$$

where ε is a noise variable distributed according to a zero-mean Gaussian. On the same page, the liklihood for this case is written down as

$$p(\boldsymbol{z}|\boldsymbol{X}, \boldsymbol{w}, \sigma^2) = \prod^N \mathcal{N}(z_n|\boldsymbol{w}^T\boldsymbol{x_n}, \sigma^2)$$

This gives the *negative* log likelihood function (also page 10)

$$-\ln p(\boldsymbol{z}|\boldsymbol{X}, \boldsymbol{w}, \sigma^2) \propto \frac{1}{\sigma^2} \sum_N \left(z_n - \boldsymbol{w}^T\boldsymbol{x_n}\right)^2$$

Substituting the Gaussian noise from above with a Laplacian noise therefore gives

$$-\ln p(\boldsymbol{z}|\boldsymbol{X}, \boldsymbol{w}, \sigma^2) \propto \frac{1}{b} \sum_N \left|z_n - \boldsymbol{w}^T\boldsymbol{x_n}\right|$$

2 A linear case

The correlation between two random variables X and Y can be measured by considering the linear relationship that may exist within sampled pairs (x_i, y_i), $i = 1 \dots N$. Assume that both X and Y have zero mean. In order to investigate the validity of the linear relationship

$$X = aY$$

we use the following sum of square objective:

$$E(a) = \sum_i (x_i - a y_i)^2.$$

Show that $E(a)$ is minimised by $\hat{a}$ with

$$\hat{a} = \frac{c}{\sigma_y^2}$$

where

$$c = \frac{1}{N} \sum_i x_i y_i, \qquad \sigma_y^2 = \frac{1}{N} \sum_i y_i^2.$$

The first derivative of $E(a)$ with respect to a is

$$\sum_i 2(x_i - a y_i)(-y_i) = \sum_i (-x_i y_i + a y_i^2)$$

Equating this to zero (we are looking for the minimizer):

$$\hat{a} = \frac{\sum_i x_i y_i}{\sum_i y_i^2}$$

Possible addon question here: **What** can you conclude if $E(\hat{a}) = 0$?

3 Degenerate and covaryin'.

The *pdf* of a multivariate Gaussian is given by

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

By only looking at the normalisation factor, argue why a low rank covariance matrix makes the pdf useless to us.

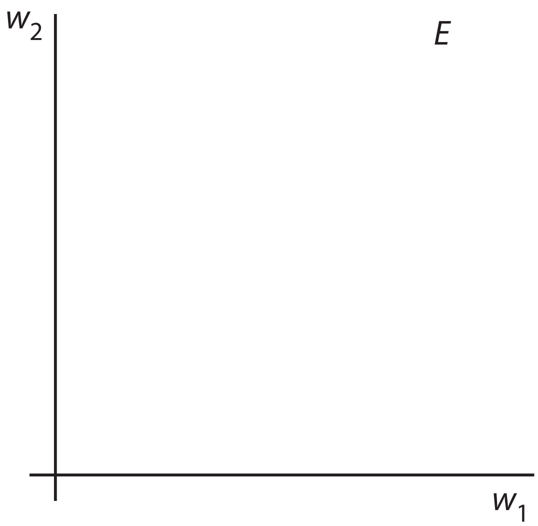
Low rank means that a matrix does not have full rank. The determinant of a matrix that has not full rank is 0.

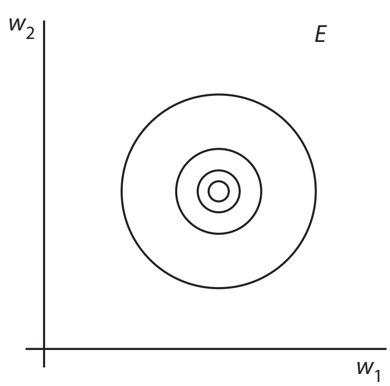
4 Optimisation

Suppose we can write an error value as $E(\boldsymbol{w}) := c - \boldsymbol{g}^T(\boldsymbol{w} - \boldsymbol{w}_0) + \frac{1}{2}(\boldsymbol{w} - \boldsymbol{w}_0)^T H(\boldsymbol{w} - \boldsymbol{w}_0)$ where $\boldsymbol{w}$ is a parameter set, $\boldsymbol{g}$ the gradient. H is the Hessian, the matrix of second-order derivatives of E.

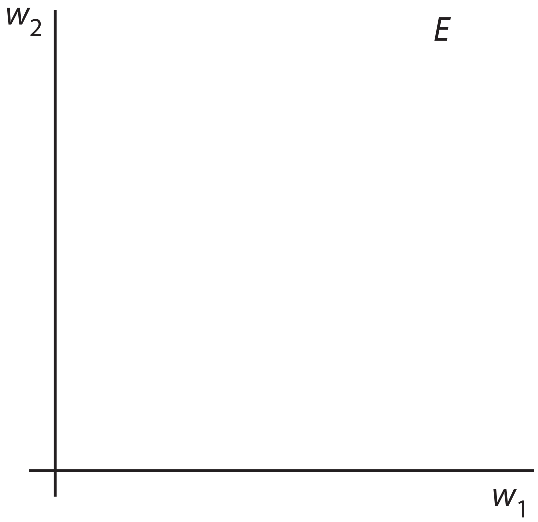
The condition number κ of H can be computed by dividing the value of the largest eigenvalue of H by its smallest eigenvalue.

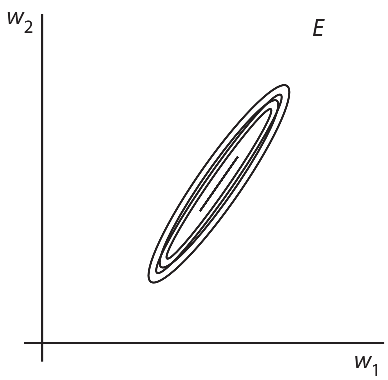
Problem 1. In the case where $\boldsymbol{w}$ is two-dimensional, draw the approximate form of E when $\kappa = 1$ in the below figure using a contour plot.





Problem 2. In the case where $\boldsymbol{w}$ is two-dimensional, draw the approximate form of E when $\kappa = 20$ in the below figure using a contour plot.


w_2
E
w_1


w_2
E
w_1

5 Matrix multiplication

Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be matrices with such dimensions that the matrix product $\boldsymbol{AB}$ is defined. Can the rank of the product $\boldsymbol{AB}$ be larger than the rank of $\boldsymbol{A}$ or $\boldsymbol{B}$? Explain your answer.

The rank of matrix $\boldsymbol{A}$ is the number of linearly independent column vectors of $\boldsymbol{A}$. Every column vector of the matrix product $\boldsymbol{AB}$ is a linear combination of column vectors of the matrix $\boldsymbol{A}$. Thus every column vector of $\boldsymbol{AB}$ is still in the span of the column vectors of $\boldsymbol{A}$. Therefore there cannot be more linearly independent column vectors in $\boldsymbol{AB}$ than in $\boldsymbol{A}$ and the rank of $\boldsymbol{AB}$ can only be smaller or equal to the rank of $\boldsymbol{A}$.

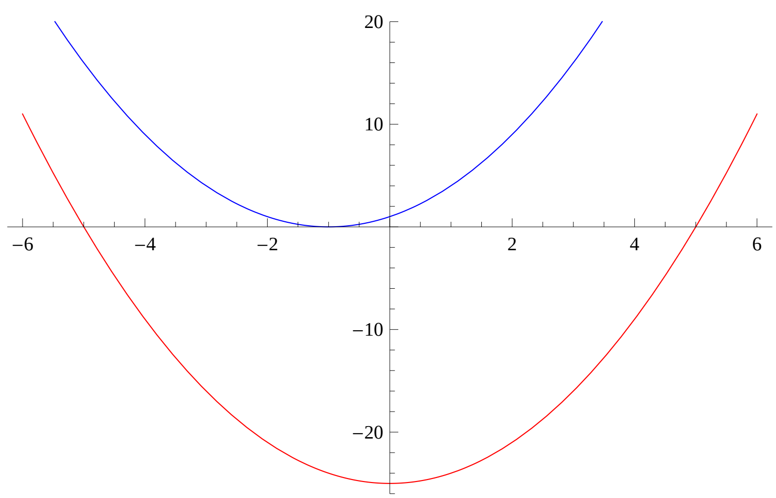
6 Constrained convex optimisation

Find the optimal value p^* and the minimiser x^* of the constrained optimisation problem

$$\begin{aligned} \text{minimise } & f_0(x) = (x+1)^2 \\ \text{subject to } & f_1(x) = (x-5)(x+5) \leq 0\,. \end{aligned}$$

What is the value of the Lagrange multiplier corresponding to the given constraint?

Hint: Before you apply the recipe for solving constrained optimisation problems consider if there is a faster way to obtain the solution in this case (hint hint: draw).



The objective $f_0(x)$ is in blue, the constraint $f_1(x)$ in red. The feasible region is $[-5, 5]$ because $f_1(x)$ has its roots at these points. The unconstrained minimum of $f_0(x)$ is obtained at $x = -1$. This point is in the feasible region, thus it is also the minimum of the constrained optimisation problem. So we have $p^* = f_0(-1) = 0$ and $x^* = -1$.

We have $f_1(x^*) < 0$ because $x^* = -1$ is not a root of $f_1(x)$, thus by complementary slackness the Lagrange multiplier must be zero, $\alpha_1 = 0$. Hence the constraint is inactive.

7 Kernels

Show that for $c \geq 0$ and $d \in \mathbb{N}^+$ the function

$$K(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}^T \boldsymbol{y} + c)^d$$

is a kernel.

The term $\boldsymbol{x}^T \boldsymbol{y}$ is a kernel because it is the scalar product of the input vectors. The constant $c \geq 0$ is a kernel because we can define the feature map $\phi(\boldsymbol{z}) = \sqrt{c}$ and obtain this kernel by calculating the scalar product in feature space $\phi(\boldsymbol{x})^T \phi(\boldsymbol{y}) = \sqrt{c}^2 = c$. Since the constant d is a natural number we can write the exponentiation as the iterated product of the kernel $(\boldsymbol{x}^T \boldsymbol{y} + c)$ with itself. The multiplication of two kernels is a kernel. Hence it follows that $K(\boldsymbol{x}, \boldsymbol{y})$ is a kernel.