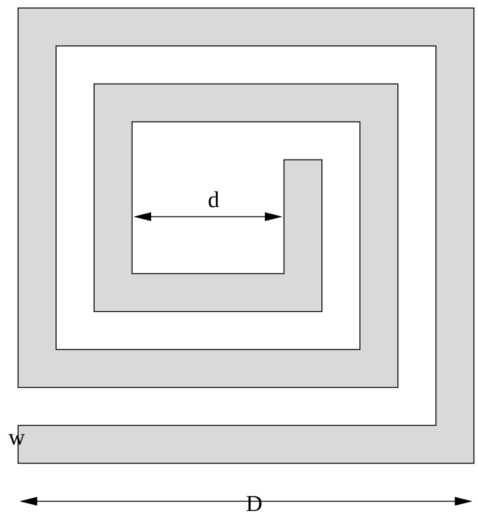
# Homework 2

1. *Approximate inductance formula.* The figure below shows a *planar spiral inductor*, implemented in CMOS, for use in RF circuits.




The inductor is characterized by four key parameters:

- $n$, the number of turns (which is a multiple of 1/4, but that needn't concern us)
- $w$, the width of the wire
- $d$, the inner diameter
- $D$, the outer diameter

The inductance $L$ of such an inductor is a complicated function of the parameters $n$, $w$, $d$, and $D$. The inductance $L$ can be found by solving Maxwell's equations, which takes considerable computer time, or by fabricating the inductor and measuring the inductance. In this problem you will develop a simple approximate inductance model of the form

$$\hat{L} = \alpha n^{\beta_1} w^{\beta_2} d^{\beta_3} D^{\beta_4},$$

where $\alpha, \beta_1, \beta_2, \beta_3, \beta_4 \in \mathbf{R}$ are constants that characterize the approximate model. (since $L$ is positive, we have $\alpha > 0$, but the constants $\beta_2, \ldots, \beta_4$ can be negative.) This simple approximate model, if accurate enough, can be used for design of planar spiral inductors. The file `inductor_data.json` on the course web site contains data `n, w, d, D, L` for 50 inductors. (The data is real, not that it would affect how you solve the problem ...) For inductor $i$, we give parameters $n_i$, $w_i$, $d_i$, and $D_i$ (all in $\mu$m), and also, the inductance $L_i$ (in nH) obtained from measurements. (The data are organized as vectors of length 50. Thus, for example, $w_{13}$ gives the wire width of inductor 13.) Your task, *i.e.*, the problem, is to find $\alpha$, $\beta_1, \ldots, \beta_4$ so that

$$\hat{L}_i = \alpha n_i^{\beta_1} w_i^{\beta_2} d_i^{\beta_3} D_i^{\beta_4} \approx L_i \qquad \text{for } i = 1, \ldots, 50.$$

Your solution must include a clear description of how you found your parameters, as well as their actual numerical values. Note that we have not specified the criterion that you use to judge the approximate model (*i.e.*, the fit between $\hat{L}_i$ and $L_i$); we leave that to your engineering judgment. But be sure to tell us what criterion you use. We define the *percentage error* between $\hat{L}_i$ and $L_i$ as

$$e_i = 100|\hat{L}_i - L_i|/L_i.$$

Find the average percentage error for your model, *i.e.*, $(e_1 + \cdots + e_{50})/50$. (We are only asking you to find the average percentage error for your model; we do not require that your model minimize the average percentage error.) *Hint:* you might find it easier to work with $\log L$.

2. *Fitting with rational functions.* In this problem we consider a function $f : \mathbf{R} \to \mathbf{R}$ of the form

$$f(x) = \frac{a_0 + a_1 x + \cdots + a_m x^m}{1 + b_1 x + \cdots + b_m x^m},$$

where $a_0, \ldots, a_m$, and $b_1, \ldots, b_m$ are parameters, with either $a_m \neq 0$ or $b_m \neq 0$. Such a function is called a *rational function of degree $m$*. We are given data points $x_1, \ldots, x_n \in \mathbf{R}$ and $y_1, \ldots, y_n \in \mathbf{R}$, where $y_i \approx f(x_i)$.

The problem is to find a rational function of degree $m$ that is consistent with this data. In other words, you are to find $a_0, \ldots, a_m$, $b_1, \ldots, b_m$, which satisfy $f(x_i) \approx y_i$. Explain how you will solve this problem, and then carry out your method for several different $m$'s, maybe 3 to 10, on the problem data given in `rational_fit_data.json`. (This contains the data size, `n`, with two vectors, `x` and `y`, that give the values $x_1, \ldots, x_n$, and $y_1, \ldots, y_n$, respectively.)

Which $m$ would you choose? Justify your answer by displaying your approximations with the raw data on a single plot.

3. *Regularized least squares and features.* The following problem will use `U`, and `v` found in `prostate_cancer_data.json`. For the problem below, use the first 70 data as the training set, and the rest as the test set.

    (a) Explain how to formulate the problem of fitting regularized least squares given a matrix $U$ and regularization parameter $\lambda > 0$, where the first feature is the constant feature $x_1 = 1$.

    (b) Standardize all of the features and then fit a model to the data, adding only a constant feature. Sweep your regularization parameter $\lambda$ over the range $[10^{-5}, 10^5]$

and plot the corresponding training and test errors. Choose an appropriate value for $\lambda$, *i.e.*, the largest value that achieves approximately minimum test error. Give the model, and the corresponding test error.

(c) If you look into the data matrix $U$, you'll notice that the last two columns actually take on only a few values. Embed both columns using a one-hot encoding, keeping the rest of the values the same. Now do the same sweep you did in part (b), and plot the corresponding training and testing errors. Compare the final test RMSE of the one-hot encoding vs. the original encoding, with appropriately chosen $\lambda$ for each.

*Hint.* You can make use of the Julia file `to_one_hot.jl`, which contains the function `to_one_hot(u)`. This function takes as input an $n$-vector $u$ whose entries are one of $k$ categories and embeds it using a one-hot embedding into $\mathbf{R}^{n\times k}$.

4. *Limit behaviours of the regularized least squares solutions.* In this problem we consider the optimal solution of the Tykhonov regularized least squares solutions with a full rank matrix $X \in \mathbf{R}^{n\times d}$.

$$\begin{aligned}\theta^* &= \operatorname*{argmin}_{\theta} ||X\theta - y||_2^2 + \lambda||\theta||_2^2 \\ &= \left(X^TX + \lambda I\right)^{-1} X^T y\end{aligned}$$

Suppose $X$ is skinny and full rank, *i.e.*, $\operatorname{rank}(A) = d < n$. Then it is crystal clear that $\theta^* \to 0$ as $\lambda \to \infty$, and $\theta^* \to (X^TX)^{-1}X^Ty$ as $\lambda \to 0$, that is, $\theta^*$ approaches to zero if $\lambda$ is extremely large, and $\theta^*$ approaches to the unregularized least squares solution when $\lambda$ is tiny. No big deal.

Now consider the opposite case when $X$ is fat and full rank, *i.e.*, $\operatorname{rank}(A) = n < d$. The optimal $\theta^*$ is zero for extremely large $\lambda$. The same thing. However an interesting thing happens when $\lambda$ approaches to zero. In this problem, we are going to look at that. First note that $X^TX$, which is the limit of $(X^TX+\lambda I)$ as $\lambda \to 0$, is not invertible in this case, hence the expression $(X^TX)^{-1}X^Ty$ doesn't make sense at all.

(a) Show that the following holds whenever the appearing matrix products and inverses make sense. It is called the *Push-through identity.*

$$A\,(I + BA)^{-1} = (I + AB)^{-1}\,A$$

(b) Find the optimal $\theta^*$ by applying the above to your optimal regularized least squares solution, and taking the limit, $\lambda \to 0$. What is it?

(c) Show that your solution satisfies $X\theta^* = y$.

(d) Show that $||\theta^*||_2 \leq ||\theta||_2$ for any $\theta \in \mathbf{R}^d$.

The solution you found achieves the minimum norm among the infinitely many solutions satisfying $X\theta = y$, hence it is called the *least norm* solution.