To be completed individually or in groups of two people (for groups, please be sure both names and matriculation numbers are clearly included at the top of your submission). Submissions can be handwritten or in LaTeX formatting, but hard-to-read handwritten submissions will not be graded.

Please submit via Ilias. Submissions should be a single PDF document (note that Jupyter notebooks can and should also be downloaded as PDFs, and not submitted as .ipynb files).

Each question will be graded "pass" (full points) or "fail" (no points). We award .5 bonus points for the exam for each theory and practical question solved. You must complete 50% of all exercises to enter the final exam.

1. **EXAMple Question** Consider the Gaussian random variable $\boldsymbol{w} \in \mathbb{R}^F$ with probability density function $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}; \boldsymbol{\mu}, \Sigma)$ where $\boldsymbol{\mu} \in \mathbb{R}^F$ and symmetric positive definite $\Sigma \in \mathbb{R}^{F \times F}$. You have access to data $\boldsymbol{y} \in \mathbb{R}^N$ assumed to be generated from $\boldsymbol{w}$ through a linear map $\Phi \in \mathbb{R}^{F \times N}$ according to the likelihood

$$p(\boldsymbol{y}|\boldsymbol{w}) = \mathcal{N}(\boldsymbol{y}; \Phi^T \boldsymbol{w}, \Lambda),$$

where $\Lambda \in \mathbb{R}^{N \times N}$ is symmetric positive definite. What is:

   (a) the pdf of the *marginal* $p(\boldsymbol{y}) = \int p(\boldsymbol{y}|\boldsymbol{w}) p(\boldsymbol{w}) d\boldsymbol{w}$?

   (b) the pdf of the *posterior* $p(\boldsymbol{w}|\boldsymbol{y})$?

2. **Theory Question** Consider the model defined in the EXAMple Question (see above), for the special case $\Lambda = \sigma^2 I$ with $\sigma^2 \in \mathbb{R}_+$ (that is, iid. observation noise).

   (a) Show that the **maximum likelihood estimator** for $\boldsymbol{w}$ is given by the **ordinary least-squares** estimate

$$\boldsymbol{w}_{ML} = (\Phi\Phi^T)^{-1}\Phi\boldsymbol{y}.$$

   To do so, use the explicit form of the Gaussian pdf to write out $\log p(\boldsymbol{y}|\boldsymbol{w})$, take the gradient with respect to the elements $[\boldsymbol{w}]_i$ of the vector $\boldsymbol{w}$ and set it to zero. If you find it difficult to do this in vector notation, it may be helpful to write out $[\Phi^T \boldsymbol{w}]_j = \sum_i [\boldsymbol{w}]_i [\Phi]_{ij}$. Calculate the derivative of $\log p(\boldsymbol{y}|\boldsymbol{w})$ with respect to $[\boldsymbol{w}]_i$ which is scalar. Setting that to zero, you can bring it to a form $\boldsymbol{v}^T [\Phi]_{i:} = 0$ (where $[\Phi]_{i:}$ is the i-th row of Φ) for some vector $\boldsymbol{v}(\boldsymbol{w})$ that is identical for all i, and thus, stacking up the columns of Φ again, we have $\boldsymbol{v}^T \Phi = 0$. Solving that equation for $\boldsymbol{w}$ yields the desired result.

   (b) By an analogous computation on the posterior $p(\boldsymbol{w}|\boldsymbol{y})$, show that the **maximum a-posteriori estimator** is identical to the posterior mean, $\boldsymbol{w}_{\text{MAP}} = \mathbb{E}_{p(\boldsymbol{w}|\boldsymbol{y})}(\boldsymbol{w})$.

   (c) There exists an important relationship between the regularization of least squares estimates and the choice of the prior in probabilistic linear regression. Given the Gaussian prior $p(\boldsymbol{w})$ for the particular choice $\boldsymbol{\mu} = 0, \Sigma = I_F, z, \Lambda = \sigma^2 I$, show that the MAP estimator calculated in part (b) is equivalent to the $\boldsymbol{l}_2$**-regularized least-squares** estimator (aka ridge regression)

$$\boldsymbol{w}_{\boldsymbol{l}_2} = (\Phi\Phi^T + \alpha I)^{-1}\Phi\boldsymbol{y},$$

   and give the corresponding value of the regularization parameter α.

   (d) Which choice of prior would a LASSO ($\boldsymbol{l}_1$) regularization correspond to?

3. **Practical Question** See `Exercise_03.ipynb`.