Deadline: Fr. Mai 10 Drop your printed or legible handwritten submissions into the boxes at Samelsonplatz, or upload a Jupyter notebook (`.ipynb`) or a `.pdf` file via LearnWeb.

1 GPs & Linear Regression (12 points)

(literature: Rasmussen Chapter 6, Murphy Chapter 15.4)

For a standard linear regression model the likelihood is given by

$$p(f_* \mid x_*, (X, Y)) = \mathcal{N}(Y \mid X\beta, \sigma^2 I)$$

With the optimal choice of parameters being $\hat{\beta} = (X^\intercal X)^{-1} X^\intercal Y$. A Bayesian linear regression (BLR) model (Murphy 7.6) extends this by factoring in a prior belief about β

$$p(\beta) = \mathcal{N}(\mu_\beta, \Sigma_\beta)$$

yielding the following posterior for the parameters

$$\begin{aligned} p(\beta|X, y) &\propto p(y|X, \beta)p(\beta) \\ &= \mathcal{N}(y \mid X\beta, \sigma^2 I)\mathcal{N}(\beta \mid \mu_\beta, \Sigma_\beta) = \mathcal{N}(\beta \mid \mu'_\beta, \Sigma'_\beta) \end{aligned}$$

and subsequently the posterior predictive distribution

$$\begin{aligned} p(y_* \mid X_*) = \mathbb{E}_\beta[p(y \mid \beta, X_*)] &= \int p(y \mid X_*, \beta)p(\beta)\mathrm{d}\beta \\ &= \int \mathcal{N}(y \mid X_*\beta, \sigma^2 I)\mathcal{N}(\beta \mid \mu'_\beta, \Sigma'_\beta)\mathrm{d}\beta \end{aligned}$$

A. [4p] Show that mean and variance of $p(\beta \mid X, y)$ are given by

$$\begin{aligned} \mu'_\beta &= (\tfrac{1}{\sigma^2} X^\intercal X + \Sigma_\beta^{-1})^{-1}(\tfrac{1}{\sigma^2} X^\intercal Y + \Sigma_\beta^{-1}\mu_\beta) \\ \Sigma'_\beta &= (\tfrac{1}{\sigma^2} X^\intercal X + \Sigma_\beta^{-1})^{-1} \end{aligned}$$

Note that we have already encountered such a model before! Ridge regression can be interpreted as a BLR model with $\mu_\beta = 0, \Sigma_\beta = \lambda I$.

B. [4p] Show that $p(y_* \mid X_*) = \mathcal{N}(y \mid X_*\mu'_\beta,\ \sigma^2 I + X_*\Sigma'_\beta X_*^\intercal)$

C. [4p] Show that the BLR is equivalent to a GP with a bi-linear kernel function

$$\kappa(x, x') = x^T \Lambda x'$$

Hint: You will need the following formulae

Theorem 1. *If* $p(x) = \mathcal{N}(x \mid \mu_x, \Sigma_x)$ *and* $p(y \mid x) = \mathcal{N}(y \mid Ax + b, \Sigma_y)$, *then*

$$\begin{aligned} p(x \mid y) = \mathcal{N}(x \mid \mu_{x|y}, \Sigma_{x|y}) \quad \textit{with} \quad & \mu_{x|y} = \Sigma_{x|y}\left[A^\intercal \Sigma_y^{-1}(y - b) + \Sigma_x^{-1}\mu_x\right] \\ & \Sigma_{x|y}^{-1} = \Sigma_x^{-1} + A^\intercal \Sigma_y^{-1} A \\ p(y) = \mathcal{N}(y \mid A\mu_x + b, \Sigma_y + A\Sigma_x A^\intercal) & \end{aligned}$$

(Murphy, 4.125 & 4.126)

Theorem 2. *Matrix Inversion Lemma*

$$(\mathbf{A} + \mathbf{UBV})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}\left(\mathbf{B}^{-1} + \mathbf{VA}^{-1}\mathbf{U}\right)^{-1}\mathbf{VA}^{-1}$$

2 GP Classification (8 points)

A. [4p] Explain briefly how classification with GPs works. What are the main difficulties?

B. [4p] Explain briefly how the Laplace Approximation method works