Probabilistic Inference & Learning

Exercise Sheet #4

Parametric Regression

1. **Least-Squares Estimation** The parametric regression model from Lectures 4 and 5 used the Gaussian likelihood and prior (for $y = [y_1, \dots, y_n] \in \mathbb{R}^n, X = [x_1, \dots, x_n] \in \mathbb{X}^n, \phi_X = [\phi(x_1), \dots, \phi(x_n)] \in \mathbb{R}^{F \times n}$ and $w \in \mathbb{R}^F, \sigma \in \mathbb{R}, \mu \in \mathbb{R}^F, \Sigma \in \mathbb{R}^{F \times F}$)

$$p(y \mid X, w) = \mathcal{N}(y; \phi_X^\intercal w, \sigma^2 I_n) \qquad p(w) = \mathcal{N}(w; \mu, \Sigma).$$

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

$$w_{\text{ML}} = (\phi_X \phi_X^\intercal)^{-1} \phi_X y.$$

   To do so, use the explicit form of the Gaussian pdf to write out $\log p(y \mid X, w)$, take the gradient with respect to the elements $[w]_i$ of the vector w and set it to zero. If you find it difficult to do this in vector notation, it may be helpful to write out $\phi_X^\intercal w = \sum_i w_i [\phi_X]_{i:}$ where $[\phi_X]_{i:}$ is the i-th column of ϕ_X. Calculate the derivative of $\log p(y \mid X, w)$ with respect to w_i, which is scalar. Setting that to zero, you can bring it to a form $v^\intercal [\phi_X]_{i:} = 0$ for some vector $v(w)$ that is identical for all i, and thus, stacking up the columns of ϕ_X again, we have $v^\intercal \phi_X = 0$. Solving that equation for w yields the desired result. *20 points*

   (b) By an analogous computation on the posterior $p(w \mid y, X)$, show that the **maximum a-posteriori estimator** is identical to the posterior mean

$$w_{\text{MAP}} = \mathbb{E}_{p(w \mid y, X)}(w) = (\Sigma^{-1} + \sigma^{-2} \phi_X \phi_X^\intercal)^{-1} (\Sigma^{-1} \mu + \sigma^{-2} \phi_X y).$$

   This result shows that, for the particular choice $\mu = 0, \Sigma = I_F$, the posterior mean is the **ℓ_2-regularized least-squares** estimator $w_{\text{MAP}} = (\sigma^2 I_F + \phi_X \phi_X^\intercal)^{-1} \phi_X y$. *30 points*

2. **Type-II maximum likelihood** Lecture 5 introduced hierarchical Bayesian inference (hyperparameter optimization) for general linear regression. Download `Gaussian_Linear_Regression.ipynb` from the "code" folder on Ilias and open it in jupyter lab. Uncomment the line in the second cell to load the toy dataset "nlindata.mat". Pick any set of features $\phi(x)$ you like (e.g. the ones listed in the final cell of the notebook, or any other), as long as $\phi(x) \in \mathbb{R}^F$ with $F \geq 3$.

   (a) add a line at the end of the cell to compute the log marginal likelihood (log evidence)

$$\log p(y \mid \phi) = \log \mathcal{N}(y; \phi_X^\intercal \mu, \phi_X^\intercal \Sigma \phi_X + \sigma^2 I)$$

   (note that the mean and the Cholesky decomposition of the covariance are already computed elsewhere in the script as the variables `M` and `G`; use these for efficiency). *10 points*

   (b) Identify at least 3 hyperparameters $\theta \in \mathbb{R}^3$ in your choice $\phi_\theta(x)$ (see slide 5 in Lecture 5 for an example). Using the results from slide 13 in the lecture, implement the gradient

$$\nabla_\theta (-2 \log p(y \mid \phi)).$$

   You can make your task easier by choosing (as is done in the notebook) $\mu = 0$, in which case the gradient elements can be computed using $K := \phi_X^\intercal \Sigma \phi_X$, $G := K + \sigma^2 I, \Gamma := G^{-1} y$, as

$$-2 \frac{\partial \log p(y \mid \phi_\theta)}{\partial \theta_i} = -\Gamma^\intercal \frac{\partial K}{\partial \theta_i} \Gamma + \operatorname{tr}\left(G^{-1} \frac{\partial K}{\partial \theta_i} \right) \text{ with } \frac{\partial K}{\partial \theta_i} = \left(\frac{\partial \phi_X}{\partial \theta_i} \right)^\intercal \Sigma \phi_X + \phi_X^\intercal \Sigma \left(\frac{\partial \phi_X}{\partial \theta_i} \right),$$

   and $\partial \phi_X / \partial \theta_i$ is the matrix of element-wise derivatives of your(!) choice of ϕ_X with respect to θ_i. Using this gradient and `scipy.optimize.minimize` (check its documentation), optimize the evidence term from 2.(a) for your features, and make a plot of the posterior on f arising from the optimal (in this sense) choice of θ. *40 points*