1. **Theory Question 1** Assume that you want to estimate the room temperature. To do so, you by three thermometers. These three thermometers measure the room temperature at different accuracies; the manufacturers specify accuracy as the standard deviation of the measurements (in degrees):

   T1: $\sigma_1 = 5.0$, T2: $\sigma_2 = 1.0$, T3: $\sigma_3 = 0.1$

   Upfront, you assume an improper (i.e. unnormalized) uniform prior across all temperatures.

   (a) You measure the temperature with T1 and the measured value is 32 degrees. What is the posterior distribution over the room temperature?

   (b) T2 measures a temperature of 31 degrees. Given measurements of T1 and T2, what is the posterior distribution over the room temperature?

   (c) T3 measures a temperature of 22 degrees. Given measurements of T1, T2, and T3, what is the posterior distribution over the room temperature?

   (d) Do you trust the inference result? Do you believe that the accuracies reported by the manufacturers are accurate?

   Solution:

   (a) $\mathcal{N}(T; 32, 5.0)$

   (b) $\mathcal{N}(T; 31.038, 0.98058)$

   (c) $\mathcal{N}(T; 22.09, 0.09948)$

   (d) No, because the evidence of these observations is practically zero. Most likely some of the accuracies of the thermomenters are wrong, or they are biased.

2. **Theory Question 2** 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.

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

   (b) Show that the **maximum a-posteriori estimator** is identical to the posterior mean, $\boldsymbol{w}_{\text{MAP}} = \mathbb{E}_{p(\boldsymbol{w}|\boldsymbol{y})}(\boldsymbol{w})$ (you can use the fact that the posterior is Gaussian).

   (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 $\boldsymbol{\mu} = 0, \Sigma = I_F, \Lambda = \sigma^2 I$, show that the MAP estimator calculated in part (b) is equivalent to the **$\boldsymbol{l}_2$-regularized least-squares** estimator (aka ridge regression)

   $$\boldsymbol{w}_{l_2} = (\Phi\Phi^T + \alpha I)^{-1}\Phi\boldsymbol{y},$$

   and give the corresponding value of the regularization parameter α.

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

Solution:

(a)

$$\begin{aligned}\log p(y|w) &= \log\left(\frac{\exp(-\frac{1}{2}(y-\Phi^\top w)^\top \sigma^{-2} I (y-\Phi^\top w))}{\sqrt{(2\Phi)^N |\sigma^2 I|}}\right)\\ &= -\frac{\sigma^{-2}}{2}(w^\top \Phi\Phi^\top w - 2y^\top \Phi^\top w + y^\top y) - \log(\sqrt{(2\phi)^N|\sigma^2 I|})\end{aligned}$$

Hence, omitting constant terms and using symmetry we have

$$\begin{aligned}\frac{\partial w^\top \Phi\Phi^\top w}{\partial w} &= 2\Phi\Phi^\top w\\ \frac{\partial y^\top \Phi^\top w}{\partial w} &= \Phi y.\end{aligned}$$

Taking the gradient and setting it to zero yields the desired result:

$$\begin{aligned}\frac{\partial}{\partial w}\log p(y|w) = 0 &\iff -\sigma^{-2}\Phi\Phi^\top w_{ML} + \sigma^{-2}\Phi y = 0\\ &\iff \Phi\Phi^\top w_{ML} = \Phi y\\ &\iff w_{ML} = (\Phi\Phi^\top)^{-1}\Phi y.\end{aligned}$$

(b) Since we can assume that the posterior is Gaussian, we only have to prove that the mean of a Gaussian is its mode. Assume a Gaussian $\mathcal{N}(\theta;\mu,\Sigma)$. Then we have to show that

$$\arg\max_\theta \log \mathcal{N}(\theta;\mu,\Sigma) = \mu$$

The computation is analogous to exercise (a).

(c) Using the formula for the posterior mean for Gaussian prior and likelihood, we get:

$$\begin{aligned}w_{MAP} &= (I + \sigma^{-2}\Phi\Phi^\top)^{-1}\sigma^{-2}\Phi y\\ &= (\Phi\Phi^\top + \sigma^2 I)^{-1}\Phi y,\end{aligned}$$

such that $\alpha = \sigma^2$.

(d) A Laplacian prior.

3. **Practical Question** See `Exercise_03_solution.ipynb`.