Course on probabilistic graphical models
Master MVA
Practice exercises 4: solutions

These exercises are not meant to provide an exhaustive coverage of the material to review for the final exam. To some extend they focus more specifically on material that is not covered in the homeworks. Also, all these exercises should not be taken as representative of the difficulty of the questions posed at the exam, although several questions of the exam are likely to have a similar style. Some exercises are easy, and a few can be much harder so don't be discouraged if you find some of them difficult. They are primarily designed to help you review and consolidate your understanding of the course.

Bayesian regression

Consider the Gaussian probabilistic conditional model seen in class for linear regression in which given a pair of variables (X, Y) with X taking values in $\mathbb{R}^d$ and Y in $\mathbb{R}$, we model the conditional distribution of Y given $X = x$ by a Gaussian distribution $\mathcal{N}(w^\top x, \sigma^2)$ parametrized by w and σ^2. Assume that σ^2 is fixed and w unknown and that the problem of learning the linear regression is approached from a Bayesian point of view, by placing a Gaussian prior distribution on w of the form $\mathcal{N}(0, \tau^2 I_d)$.

In the text, a single pair variables (x, Y) is considered. I will write the solution in the more general case where we assume that a sample of size n of the form $(x_1, y_1), \dots, (x_n, y_n)$ has been observed and one wants to consider the posterior distribution $p(w|x_1, y_1, \dots, x_n, y_n)$. Since no distribution has been specified for $x_1, \dots, x_n$ and since we are only interested in modeling the conditional distribution of Y given $X = x$, we will treat the observations $x_1, \dots, x_n$ as fixed and therefore as non random quantities. This entails that $p(w|x_1, y_1, \dots, x_n, y_n)$ is in fact the same as $p(w|y_1, \dots, y_n)$ (you can think of $x_1, \dots, x_n$ as acting like fixed parameters). We will write $\mathbf{y} = (y_1, \dots, y_n)^\top$. Normally we should write it with capital letters because it is a random variable but to avoid confusions, let's keep it uncapitalized, while keeping in mind that it is a random variable. Likewise, we will write $\mathbf{X} \in \mathbb{R}^{n \times d}$ the design matrix. This one is fixed. The use of bold letters is just to differentiate the variables that have all the data stacked, from a single observation. Finally we have $w \in \mathbb{R}^d$ which is also a random variable. I will use $\mathbf{w}$ for w, because it is strange to have just this one not bold.

1. Compute the parameters of the joint distribution of (w, Y).

   So, we will compute the joint distribution of $(\mathbf{w}, \mathbf{y})$. This would reduce to the question asked, if there was only a single observation $\mathbf{y} = y_1 = y$. First note that all the distributions that we are going to manipulate are Gaussian, in particular the joint distribution on $(\mathbf{w}, \mathbf{y})$ is Gaussian. To characterize this distribution, since it is Gaussian, we need to compute its expectation and its covariance matrix. The first thing to notice is that we can write $\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon}$, with $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2)$ and independent of w.

   We start be computing expectations. We know that $\mathbb{E}[\mathbf{w}] = 0$ by construction. Now $\mathbb{E}[\mathbf{y}] = \mathbf{X}\mathbb{E}[\mathbf{w}] + \mathbb{E}[\boldsymbol{\varepsilon}] = 0$.

   We have thus proved that

$$\mathbb{E}\begin{bmatrix}\mathbf{w}\\ \mathbf{y}\end{bmatrix} = 0$$

Let's consider the covariance matrix. Since the expectation is 0, the covariance matrix is also the matrix of second moments:

$$\mathbb{E}\begin{bmatrix}\mathbf{w}\mathbf{w}^\top & \mathbf{w}\mathbf{y}^\top \\ \mathbf{y}\mathbf{w}^\top & \mathbf{y}\mathbf{y}^\top\end{bmatrix}$$

We know that $\mathbb{E}[\mathbf{w}\mathbf{w}^\top] = \tau^2 I_d$. *We have*

$$\mathbb{E}[\mathbf{y}\mathbf{w}^\top] = \mathbb{E}[\mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon}\mathbf{w}^\top] = \mathbf{X}\mathbb{E}[\mathbf{w}\mathbf{w}^\top] + \mathbb{E}[\boldsymbol{\varepsilon}]\mathbb{E}[\mathbf{w}]^\top = \mathbf{X}\mathbb{E}[\mathbf{w}\mathbf{w}^\top] = \tau^2\mathbf{X}.$$

and

$$\mathbb{E}[\mathbf{y}\mathbf{y}^\top] = \mathbb{E}[(\mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon})(\mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon})^\top] = \mathbf{X}\mathbb{E}[\mathbf{w}\mathbf{w}^\top]\mathbf{X}^\top + \mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top] = \tau^2\mathbf{X}\mathbf{X}^\top + \sigma^2 I_n.$$

So that finally

$$\Sigma = \mathbb{E}\begin{bmatrix}\mathbf{w}\mathbf{w}^\top & \mathbf{w}\mathbf{y}^\top \\ \mathbf{y}\mathbf{w}^\top & \mathbf{y}\mathbf{y}^\top\end{bmatrix} = \tau^2\begin{bmatrix} I_d & \mathbf{X}^\top \\ \mathbf{X} & (\mathbf{X}\mathbf{X}^\top + \lambda I_n)\end{bmatrix} \qquad \textit{with} \quad \lambda := \frac{\sigma^2}{\tau^2}.$$

2. Compute de posterior distribution on w. *Now that we have computed the joint distribution. We can easily compute the conditional distribution of* $\mathbf{w}$ *given* $\mathbf{y}$. *Indeed denoting* $\Sigma_{\mathbf{w},\mathbf{w}}, \Sigma_{\mathbf{w},\mathbf{y}}, \Sigma_{\mathbf{y},\mathbf{w}}$ *and* $\Sigma_{\mathbf{y},\mathbf{y}}$, *the four blocks of the covariance matrix, reading them in row first order, we have shown in the course that (using the fact that marginal expectations are equal to 0) we have*

$$\mathbb{E}[\mathbf{w} \mid \mathbf{y}] = \Sigma_{\mathbf{w},\mathbf{y}}\Sigma_{\mathbf{y},\mathbf{y}}^{-1}\mathbf{y} = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \lambda I_n)^{-1}\mathbf{y} = (\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}\mathbf{X}^\top\mathbf{y},$$

where the last identity is not completely obvious, can be shown using the matrix inversion lemma for example and was not expected as part of the desired answer. It has however the merit of showing us that we retrieve the same form as for ridge regression.

Then for the conditional covariance, we have that

$$Cov(\mathbf{w}|\mathbf{y}) = \Sigma_{\mathbf{w},\mathbf{w}} - \Sigma_{\mathbf{w},\mathbf{y}}\Sigma_{\mathbf{y},\mathbf{y}}^{-1}\Sigma_{\mathbf{y},\mathbf{w}} = \tau^2\big(I_d - \mathbf{X}^\top(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}\mathbf{X}\big) = \sigma^2(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}.$$

3. Compute the predictive distribution over a new output variable y' given a new input x'.

Finally, for a new y' *we have* $y' = {x'}^\top\mathbf{w} + \varepsilon'$. *The joint distribution of* $(\mathbf{w}, \mathbf{y}, y')$ *is Gaussian (here implicitly given* $\mathbf{X}$ *and* x'*), which entails that the predictive distribution of* y', *which is by definition the marginal distribution of* y' *given the data is* $p(y' \mid \mathbf{y})$. *This last distribution must also be Gaussian, and so, one more time it is characterized by its mean and covariance.*

But using $y' = {x'}^\top\mathbf{w} + \varepsilon'$, *we have*

$$\mathbb{E}[y'|\mathbf{y}] = {x'}^\top\mathbb{E}[\mathbf{w}|\mathbf{y}] + \mathbb{E}[\varepsilon'|\mathbf{y}] = {x'}^\top\mathbb{E}[\mathbf{w}|\mathbf{y}] + \mathbb{E}[\varepsilon'] = {x'}^\top(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}\mathbf{X}^\top\mathbf{y}.$$

For the conditional variance, first we have

$$Var[y'|\mathbf{y}] = Var[\mathbf{w}^\top x'|\mathbf{y}] + Var[\varepsilon'|\mathbf{y}]$$

because $\varepsilon' \perp\!\!\!\perp \mathbf{w} \mid \mathbf{y}$ *for the simple reason that* $\varepsilon' \perp\!\!\!\perp (\mathbf{w}, \mathbf{y})$. *and so*

$$Var[y'|\mathbf{y}] = {x'}^\top Cov(\mathbf{w}|\mathbf{y})x' + Var(\varepsilon') = \sigma^2\big({x'}^\top(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}x' + 1\big)$$

Linear regression and Gaussian likelihood

(a) Let $\psi : H \to \Theta$ be a surjective mapping from $H \in \mathbb{R}^p$ to $\Theta \in \mathbb{R}^p$. Consider the two statistical models

$$\mathcal{P} := \{p_{\psi(\eta)} \mid \eta \in H\}, \quad \text{and} \quad \mathcal{P}' := \{p_\theta \mid \theta \in \Theta\},$$

with $p_\theta = p_{\psi(\eta)}$ when $\theta = \psi(\eta)$. Assume that based on a sample $\mathcal{S}$, the maximum likelihood estimator $\hat{\eta}$ for η in $\mathcal{P}$ exists and is unique. Show that the maximum likelihood estimator $\hat{\theta}$ for θ in $\mathcal{P}'$ exists, is unique, and that $\hat{\theta} = \psi(\hat{\eta})$.

Let $\tilde{\theta} = \psi(\hat{\eta})$. The surjectivity of ψ entails that, for all $\theta' \in \Theta$ there exists $\eta' \in H$ such that $\theta' = \psi(\eta')$ and we have

$$p_{\theta'} = p_{\psi(\eta')} \leq p_{\psi(\hat{\eta})} = p_{\tilde{\theta}}.$$

Since this is true for all θ', this shows that $\tilde{\theta}$ maximizes the likelihood. This is moreover the unique maximizer, because if there existed another maximizer $\tilde{\theta}' \neq \tilde{\theta}$, given that ψ is surjective, there would exist $\tilde{\eta}'$ such that $\tilde{\theta}' = \psi(\tilde{\eta}')$ and $\tilde{\eta}'$ would be a second maximizer of the likelihood in $\mathcal{P}$ which would contradict the uniqueness of $\hat{\eta}$. We thus have shown that $\tilde{\theta}$ is the unique maximizer of the likelihood and so $\hat{\theta} = \tilde{\theta}$.

(b) Consider a pair of random variables (X, Y) with $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}$ and with some joint distribution P. Assume that an i.i.d. sample $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ drawn from P is available. Assume that P is sufficiently nice and n sufficiently large, so that the maximum likelihood estimator for the parameters of the Gaussian model exists and is unique. Let

$$\hat{\mu} = \begin{pmatrix} \hat{\mu}_x \\ \hat{\mu}_y \end{pmatrix} \quad \text{and} \quad \widehat{\Sigma} = \begin{pmatrix} \widehat{\Sigma}_{xx} & \widehat{\Sigma}_{xy} \\ \widehat{\Sigma}_{yx} & \widehat{\sigma}_y^2 \end{pmatrix}$$

be respectively the maximum likelihood estimator of the mean and covariance matrix with $\hat{\mu}_x \in \mathbb{R}^p$, $\hat{\mu}_y \in \mathbb{R}$, $\widehat{\Sigma}_{xx} \in \mathbb{R}^{p \times p}$, $\widehat{\Sigma}_{xy} = \widehat{\Sigma}_{yx}^\top \in \mathbb{R}^p$, and $\widehat{\sigma}_y^2 \in \mathbb{R}_+$. In particular we assume that $\widehat{\Sigma}$ is invertible.

Assume now that still based on the same sample $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, a linear regression model of the form $y = w^\top x + b$ is estimated using ordinary least-squares regression (where w and b are the parameters for the uncentered and unnormalized data). Let $\hat{w}$ and $\hat{b}$ be obtained estimators and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \hat{w}^\top x_i - \hat{b})^2$ the average squared residuals. Show that $\hat{w}$, $\hat{b}$ and $\hat{\sigma}^2$ can be expressed as a function of $\hat{\mu}_x, \hat{\mu}_y, \widehat{\Sigma}_{xx}, \widehat{\Sigma}_{xy}$ and $\widehat{\sigma}_y^2$. Explain why and provide the formulas.

The linear regression function $z \mapsto \hat{w}^\top z + \hat{b}$ obtained by ordinary least-squares and the average squared residuals $\hat{\sigma}^2$ are actually the maximum likelihood estimates, respectively for the conditional expectation and for the conditional variance of a conditional Gaussian model of Y given X. A joint Gaussian model on (X, Y) can thus be reparameterized by splitting it into a Gaussian marginal distribution of X with mean $\mu_x^{marg} = \mu_x$ and covariance $\Sigma_{xx}^{marg} = \Sigma_{xx}$ and a conditional Gaussian distribution of Y given X with

$$\mathbb{E}[Y \mid X] = w^\top X + b = \Sigma_{yx}\Sigma_{xx}^{-1}(X - \mu_x) + \mu_y$$
$$\mathrm{Var}(Y \mid X) = \sigma^2 = \sigma_y^2 - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}$$

$$\begin{bmatrix} \mu_x^{marg} \\ \Sigma_{xx}^{marg} \\ w \\ b \\ \sigma^2 \end{bmatrix} = \psi(\mu, \Sigma) = \begin{bmatrix} \mu_x \\ \Sigma_{xx} \\ \Sigma_{xx}^{-1}\Sigma_{xy} \\ \mu_y - \Sigma_{yx}\Sigma_{xx}^{-1}\mu_x \\ \sigma_y^2 - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy} \end{bmatrix}.$$

Clearly ψ is a bijection since

$$\begin{bmatrix} \Sigma_{yx} \\ \mu_y \\ \sigma_y^2 \end{bmatrix} = \begin{bmatrix} \Sigma_{xx} w \\ b + w^\top \mu_x \\ \sigma^2 + w^\top \Sigma_{xx} w \end{bmatrix}$$

Applying the result of the previous question, if the MLE is unique for μ and Σ then it is also unique for w, b and σ^2 and

$$\begin{bmatrix} \hat{w} \\ \hat{b} \\ \hat{\sigma}^2 \end{bmatrix} = \begin{bmatrix} \widehat{\Sigma}_{xx}^{-1} \widehat{\Sigma}_{xy} \\ \mu_y - \widehat{\Sigma}_{yx} \widehat{\Sigma}_{xx}^{-1} \mu_x \\ \widehat{\sigma}_y^2 - \widehat{\Sigma}_{yx} \widehat{\Sigma}_{xx}^{-1} \widehat{\Sigma}_{xy} \end{bmatrix}.$$

Gaussian Markov Chain

The following exercise does not require to do any tedious calculations. If you get into complicated calculation, it means you have the wrong approach...

Assume that the ε_i are i.i.d. with $\varepsilon_i \sim \mathcal{N}(0,1)$ and let $X_1 = \varepsilon_1,\ X_2 = \rho X_1 + \varepsilon_2,\ X_3 = \rho X_2 + \varepsilon_3$.

(a) What is the precision matrix of the joint distribution of (X_1, X_2, X_3)?

(b) Compute $\mathbb{E}[X_2 \mid X_1, X_3]$ and $\mathrm{Var}(X_2 \mid X_1, X_3)$.

The joint density is

$$\propto \exp -\frac{1}{2}\left[x_1^2 + (x_2 - \rho x_1)^2 + (x_3 - \rho x_2)^2\right],$$

which reveals by identification with

$$\exp(-\frac{1}{2} x^\top \Lambda x + \eta^\top x - A(\eta, \Lambda))$$

that

$$\Lambda = \begin{bmatrix} 1+\rho^2 & -\rho & 0 \\ -\rho & 1+\rho^2 & -\rho \\ 0 & -\rho & 1 \end{bmatrix} \quad \textit{and} \quad \eta = \mu = 0.$$

So that

$$\mathit{Var}(X_2 \mid X_1, X_3) = \Lambda_{22}^{-1} = \frac{1}{1+\rho^2}$$

and

$$\mathbb{E}[X_2 \mid X_1, X_3] = \eta_2 - \Lambda_{22}^{-1} \Lambda_{2,(1,3)} (X_1, X_3)^\top = \frac{\rho}{1+\rho^2}(X_1 + X_3).$$

The computation of the conditional mean variance can also be obtained from completing the square in the expression of the density:

$$p(x_2|x_1, x_3) \propto \exp -\frac{1}{2}\left[(1+\rho^2)x_2^2 - 2\rho x_1 x_2 - 2\rho x_3 x_2\right] \propto \exp -\frac{1}{2}(1+\rho^2)\big(x_2 - 2\tfrac{\rho}{1+\rho^2}(x_1 + x_3)\big)^2.$$