



TD: Basic knowledge

1 Matrix Calculus

Question 1 : Let A and X be two real-valued matrices, computes

$$\frac{\partial \text{Tr}(XA)}{\partial X}$$

$$\begin{aligned} \frac{\partial \text{Tr}(XA)}{\partial X} &= \left(\frac{\partial \text{Tr}(XA)}{\partial X_{ij}} \right)_{1 \leq i, j \leq d} \\ &= \left(\frac{\partial}{\partial X_{ij}} \sum_k [XA]_{kk} \right)_{1 \leq i, j \leq d} \\ &= \left(\frac{\partial}{\partial X_{ij}} \sum_k \sum_l X_{kl} A_{lk} \right)_{1 \leq i, j \leq d} \\ &= \left(\frac{\partial}{\partial X_{ij}} \sum_k \sum_l X_{kl} A_{lk} \right)_{1 \leq i, j \leq d} \\ &= (A_{ji})_{1 \leq i, j \leq d} \\ &= A^T \end{aligned}$$

Question 2 : Let X be an invertible real-valued matrix, computes

$$\frac{\partial \det(X)}{\partial X}$$

$$\begin{aligned} \frac{\partial \det(X)}{\partial X} &= \left(\frac{\partial \det(X)}{\partial X_{ij}} \right)_{1 \leq i, j \leq d} \\ &= \left(\frac{\partial}{\partial X_{ij}} \det(X) \right)_{1 \leq i, j \leq d} \\ &= \left(\frac{\partial}{\partial X_{ij}} \sum_i \sum_j X_{ij} C_{ij} \right)_{1 \leq i, j \leq d} \quad \text{where } C_{ij} = (-1)^{i+j} \det(X_{-i, -j}) \text{ is the cofactor of } X_{ij} \\ &= (C_{ij})_{1 \leq i, j \leq d} \\ &= \text{Cof}(X)^T \quad \text{where } \text{Cof}(X) \text{ is the cofactor matrix of } X \\ &= \det(X)(X^{-1})^T \end{aligned}$$

2 Joint and Posterior Distributions

Use Bayesian formula to show that if $v \sim \mathcal{N}(\mu, K)$, $u|v \sim \mathcal{N}(Lv + m, \Sigma)$, then

$$v|u \sim \mathcal{N}(\mu + J[u - (Lv + m)], K - JLK)$$

where $J = K^T L^T (\Sigma + LKL^T)^{-1}$.

We can express u as a linear transformation of v and a noise term:

$$u = Lv + m + \varepsilon \quad \text{where } \varepsilon \sim \mathcal{N}(0, \Sigma)$$

- $\mathbb{E}(v) = \mu$
- $\mathbb{E}(u) = \mathbb{E}(Lv + m + \varepsilon) = L\mathbb{E}(v) + m = L\mu + m$
- $\text{Cov}(v) = K$
- $\text{Cov}(u) = \text{Cov}(Lv + m + \varepsilon) = L\text{Cov}(v)L^T + \text{Cov}(\varepsilon) = LKL^T + \Sigma$
- $\text{Cov}(v, u) = \text{Cov}(v, Lv + m + \varepsilon) = \text{Cov}(v, Lv + m) = K^T L^T$
- $\text{Cov}(u, v) = \text{Cov}(Lv + m + \varepsilon, v) = \text{Cov}(v, Lv + m + \varepsilon)^T = LK$

Therefore, the joint distribution of u and v is

$$\begin{pmatrix} v \\ u \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} \mu \\ L\mu + m \end{pmatrix}, \begin{pmatrix} K & K^T L^T \\ LK & LKL^T + \Sigma \end{pmatrix} \right)$$

Using the conditional distribution formula for Gaussian vectors, we have

$$\mathbb{E}(v|u) = \mathbb{E}(v) + \text{Cov}(v, u)\text{Cov}(u)^{-1}(u - \mathbb{E}(u)) = \mu + K^T L^T (\Sigma + LKL^T)^{-1}(u - L\mu - m)$$

$$\text{Cov}(v|u) = \text{Cov}(v) - \text{Cov}(v, u)\text{Cov}(u)^{-1}\text{Cov}(u, v) = K - K^T L^T (\Sigma + LKL^T)^{-1} LK$$

Let $J = K^T L^T (\Sigma + LKL^T)^{-1}$, we have

$$v|u \sim \mathcal{N}(\mu + J[u - (Lv + m)], K - JLK)$$

3 EM Algorithm

In the slide on the justification of EM algorithm using log partition function, explain how the key idea of EM is related to

$$\log Z = \max_q \int f(x)q(x)dx - \int \log q(x)q(x)dx$$

You should specify the function f and the density q in order to make the connection.

The Expectation-Maximization (EM) algorithm is a powerful iterative method used to perform maximum likelihood estimation (MLE) in the presence of latent (unobserved) variables.

The EM algorithm alternates between two steps:

1. E-step: Compute the expected value of the log-likelihood function with respect to the conditional distribution of the latent variables given the observed data and the current estimate of the parameters.
2. M-step: Maximize the expected value of the log-likelihood function with respect to the parameters.

In the context of the log partition function $\log Z$, we have

- $f(x) = \log p(x, z|\theta)$ is the log-likelihood function of the complete data x and latent variable z given the parameters θ .
- $q(x) = p(z|x, \theta^{(t)})$ is the conditional distribution of the latent variable z given the observed data x and the current estimate of the parameters $\theta^{(t)}$.

The connection to the log partition function is that the EM algorithm can be seen as maximizing a lower bound on the log-likelihood, which is analogous to maximizing the log partition function $\log Z$. The optimal $q(x)$ is given by

$$q(x) = \frac{e^{f(x)}}{Z}$$

where $Z = \int e^{f(x)}dx$ is the partition function.