Assignment 2E, 2025

DD2434 Machine Learning, Advanced Course

Riccardo Alfonso Cerrone cerrone@kth.se

Bruno Carchia carchia@kth.se

December 7, 2025

2.1 Exponential Family

Question 2.1.1

The Poisson distribution is

$$p(x \mid \lambda) = \frac{\lambda^x}{x!} e^{-\lambda}.$$

We can rewrite it as:

$$p(x \mid \lambda) = \frac{1}{x!} e^{(x \log \lambda - \lambda)}, = h(x) e^{\eta^T t(x) - a(\eta)}$$

We assume

$$\eta = \log \lambda$$

$$\lambda = e^{\eta}$$

$$h(x) = \frac{1}{x!}, \qquad t(x) = x, \qquad a(\eta) = e^{\eta}.$$

Final result

$$p(x|\lambda) = \frac{1}{x!} e^{x\eta - e^{\eta}}$$

Question 2.1.2

By definition , we know that:

$$\begin{aligned} G(\lambda) &= \mathbb{E}_{\lambda}\Big[(\nabla_{\lambda} \log p(\beta \mid \lambda))(\nabla_{\lambda} \log p(\beta \mid \lambda))^{\top}\Big] \\ &= \mathbb{E}_{\lambda}\Big[(\, t(\beta) - \mathbb{E}_{\lambda}[t(\beta)]\,)(\, t(\beta) - \mathbb{E}_{\lambda}[t(\beta)]\,)^{\top}\Big] \\ &= \nabla^2_{\lambda} a_g(\lambda) \end{aligned}$$

In our case, we express it wrt the natural parameters

$$I(\eta) = \mathbb{E}_{\eta}\Big[(\nabla_{\eta} \log p(\beta \mid \eta))(\nabla_{\eta} \log p(\beta \mid \eta))^{\top}\Big]$$

We can exploit the properties of the exponential family

$$P(x|\eta) = h(x)e^{\eta^T t(x) - a(\eta)}$$

$$\log P(x|\eta) = \log h(x) + \eta^T t(x) - a(\eta)$$

$$\nabla_\eta \log P(x|\eta) = t(x) - \nabla_\eta a(\eta)$$

We use it for the definition of the Fisher Information

$$I(\eta) = [(t(x) - \nabla_\eta a(\eta))(t(x) - \nabla_\eta a(\eta))^T]$$

Knowing that $E_{p(x|\eta)}[t(x)] = \nabla a(\eta)$, we can rewrite it as

$$I(\eta) = [(t(x) - E_{p(x|\eta)}[t(x)])(t(x) - E_{p(x|\eta)}[t(x)])^T]$$

$$I(\eta) = \nabla_\eta^2 a(\eta)$$

In our case:

$$a(\eta) = e^\eta$$

$$\eta = \log \lambda$$

$$\lambda = e^\eta$$

We get

$$I(\eta) = e^\eta = e^{\log \lambda} = \lambda$$

We need to use the transformation function

$$I_\lambda(\lambda) = I_\eta(\eta(\lambda))(\frac{\partial \eta}{\partial \lambda})^2$$

$$I_\lambda(\lambda) = \lambda \cdot \frac{1}{\lambda^2} = \frac{1}{\lambda}$$

This formula can be obtained by using the chain rule in the derivation process as showed below

$$I(\eta) = \mathbb{E}\Big[(\nabla_\eta \log p(\beta \mid \eta))(\nabla_\eta \log p(\beta \mid \eta))^\top\Big]$$

$$\nabla_\eta \log p(\beta \mid \eta) = \nabla_\lambda \log p(\beta \mid \lambda) \cdot \frac{d\lambda}{d\eta}$$

$$I(\eta) = \mathbb{E}\Big[(\frac{d\lambda}{d\eta})^2(\nabla_\lambda \log p(\beta \mid \lambda))(\nabla_\lambda \log p(\beta \mid \lambda))^T\Big]$$

$$I(\eta) = (\frac{d\lambda}{d\eta})^2 \mathbb{E}_\lambda\big[(\nabla_\lambda \log p(\beta \mid \lambda))(\nabla_\lambda \log p(\beta \mid \lambda))^T\big]$$

$$I(\eta) = (\frac{d\lambda}{d\eta})^2 \cdot I(\eta)$$

Question 2.1.3

We need just to replace each component inside the definition of exponential family

$$
\begin{aligned}
& e^{(\theta_1-1,-\theta_2)\binom{\log x}{x}-\log\Gamma(\eta_1+1)+(\eta_1+1)\log(-\eta_2)} \\
\overset{\eta}{=} \quad & e^{(\theta_1-1,-\theta_2)\binom{\log x}{x}-\log\Gamma(\theta_1-1+1)+(\theta_1-1+1)\log(-(-\theta_2|))} \\
= \quad & e^{(\theta_1-1-\theta_2)\binom{\log x}{x}-\log\Gamma(\theta_1)+\theta_1\log(\theta_2)} \\
\overset{\theta}{=} \quad & e^{(\alpha-1,-\beta)\binom{\log x}{x}-\log\Gamma(\alpha)+\alpha\log(\beta)} \\
= \quad & e^{(\alpha-1)\log x-\beta x-\log\Gamma(\alpha)+\alpha\log(\beta)} \\
= \quad & \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x} \sim \text{Gamma}(\alpha,\beta)
\end{aligned}
$$

Question 2.1.4

$$
\begin{aligned}
& \frac{1}{x\sqrt{2\pi}}e^{\left(\frac{\theta_1}{\theta_2},-\frac{1}{2\theta_2}\right)\cdot\binom{\log x}{(\log x)^2}-\left(-\frac{\eta_1^2}{4\eta_2}-\frac{1}{2}\log(-2\eta_2)\right)} \\
\overset{\eta}{=} & \frac{1}{x\sqrt{2\pi}}e^{\left(\frac{\theta_1}{\theta_2},-\frac{1}{2\theta_2}\right)\cdot\binom{\log x}{(\log x)^2}-\frac{\theta_1^2}{2\theta_2}+\frac{1}{2}\log\left(\frac{1}{\theta_2}\right)} \\
\overset{\theta}{=} & \frac{1}{x\sqrt{2\pi}}e^{\left(\frac{\mu}{\sigma^2},-\frac{1}{2\sigma^2}\right)\cdot\binom{\log x}{(\log x)^2}-\frac{\mu^2}{2\sigma^2}+\frac{1}{2}\log\left(\frac{1}{\sigma^2}\right)} \\
= & \frac{1}{x\sqrt{2\pi\sigma^2}}e^{\left(\frac{\mu}{\sigma^2},-\frac{1}{2\sigma^2}\right)\cdot\binom{\log x}{(\log x)^2}-\frac{\mu^2}{2\sigma^2}} \\
= & \frac{1}{x\sqrt{2\pi\sigma^2}}e^{\left(\frac{\mu}{\sigma^2},-\frac{1}{2\sigma^2}\right)\cdot\binom{\log x}{(\log x)^2}-\frac{\mu^2}{2\sigma^2}} \\
= & \frac{1}{x\sqrt{2\pi\sigma^2}}e^{\frac{\mu}{\sigma^2}\log x-\frac{1}{2\sigma^2}(\log x)^2-\frac{\mu^2}{2\sigma^2}} \\
= & \frac{1}{x\sqrt{2\pi\sigma^2}}e^{\frac{-1}{2\sigma^2}\left(-2\mu\log x+(\log x)^2-\mu^2\right)} \\
= & \frac{1}{x\sqrt{2\pi\sigma^2}}e^{-\frac{(\log x-\mu)^2}{2\sigma^2}} \sim \text{ logNormal } (\mu,\sigma^2)
\end{aligned}
$$

2.3 BBVI

Question 2.3.8

We know that:

$$P(X_n|\lambda) \sim \text{Poisson}(\lambda) \qquad P(\lambda) \sim \text{Gamma}(\alpha,\beta)$$

$$z_s \sim q(\lambda) = \text{Exponential}(\theta)$$

$$p(x,\lambda) = p(\lambda)\cdot\prod_{n=1}^{N}p(x_n|\lambda) = \left(\frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}\right)\cdot\prod_{n=1}^{N}\left(\frac{\lambda^{x_n}e^{-\lambda}}{x_n!}\right)$$

$$q(\lambda|\theta) = \theta e^{-\theta \cdot \lambda}$$

We want to write an expression for the Naive BBVI gradient estimate w.r.t θ , starting from its general definition (considering the paper notation , we apply the following changes $z = \lambda$ and $\lambda = \theta$)

$$\nabla_\theta \mathcal{L} = \mathbb{E}_{q(\lambda|\theta)}[\nabla_\theta \log q(\lambda|\theta) \left(\log p(x, \lambda) - \log q(\lambda|\theta)\right)]$$

We compute noisy unbiased gradients of the ELBO with Monte Carlo samples from the variational distribution

$$\nabla_\theta \mathcal{L} \approx \frac{1}{S} \sum_{s=1}^{S} \nabla_\theta \log q(\lambda_s|\theta) \left(\log p(x, \lambda_s) - \log q(\lambda_s|\theta)\right)$$

with $\lambda_s \sim q(\lambda|\theta)$ Where:

- $\log q(\lambda_s|\theta) = \log \theta - \theta \lambda_s$
- $\nabla_\theta \log q(\lambda_s|\theta) = \frac{1}{\theta} - \lambda_s$
- $\log p(x, \lambda_s) = \log \frac{\beta^\alpha}{\Gamma(\alpha)} + (\alpha - 1) \log \lambda_s - \beta \lambda_s + \sum_{n=1}^{N} \left(x_n \log \lambda_s - \lambda_s - \log x_n!\right)$

Question 2.3.9

In the BBVI (Black Box Variational Inference) paper, Control Variates are used to reduce the variance of the gradient estimators of the ELBO while keeping them unbiased.

2.4 Variational Autoencoders

Question 5.1

We want to prove that the elbo can be rewritten as $E_{q(z)}\big[\log p(x|z)\big] - D_{KL}\big(q(z)||p(z)\big)$

$$\begin{aligned}
\mathcal{L} &= \mathbb{E}_{q(z)}\left[\log \frac{p(x,z)}{q(z)}\right] = \int q(z) \cdot \log \frac{p(x,z)}{q(z)}\, dz \\
&= \int q(z) \cdot \log \frac{p(x|z)p(z)}{q(z)}\, dz \\
&= \int q(z) \cdot \log p(x|z)\, dz + \int q(z) \log \frac{p(z)}{q(z)}\, dz \\
&= \int q(z) \cdot \log p(x|z)\, dz - \int q(z) \log \frac{q(z)}{p(z)}\, dz \\
&= \mathbb{E}_{q(z)}\left[\log p(x|z)\right] - D_{KL}\left(q(z) \| p(z)\right)
\end{aligned}$$

Question 5.2

Kullback–Leibler divergence can be computed using the closed-form analytic expression when both the variational and the prior distributions are Gaussian. We write down this KL divergence in terms of the parameters of the prior and the variational distributions by considering a generic case where the latent space is K-dimensional.

$$\begin{aligned}
-D_{\mathrm{KL}}(q_\phi(z|x)||p_\theta(z)) &= \int q_\theta(z)\left(\log p_\theta(z) - \log q_\theta(z)\right) dz \\
&= \int q_\theta(z)\left(\log p_\theta(z)\right) dz + \int q_\theta(z)\left(\log q_\theta(z)\right) dz \\
&= \int \mathcal{N}(z;\mu,\sigma^2)\log\mathcal{N}(z;0,I)dz + \int \mathcal{N}(z;\mu,\sigma^2)\log\mathcal{N}(z;\mu,\sigma^2)dz \\
&= -\frac{K}{2}\log(2\pi) - \frac{1}{2}\sum_{k=1}^{K}(\mu_k^2+\sigma_k^2) - \frac{K}{2}\log(2\pi) - \frac{1}{2}\sum_{k=1}^{K}(1+\log\sigma_k^2) \\
&= \frac{1}{2}\sum_{k=1}^{K}\left[1+\log(\sigma_k^2) - \mu_k^2 - \sigma_k^2\right]
\end{aligned}$$

0.1 Question 5.3

To compute the ELBO, we note that the KL divergence component was already derived in the previous question. We now proceed to compute the remaining expectation term $E_{q(z)}[\log p(x|z)]$, as shown below.

$$E_{q(z)}[\log p(x|z)] = \int q(z|x)\log p(x|z)dz$$

$$\underset{\substack{\uparrow \\ \text{MONTE} \\ \text{CARLO} \\ \text{SAMPLES}}}{\approx} \log p(x|z_1) \quad \text{WHERE } z_1 \sim q(z|x)$$

$$P(x|z_1) = \text{Bernoulli}(g(z_1)) = \prod_{i=1}^{D}(1-g(z_1)_i)^{1-x_i}\cdot g(z_1)_i^{x_i}$$

$$\log p(x|z_1) = \sum_{i=1}^{D}(1-x_i)\log(1-g(z_1)_i) + x_i\log(g(z_1)_i)$$

$$E_{q(z)}[\log p(x|z)] \approx \log p(x|z_1) = \sum_{i=1}^{D}(1-x_i)\log(1-g(z_1)_i) + x_i\log(g(z_1)_i)$$

We assumed only one monte carlo sampling $L = 1$ as suggested by the TA

2.2 SVI - LDA

Question 2.2.5

Given the model described in the Hoffman paper, let us define x_n as the n-th observation, z_n as the n-th local hidden variable, β as the global hidden variables, and α as the fixed hyperparameters. According to this notation and the corresponding graphical model, the joint distribution can be written as

$$p(X, Z, \beta \mid \alpha) = p(\beta \mid \alpha) \prod_{n=1}^{N} p(x_n, z_n \mid \beta). \tag{1}$$

The equality

$$p(x_n, z_n \mid x_{-n}, z_{-n}, \beta) = p(x_n, z_n \mid \beta).$$

holds because of the fork structure in the graphical model: given the global variables β, each pair (x_n, z_n) is conditionally independent of all other pairs (x_{-n}, z_{-n}), therefore it is dependent on the global variables only.

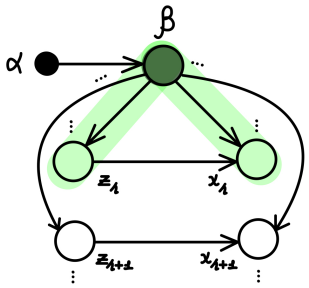

Figure 1: V structure in the given model

The global variables β are parameters equipped with a prior distribution $p(\beta)$, while each local variable z_n contains the latent structure associated with the n-th observation.

The variables z_n are called *local* because each of them is tied solely to its corresponding observation x_n and does not interact with the other local hidden variables.

Question 2.2.6

Question 2.2.7

For this question, we were asked to implement the SVI updates and the SVI algorithm for the LDA model. To do so, we followed the formulation of Latent Dirichlet Allocation given in Section 3.2 of Hoffman et al.

Algorithm 1 Stochastic Variational Inference for LDA

Initialize $\lambda^{(0)}$ randomly.
Set the step-size schedule ρ_t appropriately.
repeat
  Sample a document w_d uniformly from the data set.
  Initialize $\gamma_{dk} = 1$, for $k \in \{1, \ldots, K\}$.
  repeat
    for $n \in \{1, \ldots, N\}$ **do**
      $\phi_{dn}^k \propto \exp\{\mathbb{E}[\log \theta_{dk}] + \mathbb{E}[\log \beta_{k,w_{dn}}]\}, \quad k \in \{1, \ldots, K\}.$
    end for
    $\gamma_d = \alpha + \sum_n \phi_{dn}$.
  until local parameters ϕ_{dn} and γ_d converge.
  for $k \in \{1, \ldots, K\}$ **do**
    $\hat{\lambda}_k = \eta + D \sum_{n=1}^{N} \phi_{dn}^k w_{dn}$.
  end for
  $\lambda^{(t)} = (1 - \rho_t)\lambda^{(t-1)} + \rho_t \hat{\lambda}$.
until forever

Update $q(Z)$

Update $q(\theta)$

Update $q(\beta)$