

HW1 – Solutions

Thomas Marchioro

March 2023

Exercise 1

- (a) We know that $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y = e^X = g(X)$. The PDF of a normal random variable with mean μ and variance σ^2 is

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad (1)$$

The PDF of Y can be obtained by applying the change of variable formula

$$p_Y(y) = p_X(g^{-1}(y)) \left| \frac{\partial g^{-1}}{\partial y}(y) \right| \quad (2)$$

with $g^{-1}(y) = \log(y)$. Notice that the g^{-1} is a logarithm function, which is defined only when its argument (namely, y) is positive. This means that Y can never take values that are lower or equal to zero. Therefore, the PDF of Y must have the following structure:

$$p_Y(y) = \begin{cases} \text{something} & \text{if } y > 0 \\ 0 & \text{if } y \leq 0 \end{cases} \quad (3)$$

To determine the value of the “something” in the $y > 0$ case, we just plug $\log(y)$ in the change of variable formula, obtaining

$$p_Y(y) = \begin{cases} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\log(y)-\mu)^2}{2\sigma^2}\right) \cdot \frac{1}{y} & \text{if } y > 0 \\ 0 & \text{if } y \leq 0 \end{cases} \quad (4)$$

where there is no need to keep the absolute value for the derivative $1/y$, since that is always positive for $y > 0$.

- (b) A comparison between the empirical distribution of Y and its PDF is shown in figure 1.

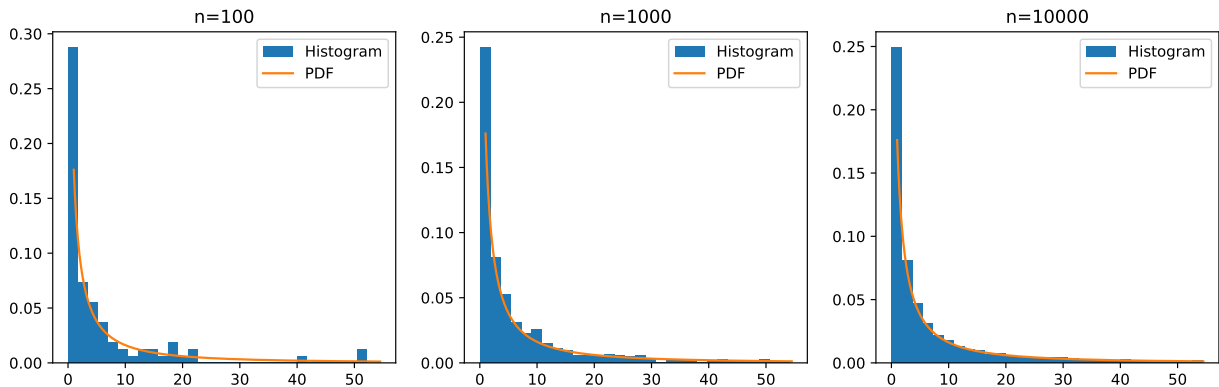


Figure 1: Histogram validation of the PDF obtained in exercise 1a.

- (c) We can start by observing that the transformation $Z = -\lambda^{-1} \log(U)$, $\lambda > 0$ with $U \sim \mathcal{U}(0,1)$ follows an exponential distribution $Z \sim \text{Exp}(\lambda)$. This is a known Probability 101 fact, but we can also easily prove it using the change of variable formula, with $Z = g(U) = -\lambda^{-1} \log(U)$. We first

calculate the inverse $g^{-1}(z) = -\exp(-\lambda z)$. Also, notice that since the range of U is $(0, 1)$, its logarithm can never take negative values, meaning that

$$p_Z(z) = \begin{cases} \text{something} & \text{if } z \geq 0 \\ 0 & \text{if } z < 0 \end{cases} \quad (5)$$

Again, the “something” is computed using the change of variable formula

$$p_Z(z) = \begin{cases} \lambda \exp(-\lambda z) & \text{if } z \geq 0 \\ 0 & \text{if } z < 0 \end{cases} \quad (6)$$

which is the PDF of an exponential distribution with parameter $\lambda > 0$. The random variable $Y = -\lambda_1^{-1} \log(U_1) - \lambda_2^{-1} \log(U_2)$ is essentially the sum of two independent random variables $Z_1 \sim \text{Exp}(\lambda_1)$ and $Z_2 \sim \text{Exp}(\lambda_2)$. This can be computed as the convolution of the PDFs of Z_1 and Z_2

$$p_Y(y) = p_{Z_1} * p_{Z_2}(y) = \int_{-\infty}^{\infty} p_{Z_1}(z) p_{Z_2}(y - z) dz \quad (7)$$

$$= \int_0^y \lambda_1 \exp(-\lambda_1 z) \cdot \lambda_2 \exp(-\lambda_2(y - z)) dz \quad (8)$$

$$= \lambda_1 \lambda_2 \exp(-\lambda_2 y) \int_0^y \exp(-\lambda_1 z) \cdot \exp(\lambda_2 z) dz \quad (9)$$

$$= \lambda_1 \lambda_2 \exp(-\lambda_2 y) \int_0^y \exp((\lambda_2 - \lambda_1)z) dz \quad (10)$$

$$= \frac{\lambda_1 \lambda_2}{\lambda_1 - \lambda_2} \exp(-\lambda_2 y) (1 - \exp((\lambda_2 - \lambda_1)y)) \quad (11)$$

$$= \frac{\lambda_1 \lambda_2}{\lambda_1 - \lambda_2} (\exp(-\lambda_2 y) - \exp(-\lambda_1 y)), \text{ for } y \geq 0. \quad (12)$$

For $y < 0$ the PDF of Y is again 0. We can deduce this when calculating the extremes of integration or simply by noticing that the sum of two positive random variables Z_1 and Z_2 must also be positive.

(d) A comparison between the empirical distribution of Y and its PDF is shown in figure 2.

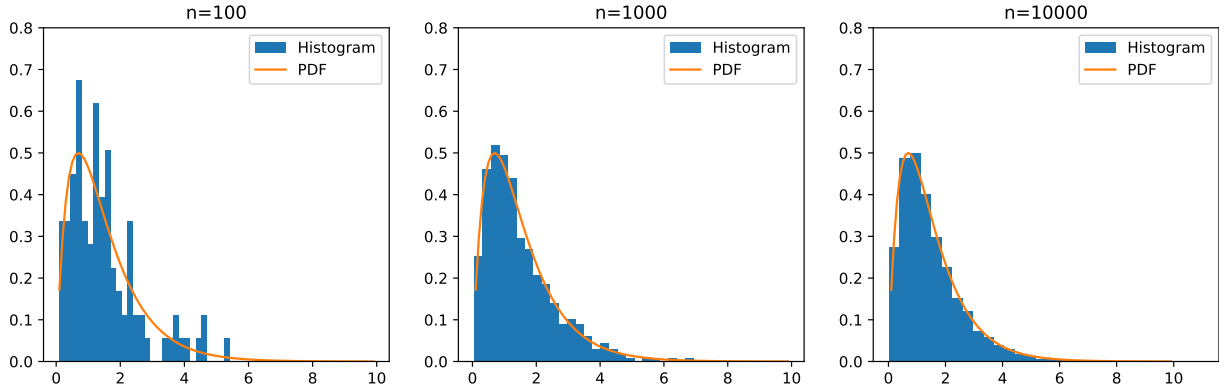


Figure 2: Histogram validation of the PDF obtained in exercise 1c.

1 Exercise 2

(a) In order to prove that the sum of two dependent Gaussian r.v.s is Gaussian, we can use the following facts:

- The sum of two independent Gaussian r.v.s is Gaussian. This can be easily proven by computing the convolution of $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ (see Appendix).
- The sum of two dependent Gaussians can be written as the sum of two independent Gaussians (see last exercise of tutorial 3).

This is enough to conclude that $Y = X_1 + X_2$ (with X_1, X_2 dependent Gaussian r.v.s) follows a Gaussian distribution. In order to compute the PDF of Y , we can take advantage of the fact that a Gaussian distribution is uniquely characterized by its mean and variance. We can compute the mean μ_Y by using the linearity of expectation

$$\mu_Y = \mathbb{E}[X_1 + X_2] = \mathbb{E}[X_1] + \mathbb{E}[X_2] = \mu_1 + \mu_2. \quad (13)$$

We also can compute the variance by using its definition

$$\sigma_Y^2 = \mathbb{E}[(X_1 + X_2 - \mu_Y)^2] \quad (14)$$

$$= \mathbb{E}[(X_1 + X_2 - \mu_1 - \mu_2)^2] \quad (15)$$

$$= \mathbb{E}[(X_1 - \mu_1)^2 + (X_2 - \mu_2)^2 + 2(X_1 - \mu_1)(X_2 - \mu_2)] \quad (16)$$

$$= \mathbb{E}[(X_1 - \mu_1)^2] + \mathbb{E}[(X_2 - \mu_2)^2] + 2\mathbb{E}[(X_1 - \mu_1)(X_2 - \mu_2)] \quad (17)$$

$$= \sigma_1^2 + \sigma_2^2 + 2\sigma_{12}. \quad (18)$$

The PDF of Y is hence the PDF of a Gaussian r.v. with mean $\mu_1 + \mu_2$ and variance $\sigma_1^2 + \sigma_2^2 + 2\sigma_{12}$, i.e.

$$p_Y(y) = \frac{1}{\sqrt{2\pi(\sigma_1^2 + \sigma_2^2 + 2\sigma_{12})}} \exp\left(-\frac{(y - \mu_1 - \mu_2)^2}{2(\sigma_1^2 + \sigma_2^2 + 2\sigma_{12})}\right). \quad (19)$$

Notice that if we denote the mean vector and covariance matrix for the joint distribution of X_1, X_2 as

$$\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \text{ and } \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \quad (20)$$

respectively, we have that:

- The mean μ_Y is the sum of the two elements of the mean vector;
- The variance σ_Y^2 is the sum of all the elements of the covariance matrix.

We will use these results in the next part of the exercise.

- (b) Let us denote (without loss of generality) the elements of the mean vector and covariance matrix for the triplet X_1, X_2, X_3 as follows

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{bmatrix}. \quad (21)$$

We can first compute the joint conditional distribution $p(x_1, x_2 | x_3)$ by using the formula from slide 29 of Lecture 2, letting $x_A = [x_1, x_2]^\top$ and $x_B = x_3$, i.e.,

$$\mu_A = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \mu_B = \mu_3, \quad (22)$$

$$\Sigma_{AA} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}, \quad \Sigma_{BB} = \sigma_3^2, \quad (23)$$

and

$$\Sigma_{AB} = \begin{bmatrix} \sigma_{13} \\ \sigma_{23} \end{bmatrix}, \quad \Sigma_{BA} = [\sigma_{13} \quad \sigma_{23}]. \quad (24)$$

A straightforward application of the formula yields:

$$\mu_{X_1, X_2 | x_3} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} \sigma_{13} \\ \sigma_{23} \end{bmatrix} \cdot \frac{1}{\sigma_3^2} \cdot (x_3 - \mu_3), \quad (25)$$

and

$$\Sigma_{X_1, X_2 | x_3} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} - \begin{bmatrix} \sigma_{13} \\ \sigma_{23} \end{bmatrix} \cdot \frac{1}{\sigma_3^2} \cdot [\sigma_{13} \quad \sigma_{23}]. \quad (26)$$

Some simple algebraic steps lead to

$$\mu_{X_1, X_2 | x_3} = \begin{bmatrix} \mu_1 + \frac{\sigma_{13}}{\sigma_3^2}(x_3 - \mu_3) \\ \mu_2 + \frac{\sigma_{23}}{\sigma_3^2}(x_3 - \mu_3) \end{bmatrix} \quad (27)$$

and

$$\Sigma_{X_1, X_2|x_3} = \begin{bmatrix} \sigma_1^2 - \frac{\sigma_{13}^2}{\sigma_3^2} & \sigma_{12} - \frac{\sigma_{13}\sigma_{23}}{\sigma_3^2} \\ \sigma_{12} - \frac{\sigma_{13}\sigma_{23}}{\sigma_3^2} & \sigma_2^2 - \frac{\sigma_{23}^2}{\sigma_3^2} \end{bmatrix}. \quad (28)$$

As a final step, we know that, since $p(x_1, x_2|x_3)$ follows a multivariate Gaussian distribution, also $p(x_1 + x_2|x_3)$ should follow a (univariate) Gaussian distribution. This can be characterized by its mean and variance, computed as shown in point (a) by summing the elements of the mean vector and covariance matrix:

$$\mu_{X_1+X_2|x_3} = \mu_1 + \mu_2 + \frac{\sigma_{13} + \sigma_{23}}{\sigma_3^2}(x_3 - \mu_3) \quad (29)$$

$$\sigma_{X_1+X_2|x_3}^2 = \sigma_1^2 + \sigma_2^2 + 2\sigma_{12} - \frac{1}{\sigma_3^2}(\sigma_{13}^2 + \sigma_{23}^2 + 2\sigma_{13}\sigma_{23}). \quad (30)$$

The PDF $p(x_1 + x_2|x_3)$ is simply a Gaussian PDF with mean $\mu_{X_1+X_2|x_3}$ and variance $\sigma_{X_1+X_2|x_3}^2$.

- (c) Comparison between the empirical distribution of Y and its PDF for $x_3 = -1, 0, 1$. The value of x_3 does not change the distribution of Y . This is due to the structure of the covariance matrix: X_1 and X_2 have opposite mean and are negative correlated with each other. Furthermore, they have the same degree of correlation with X_3 . This implies that whatever value X_3 will take, it will “even out” when X_1 and X_2 are summed (you can verify that by plugging the mean and covariance values in eq. 29 and see that x_3 gets canceled out).

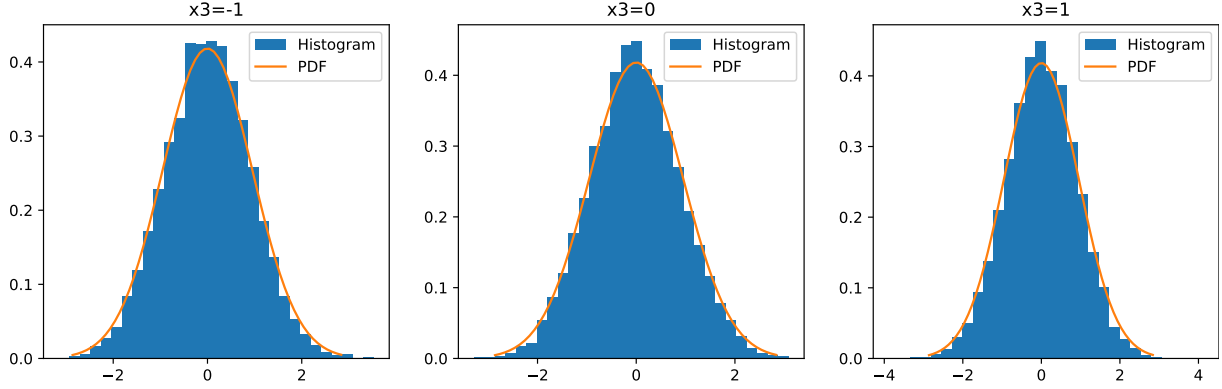


Figure 3: Histogram of Y for $x_3 = -1, 0, 1$.

Exercise 3

- (a) The maximum likelihood estimator $\hat{\theta}_{\text{MLE}}$ can be obtained by maximizing the log-likelihood w.r.t. the observations x_1, \dots, x_n :

$$\hat{\theta}_{\text{MLE}} = \arg \max_{\theta} \log \prod_{i=1}^n \left(\frac{1}{\sqrt{2\pi\theta}} \exp \left(-\frac{x_i^2}{2\theta} \right) \right) \quad (31)$$

$$= \arg \max_{\theta} \sum_{i=1}^n \left(-\frac{x_i^2}{2\theta} - \frac{1}{2} \log \theta - \frac{1}{2} \log(2\pi) \right) \quad (32)$$

$$= \arg \max_{\theta} \sum_{i=1}^n \left(-\frac{x_i^2}{2\theta} - \frac{1}{2} \log \theta \right). \quad (33)$$

To find the value of theta that maximizes the expression above, we can just look for the points where the derivative is zero by solving

$$\frac{\partial}{\partial \theta} \sum_{i=1}^n \left(-\frac{x_i^2}{2\theta} - \frac{1}{2} \log \theta \right) = -\frac{n}{2\theta} + \sum_{i=1}^n \frac{x_i^2}{2\theta^2} = 0. \quad (34)$$

This yields

$$\hat{\theta}_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^n x_i^2 \quad (35)$$

meaning that the best way of estimating the variance for a Gaussian r.v. with known mean is simply averaging the squared values of the observations.

(b) See figure 4.

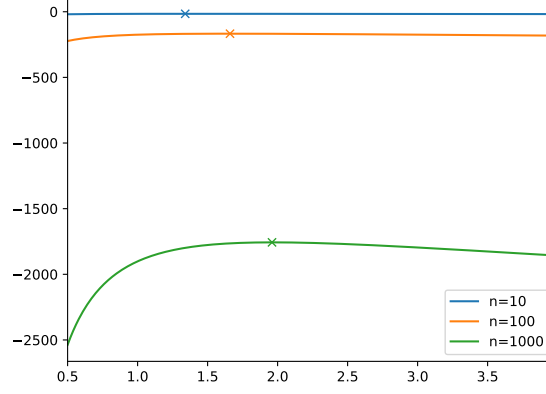


Figure 4: Results obtained for a single realization of the dataset for $n = 10, 100, 1000$.

(c) On average, the estimation $\hat{\theta}_{\text{MLE}}(n)$ for $n = 10, 100, 1000$ all approach the correct value θ^* of the variance, as shown in figure 5. However, the dispersion of the the predictions is lower for a bigger value of n , which follows the intuition that more data points imply more confidence in the estimation.

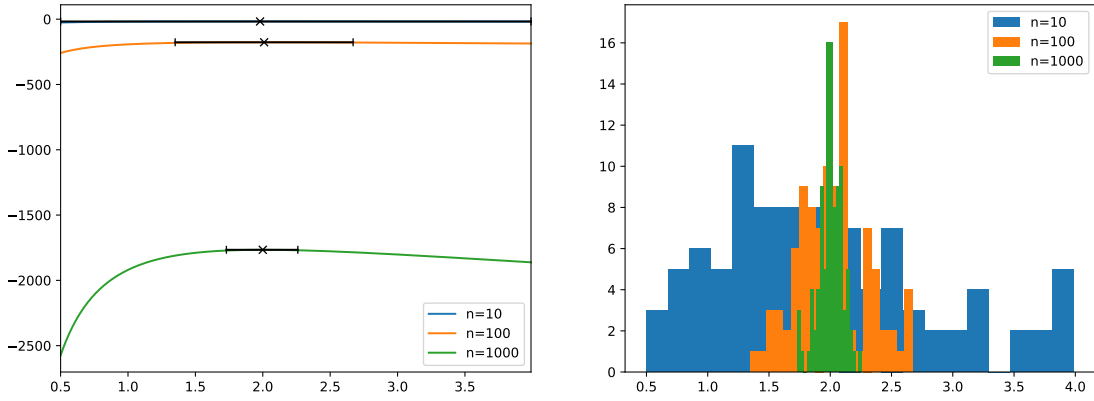


Figure 5: Results obtained averaging 100 dataset realizations for $n = 10, 100, 1000$. The histogram on the right shows the distribution of the estimated $\hat{\theta}_{\text{MLE}}(n)$.

(d) The Fisher information is

$$I(\theta) = -\mathbb{E} \left[\frac{d^2}{d\theta^2} \log p_\theta(X) \right] = -\mathbb{E} \left[\frac{d}{d\theta} \left(-\frac{1}{2\theta} + \frac{X^2}{2\theta^2} \right) \right] \quad (36)$$

$$= -\mathbb{E} \left[\frac{1}{2\theta^2} - \frac{X^2}{\theta^3} \right] = \frac{\mathbb{E}[X^2]}{\theta^3} - \frac{1}{2\theta^2} \quad (37)$$

$$= \frac{\theta}{\theta^3} - \frac{1}{2\theta^2} = \frac{1}{2\theta} \quad (38)$$

2 Exercise 4

(a) Given the logistic distribution

$$p_\theta(x) = \frac{e^{-(x-\mu)/s}}{s(1 + e^{-(x-\mu)/s})^2} \quad (39)$$

the log-likelihood for a dataset $\mathcal{D} = \{x_1, \dots, x_n\}$ of observations is

$$\log \mathcal{L}(\theta; \mathcal{D}) = \sum_{i=1}^n \log \left(\frac{e^{-(x_i - \mu)/s}}{s(1 + e^{-(x_i - \mu)/s})^2} \right) \quad (40)$$

$$= \sum_{i=1}^n \left(\frac{\mu - x_i}{s} - \log s - 2 \log(1 + z_i) \right) \quad (41)$$

where we substitute $z_i = e^{-(x_i - \mu)/s}$ to simplify the calculations. Of course we need to remember that z_i is a function of both μ and s , meaning that we need to use the chain rule when computing the partial derivatives. The partial derivatives of z_i w.r.t. the parameters μ and s are

$$\frac{\partial z_i}{\partial \mu} = \frac{1}{s} e^{-(x_i - \mu)/s} = \frac{z_i}{s} \quad \text{and} \quad \frac{\partial z_i}{\partial s} = \frac{x_i - \mu}{s^2} e^{-(x_i - \mu)/s} = \frac{x_i - \mu}{s^2} z_i \quad (42)$$

The partial derivatives of the log-likelihood w.r.t. the parameters μ and s are

$$\frac{\partial}{\partial \mu} \log \mathcal{L}(\theta; \mathcal{D}) = \sum_{i=1}^n \left(\frac{1}{s} - \frac{2}{1 + z_i} \cdot \frac{\partial z_i}{\partial \mu} \right) \quad (43)$$

$$= \frac{n}{s} - \sum_{i=1}^n \frac{2z_i}{s(1 + z_i)} \quad (44)$$

$$= \frac{n}{s} - \sum_{i=1}^n \frac{2e^{-(x_i - \mu)/s}}{s(1 + e^{-(x_i - \mu)/s})} \quad (45)$$

and

$$\frac{\partial}{\partial s} \log \mathcal{L}(\theta; \mathcal{D}) = \sum_{i=1}^n \left(\frac{x_i - \mu}{s^2} - \frac{1}{s} - \frac{2}{1 + z_i} \cdot \frac{\partial z_i}{\partial s} \right) \quad (46)$$

$$= \sum_{i=1}^n \left(\frac{x_i - \mu - s}{s^2} - \frac{2z_i(x_i - \mu)}{s^2(1 + z_i)} \right) \quad (47)$$

$$= \frac{1}{s^2} \left(-n(\mu + s) + \sum_{i=1}^n \left(x_i - \frac{2z_i(x_i - \mu)}{1 + z_i} \right) \right) \quad (48)$$

$$= \frac{1}{s^2} \left(-n(\mu + s) + \sum_{i=1}^n \left(x_i - \frac{2e^{-(x_i - \mu)/s}(x_i - \mu)}{1 + e^{-(x_i - \mu)/s}} \right) \right) \quad (49)$$

respectively. In the code implementation, it may actually be more convenient to use the expressions in eq. 44 and 48.

(b) See figure 6.

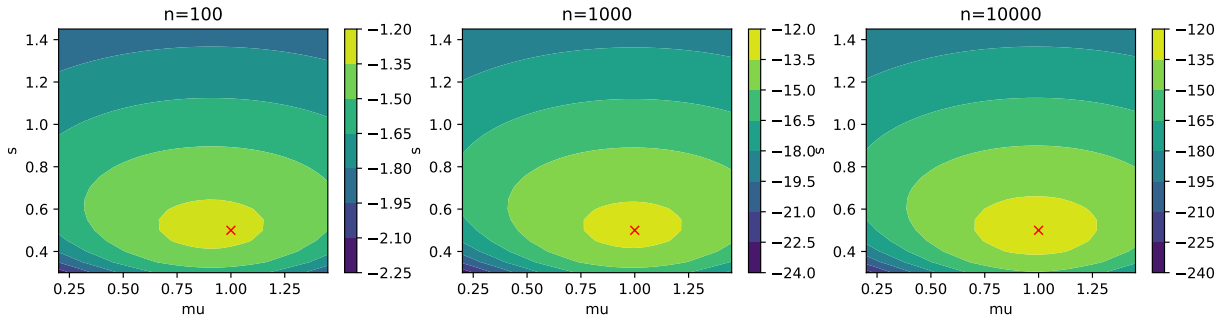


Figure 6: Contour plot of the log-likelihood for $n = 100, 1000, 10^4$. Brighter colors indicate a higher likelihood. The ground truth μ^*, σ^* is always contained in the high-likelihood region.

(c) See table 1.

n	$\hat{\mu}$	\hat{s}	MSE
100	0.9060	0.5091	4.4597×10^{-3}
1000	0.9853	0.5077	1.3768×10^{-4}
10^4	0.9837	0.4945	1.4733×10^{-4}

Table 1: Estimated values of $\hat{\mu}$ and \hat{s} using gradient ascent with learning rate $\eta = 0.1/n$ for datasets of size $n = 100, 1000, 10^4$. As expected, the prediction is less accurate for $n = 100$ in terms of MSE. The results for $n = 1000$ and 10^4 are comparable, meaning that 1000 samples are enough to get an accurate estimation of the parameters.

Exercise 5

- (a) The log-likelihood of a Bernoulli mixture with parameters π, μ for a dataset of observations $\mathcal{D} = \{x_1, \dots, x_n\}$ is

$$\log \mathcal{L}(\mathcal{D}; \pi, \mu) = \sum_{i=1}^n \log \left(\sum_{k=1}^K \pi_k \prod_{j=1}^d \mu_{kj}^{x_{ij}} (1 - \mu_{kj})^{(1-x_{ij})} \right) \quad (50)$$

not providing any simplification. The log-likelihood of each multivariate Bernoulli with parameter μ_k , instead, is

$$\log \mathcal{L}_k(\mathcal{D}; \mu_k) = \sum_{i=1}^n \sum_{j=1}^d (x_{ij} \log \mu_{kj} + (1 - x_{ij}) \log(1 - \mu_{kj})). \quad (51)$$

This suggests that the likelihood of a Bernoulli mixture model (BMM) can be maximized using the EM algorithm.

- (b) Similarly to the GMM case, let's consider a latent variable Z . Each realization of Z is a K -dimensional vectors z_i with a single element equal to 1 and all the others being 0. E.g., for $K = 6$ a realization of Z may look like $[0, 0, 1, 0, 0, 0]^\top$. The one element with value 1 is chosen according to $\pi = [\pi_1, \dots, \pi_K]^\top$.

If we assume that our dataset can be modeled by a BMM, that means that we can pair each data point x_i with a z_i , representing the multivariate Bernoulli from which x_i was sampled.

Let's focus on a single sample x_i for the moment. The expected log-likelihood for the pair (x_i, Z_i) is

$$\mathbb{E}[\log p_{X,Z|\pi,\mu}(x_i, Z_i)] = \mathbb{E}[\log p_{X,Z|\pi,\mu}(x_i, Z_i)] \quad (52)$$

$$= \mathbb{E} \left[\log \left(\prod_{k=1}^K (\pi_k \cdot p_{X|Z,\pi,\mu}(x_i))^{Z_{ik}} \right) \right] \quad (53)$$

$$= \mathbb{E} \left[\sum_{k=1}^K Z_{ik} \log(\pi_k \cdot p_{X|Z,\pi,\mu}(x_i)) \right] \quad (54)$$

$$= \sum_{k=1}^K \mathbb{E}[Z_{ik}] \log(\pi_k \cdot p_{X|Z,\pi,\mu}(x_i)) \quad (55)$$

$$= \sum_{k=1}^K \gamma_{ik} \log(\pi_k \cdot p_{X|Z,\pi,\mu}(x_i)) \quad (56)$$

and for the entire dataset becomes

$$\mathbb{E}_{\mathcal{D}_Z}[\log \mathcal{L}(\mathcal{D}, \mathcal{D}_Z; \pi, \mu)] = \sum_{i=1}^n \sum_{k=1}^K \gamma_{ik} \left(\log(\pi_k) + \sum_{j=1}^d (x_{ij} \log \mu_{kj} + (1 - x_{ij}) \log(1 - \mu_{kj})) \right). \quad (57)$$

The values of γ_{ik} represent the responsibilities of each multivariate Bernoulli for each data point x_i . In other words, they represent “how much” each component of the mixture contributes to the log-likelihood of x_i . These are computed as

$$\gamma_{ik} = \mathbb{E}[Z_{ik}] = \frac{\pi_k p_{X|\mu_k}(x)}{\sum_{k'=1}^K \pi_{k'} p_{X|\mu_{k'}}(x)} \quad (58)$$

exactly as in the GMM case. Once we have computed the responsibilities, we can use them to update the parameters as follows

$$\pi_k = \frac{n_k}{n}, \quad \mu_k = \frac{1}{n_k} \sum_{i=1}^n \gamma_{ik} x_i, \quad (59)$$

with $n_k = \sum_{i=1}^n \gamma_{ik}$. Notice that n_k represent the overall contribution of the k th Bernoulli to the dataset. Intuitively, π_k is proportional to such contribution. The parameter vector μ_k for the k th Bernoulli is updated by averaging the samples, which are weighted based on how likely they are to come from that component.

- (c) Figure 7 shows an example of image produced by the BMM, which was “trained” on the 0 digits of the MNIST dataset. Unfortunately, the result is more similar to a frog than to a zero, but at least the model correctly learns to insert white spaces in the center and on the borders of the figure.

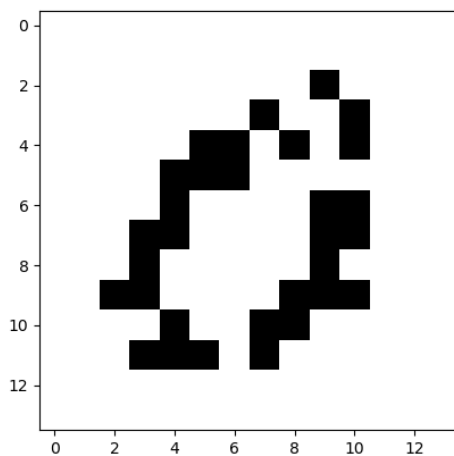


Figure 7: Frog.

Appendix

Proof: The sum of two independent Gaussians is Gaussian The PDF of $Y = X_1 + X_2$ with $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ can be computed as the convolution of the PDFs

$$p_Y(y) = \int_{-\infty}^{\infty} p_{X_1}(x)p_{X_2}(y-x)dx. \quad (60)$$

Computing this integral is feasible but needlessly long and prone to calculation mistakes. A better approach here is to use the characteristic functions $\varphi_{X_1}(\omega)$ and $\varphi_{X_2}(\omega)$, which is just how statisticians call the Fourier transforms of PDFs evaluated at $-\omega$. In the Fourier domain, convolutions become simple products, so that should be easier. The characteristic function of a Gaussian r.v. $\mathcal{N}(\mu, \sigma^2)$ is

$$\varphi_X(\omega) = \mathbb{E}[\exp(j\omega X)] \quad (61)$$

$$= \int_{-\infty}^{\infty} p_X(x) \exp(j\omega x) dx \quad (62)$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \exp(j\omega x) dx \quad (63)$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2 - 2\mu x + \mu^2}{2\sigma^2}\right) \exp(j\omega x) dx \quad (64)$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2 - 2\mu x + \mu^2 - 2j\sigma^2\omega x}{2\sigma^2}\right) dx \quad (65)$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2 - 2(\mu + j\sigma^2\omega)x + \mu^2 + 2j\mu\sigma^2\omega - \sigma^4\omega^2 - 2j\mu\sigma^2\omega + \sigma^4\omega^2}{2\sigma^2}\right) dx \quad (66)$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp(j\mu\omega - \frac{1}{2}\sigma^2\omega^2) \int_{-\infty}^{\infty} \exp\left(-\frac{x^2 - 2(\mu + j\sigma^2\omega)x + (\mu + j\sigma^2\omega)^2}{2\sigma^2}\right) dx \quad (67)$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp(j\mu\omega - \frac{1}{2}\sigma^2\omega^2) \int_{-\infty}^{\infty} \exp\left(-\frac{(x - \mu - j\sigma^2\omega)^2}{2\sigma^2}\right) dx \quad (68)$$

$$= \exp(j\mu\omega - \frac{1}{2}\sigma^2\omega^2) \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu - j\sigma^2\omega)^2}{2\sigma^2}\right) dx}_1 \quad (69)$$

$$= \exp(j\mu\omega - \frac{1}{2}\sigma^2\omega^2). \quad (70)$$

Step 66 is the well-known “completing the square” technique. The idea is to add and subtract the same quantity so that the polynomial in x inside the integral becomes a perfect square (i.e., $(x - \mu - j\sigma^2\omega)^2$). On the last step, we take advantage of this square form, since it becomes the integral in $(-\infty, \infty)$ of the PDF of a Gaussian r.v. $\mathcal{N}(\mu + j\sigma^2\omega, \sigma^2)$, which of course is 1. Now we can just multiply the characteristic function of $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ to obtain

$$\varphi_Y(\omega) = \exp(j\mu_1\omega - \frac{1}{2}\sigma_1^2\omega^2) \exp(j\mu_2\omega - \frac{1}{2}\sigma_2^2\omega^2) \quad (71)$$

$$= \exp(j\mu_1\omega - \frac{1}{2}\sigma_1^2\omega^2 + j\mu_2\omega - \frac{1}{2}\sigma_2^2\omega^2) \quad (72)$$

$$= \exp(j(\mu_1 + \mu_2)\omega - \frac{1}{2}(\sigma_1^2 + \sigma_2^2)\omega^2) \quad (73)$$

This is the characteristic function of a Gaussian r.v. with mean $\mu_1 + \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$, so we can conclude that the sum of two independent Gaussians is indeed Gaussian \square .