Further Exercises

12.1 Maximum Likelihood (4 points)

Suppose we are given a data set $x^{(1)}, \ldots, x^{(N)}$ representing p *iid* (independent and identically distributed) observations of the scalar random variable X, which follow a Gaussian distribution:

$$p(x) \sim \mathcal{N}(\mu, \sigma^2)$$

(a) Find the *maximum likelihood* estimates μ_{ML} and σ_{ML} of the distribution parameters.

(b) Show that μ_{ML} is *unbiased* and σ_{ML} is *biased*. Next, replace μ_{ML} with the true value μ in the maximum likelihood estimator σ_{ML} and verify that the resulting estimator $\hat{\sigma}_{ML}$ is *unbiased*.

Bonus Question (+2 points):
Find the *maximum a posteriori* estimate of the mean μ_{MAP} by applying Bayes' theorem, given the following prior distribution:

$$p(\mu) \sim \mathcal{N}(0, \sigma_{\text{pr}}^2)$$

12.2 Inverse CDF and Random Number Generation (4 points)

Background: If $F_X(x)$ is the cumulative distribution function (cdf) of a random variable X, then the random variable $Z = F_X(X)$ is uniformly distributed on the interval $[0, 1]$. This result provides a general receipe to generate samples $\tilde{x}$ of a random variable X with a desired probability density function (pdf) $p_X(x)$ from uniformly distributed random numbers $\tilde{z} \in [0, 1]$:

1. Compute the cdf $F_X(x)$ of the desired pdf $p_X(x)$
2. Determine the inverse transformation F^{-1}.
3. Sample uniformly distributed numbers (in $[0, 1]$), $\tilde{z}$.
4. Get the samples $\tilde{x} = F^{-1}(\tilde{z})$ from X.

The pdf of a Laplace distribution with location parameter μ (= mean), and scale parameter $b > 0$ (variance = $2b^2$) is given by

$$p_X(x) = \frac{1}{2b} \exp\left(-\frac{|x - \mu|}{b}\right).$$

Task:

(a) Following the procedure above, derive a formula to generate samples of a scalar random variable with a Laplacian distribution from uniformly distributed random numbers.

(b) Implement your procedure for verification and generate 500 samples for a Laplacian random variable X with a specific mean $\mu = 1$ and scale parameter $b = 2$. Plot a density estimate (e.g. histogram, ecdf) for these samples overlayed with the pdf $p_X(x)$ from above.

12.3 Density Transformations (6 points)

Background: Let $f(\mathbf{x}) = f(x_1, \ldots, x_n)$ be a function of $\mathbf{x} \in \Omega \subset \mathbb{R}^n$ and assume we make a change of variables to a new coordinate system by a mapping $\mathbf{u} = \mathbf{u}(\mathbf{x}) = (u_1(\mathbf{x}), \ldots, u_n(\mathbf{x}))$, whose inverse mapping $\mathbf{x} = \mathbf{x}(\mathbf{u}) = (x_1(\mathbf{u}), \ldots, x_n(\mathbf{u}))$ exists and is differentiable. As we change the coordinate system, the integral over f changes according to

$$\int_\Omega f(\mathbf{x})\mathbf{dx} = \int_{u(\Omega)} f(\mathbf{x}(\mathbf{u})) \left|\det \frac{\partial \mathbf{x}(\mathbf{u})}{\partial \mathbf{u}}\right| \mathbf{du} = \int_{u(\Omega)} f(\mathbf{x}(\mathbf{u})) \frac{1}{\left|\det \frac{\partial \mathbf{u}(\mathbf{x})}{\partial \mathbf{x}}\right|} \mathbf{du},$$

where $\frac{\partial \mathbf{x}(\mathbf{u})}{\partial \mathbf{u}}$ is the Jacobi matrix, which is the matrix of the partial derivatives

$$\frac{\partial \mathbf{x}(\mathbf{u})}{\partial \mathbf{u}} = \begin{pmatrix} \frac{\partial x_1(\mathbf{u})}{\partial u_1} & \cdots & \frac{\partial x_1(\mathbf{u})}{\partial u_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_n(\mathbf{u})}{\partial u_1} & \cdots & \frac{\partial x_n(\mathbf{u})}{\partial u_n} \end{pmatrix}$$

and whose determinant $\det \frac{\partial \mathbf{x}(\mathbf{u})}{\partial \mathbf{u}} = (\det \frac{\partial \mathbf{u}(\mathbf{x})}{\partial \mathbf{x}})^{-1}$ is called the *Jacobi determinant* (also *functional determinant*).
Remark: The absolute value of the Jacobi determinant at a point $\mathbf{u}_0$ corresponds to the factor by which the function $\mathbf{x}(\mathbf{u})$ expands or shrinks volumes near $\mathbf{u}_0$.

Implication: If $f(\mathbf{x})$ is the probability density function (pdf) of the n-dimensional random vector $\mathbf{X}$ then $f(\mathbf{x}(\mathbf{u})) \left|\det \frac{\partial \mathbf{x}(\mathbf{u})}{\partial \mathbf{u}}\right|$ is the pdf of the random vector $\mathbf{u}(\mathbf{X})$.

Task:

(a) Consider the density of a random variable X to be $p_X(x) = e^{-x},\ x \geq 0$. For the change of variables $u = u(x) = e^{-x}$ calculate the density $p_{u(X)}(u)$ of the random variable $u(X)$.

(b) Consider two independent and uniformly in the interval $[0, 1]$ distributed random variables $(X_1, X_2)^T =: \mathbf{X}$. The pdf is given by $p_{\mathbf{X}}(x_1, x_2) = 1$ in $[0, 1]^2$ and zero otherwise. Consider the variable transformation $\mathbf{u} = \mathbf{u}(\mathbf{x})$ with $u_1(\mathbf{x}) = \sqrt{-2 \log x_1} \cos(2\pi x_2)$ and $u_2(\mathbf{x}) = \sqrt{-2 \log x_1} \sin(2\pi x_2)$. Show that $\mathbf{u}(\mathbf{X})$ corresponds to two independent unit-variance zero-mean normally distributed random variables. *Remark:* This procedure to produce Gaussian samples from uniform random numbers is called the Box-Muller method.

(c) Outline how to extend the last result to n dimensions, i.e., how to generate samples from a multidimensional Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ just from uniformly distributed random numbers in $[0, 1]^n$. Use the following:

- Any symmetric positive semidefinite matrix (such as the covariance matrix $\boldsymbol{\Sigma}$) has a Cholesky decomposition $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^T$ (and that can be easily computed numerically).
- If $\mathbf{L}$ is a constant matrix and $\mathbf{X}$ a random vector then $\mathrm{Cov}(\mathbf{L}\mathbf{X}) = \mathbf{L}\,\mathrm{Cov}(\mathbf{X})\,\mathbf{L}^T$.
- The covariance matrix of independent unit-variance Gaussian variables is identity, i.e., $\mathrm{Cov}(\mathbf{X}) = \mathbf{I}$.

12.4 K-means and other clustering techniques (6 points)

The file clusters.zip contains 4 data sets which you should analyse using available packages for k-means and hierarchical (connectivity based) clustering (e.g. `fastcluster`).

(a) Use the dataset `stripes2.csv`. Apply k-means to find the two clusters. Plot the data using e.g. color to indicate cluster assignment.

(b) Use the dataset `ring.csv`. Apply k-means to find the clusters. What is the problem? Apply a linkage-based algorithm. Plot the dendrogram to decide where to "cut", i.e. how to assign the datapoints to clusters. Plot the data using e.g. color to indicate cluster assignment.

(c) Use the dataset `stripes3.csv`. Apply k-means to find the clusters. What is the problem? Apply a linkage-based algorithm. Plot the dendrogram and the cluster-sizes. Plot the data using e.g. color to indicate cluster assignment.

(d) Use the dataset `3d.csv`. Plot the data using a scatterplot matrix or a 3d plot. Cluster the data using a method of your choice. Plot the data using e.g. color to indicate cluster assignment. Justify your choice and describe the solution.

12.5 Properties of the KL-Divergence (3 points)

The Kullback-Leibler (KL) divergence can be used to measure the similarity of two probability densities or probability mass functions.

For a discrete random variable X taking values from the set $\mathcal{X} = \{x_1, ..., x_n\}$, the Kullback-Leibler (KL) divergence between two probability mass functions $p(X)$ and $q(X)$ having the support set $\mathcal{X}$ is defined as

$$D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \tag{1}$$

A real-valued function f is called *convex* if for any two points x and y and any $\lambda \in [0, 1]$

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda) f(y). \tag{2}$$

Jensen's inequality states that for a random variable x and a convex function $f(x)$

$$f(\mathbb{E}[x]) \leq \mathbb{E}[f(x)]$$

where $\mathbb{E}[.]$ denotes the expectation.

(a) **Nonnegativity of D_{KL}:** Use Jensen's inequality to show that the KL divergence is non-negative.

(b) **D_{KL} is not symmetric:** Show that the KL divergence is not symmetric by finding an example of two distributions p and q for which $D(p||q)$ is not equal to $D(q||p)$. Give the two distributions and show the explicit computation of the KL divergences. Simple toy distributions (e.g. over heads, tail) are fine.

(c) Is D_{KL} a metric? If not, why?

12.6 FastICA: Toy Signal Separation (4 points)

Generate three signals as row vectors given at time points $t = 0,\ 0.05, \ldots, 50$ by

$$\begin{aligned} s_1(t) &= 4\ \sin(t-3) \\ s_2(t) &= (t+5)\ \bmod 10 \\ s_3(t) &= \begin{cases} -14 & \text{if } \cos(2t) > 0 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

(a) Mix the signals to get $\mathbf{x} = \mathbf{A}\mathbf{s}$ where

$$\mathbf{A} = \begin{pmatrix} 2 & -3 & -4 \\ 7 & 5 & 1 \\ -4 & 7 & 5 \end{pmatrix}$$

(b) Whiten and separate the mixed (observed) signals $\mathbf{x}$ using `fastICA`.

(c) Plot the original source signals, mixtures, whitened mixtures, and unmixed signals.

(d) Repeat the same analysis using a different contrast function `G`.

(e) Repeat the same analysis using a matrix $\mathbf{A}$ which is closer to singular than the matrix above.

(f) Repeat the same analysis using an additive zero-mean Gaussian noise $\mathbf{n}$ on top of the mixture, i.e., $\mathbf{x} = \mathbf{A}\mathbf{s} + \mathbf{n}$. Try different variances of $\mathbf{n}$ (try first small variances, in another run a larger one etc $\rightsquigarrow$ how robust is the method?).

Properties of Differential Entropy and Negentropy

12.7 Differential entropy is not scale invariant (3 points)

The entropy of a random variable $\mathbf{x}$ with probability density $p(\mathbf{x})$ is defined as

$$H(\mathbf{x}) = -\int_S p(\mathbf{x}) \log\ p(\mathbf{x}) dx$$

where S is the support set of $\mathbf{x}$. Consider an invertible transformation $\mathbf{y} = \mathbf{g}(\mathbf{x})$. Using the Jacobi determinant, find the relation between $H(\mathbf{y})$ and $H(\mathbf{x})$. Use this to show that the entropy is not scale invariant, i.e. $H(a\mathbf{x}) \neq H(\mathbf{x})$, for $a = const.$

12.8 Differential entropy of the multivariate Gaussian (3 points)

Show that the entropy of a multivariate n-dimensional Gaussian random vector $\mathbf{x}$ with covariance matrix $\mathbf{\Sigma}$ has the form

$$H(\mathbf{x}_{Gauss}) = \frac{1}{2}\ \log\ |\det\ \mathbf{\Sigma}| + \frac{n}{2}\ (1 + \log\ 2\pi)$$

12.9 Negentropy is scale invariant (4 points)

The negentropy is defined as

$$J(\mathbf{x}) = H(\mathbf{x}_{Gauss}) - H(\mathbf{x})$$

where $\mathbf{x}_{Gauss}$ is a multivariate Gaussian with the same covariance matrix as $\mathbf{x}$. Show that the negentropy is invariant wrt. invertible linear transformations $\mathbf{y} = \mathbf{A}\mathbf{x}$, i.e.

$$J(\mathbf{A}\mathbf{x}) = J(\mathbf{x})$$

from which it follows that the negentropy is scale-invariant.