May 26, 2026

Second Intermediate test - Signal processing and information theory

Question A

During the lessons, it was illustrated how a spectral analysis of the sequence of nucleotides (or bases) obtained by DNA sequencing can be helpful in identifying the location of its coding regions (the exons).

1. Why can this analysis be successful?
2. How is this short-term spectral analysis conducted?
3. What are the advantages of taking a filtering-based approach, rather than performing short-term spectral analysis?

Question B

What is a simple method to remove periodic noise from an image signal?

Question C

1. What property of the two-dimensional Fourier transform is used to reconstruct a tomographic image from a sinogram?
2. Why does the simple back-projection method produce very blurry results?
3. What solution to the previous problem was illustrated in class, and why does it work?

Question D

1. How is the amount of information received from knowing the realization of a discrete random variable defined?
2. How is the average information emitted by a discrete, memoryless random source defined, what name is it given, and what extreme values can it assume?
3. How does the previous quantity vary if the symbols emitted by the source are not statistically independent?
4. What kind of limit does this quantity place on the binary rate produced by a source encoder?
5. What are the main divergent aspects when applying the same concepts to a continuous information source?

Question E

Let us have two binary random variables X and Y, whose joint pdf $p(x, y)$ is expressed by the table

	y_0	y_1
x_0	0.13	0.61
x_1	0.24	0.02

, so that it turns out that $p_{ij} = p\left(x_i, y_j\right)$.

1. Calculate the entropy values $H(X)$ and $H(X)$ of both the r.v. You can perform the calculations manually, or approximate the values by referring to the figure for the entropy of binary sources, or invoke the `entropy` function of `scipy.stats`. *(The latter is also the recommended method for continuing the exercise - if this is the case, please attach the code)*

2. Calculate the value of the joint entropy $H(X, Y)$ and verify the inequality reported at page 25 of the 9^{th} slide set

3. Derive the conditional entropy values $H(X/Y)$ and $H(Y/X)$ from the quantities already obtained. Which of the two r.v. provides the greatest reduction in uncertainty?

4. Now derive the values of the average mutual information $I(X;Y)$, again starting from the quantities already obtained. If the one described were a communication channel, how many binary symbols would be needed to transfer at least one bit of information?

5. Is there a way to calculate $I(X;Y)$ directly from the pdf (*formula (3) of the* 9^{th} *slide set*) using Python's `entropy` function? If so, try implementing it and verify the accuracy of the result obtained at the previous point.

Question F

1. What is meant by channel capacity?

2. What does it depend on solely?

3. What constraints does it impose on the information source?

'Official' answers

Question A

1. Because it has been observed that in the areas corresponding to the exons, the same base tends to always repeat itself in the same position (first, second or third) within the codons that make up the exons. This determines the presence of a periodicity of length three in the composition of the exons, which can be detected by means of a frequency analysis.

2. Starting from the base sequence, four binary indicator sequences are constructed, each indicating the presence (or absence) of each base (A, T, C, G). Partially overlapping time windows are then isolated from each indicator sequence, each containing a multiple of three bases. Typically, the sequences have 351 elements or more, as excessively long windows would compromise the spatial resolution of the result. A DFT is calculated for each window, and from these four sequences of spectral coefficients, a single one is obtained by adding their squares, element by element. If the window lies within a coding region, a spectral peak is observed near the $2/3\pi$ digital frequency.

3. A filtering approach involves designing a digital filter fed by an indicator sequence, with a bandpass frequency response centered on the digital frequency of $2/3\pi$. Since this spectral content is present only in the coding regions, higher output values will be observed in these regions. The advantage of this approach lies in the possibility of designing the digital filter with a particularly narrow bandwidth, thus increasing the sensitivity of the measurement; in fact, using an FFT is equivalent to observing the output of a filter bank, whose frequency response depends solely on the duration of the analysis window.

Question B

One can work on the two-dimensional DFT of the image. The presence of periodic noise is reflected in the appearance of spatial harmonics with frequencies multiples of the fundamental period, arranged along the direction(s) in which the periodicity evolves. The magnitude of the 2D-DFT coefficients arranged according to this pattern is then attenuated, or masked, to greatly reduce its effect. An inverse 2D-DFT is then performed to obtain the original image (almost) devoid of the original periodic component.

Question C

1. The result of the Fourier slice theorem is used, which states that the 1D Fourier transform of the projection $p(t, \theta)$ of the absorption coefficient $\mu(x, y)$ (which is what you get from the sinogram) *is equal* to a slice $P(f, \theta)$, passing through the origin and with the same angle θ, of $M(u, v)$, which in turn is the 2D Fourier transform of $\mu(x, y)$, so that

   (a) the tomographic cross-section $\mu(x, y)$ can be retrieved by adding all the slices $P(f, \theta)$ for any θ, so that $M(u, v)$ is obtained, and then get back to $\mu(x, y) = \mathcal{F}_{2D}^{-1}\{M(u, v)\}$

2. The solution illustrated in the previous point suffers from the same problem as the one obtained with the back-projection method, which was intuitively described in class through explanatory videos. However, the frequency perspective allows us to identify the nature of the problem: both methods exaggerate the low spatial frequencies, while the higher-frequency regions, necessary to reproduce the finest details, are underrepresented.

3. The solution illustrated consists of applying to $P(f, \theta)$ (the DFT of $p(t, \theta)$) a filter characterized by a frequency response $H(f) = |f|$, called a *ramp filter*. Recalling the basic properties of a Fourier transform, the result is a derivative in the inverse (spatial) domain, i.e. for $p(t, \theta)$.

   The resulting method is called *filtered backprojection*, in which, instead of backprojecting the $p(t, \theta)$ obtained from the sinogram, its filtered version (i.e., after the IDFT of $P(f, \theta) H(f)$) is backprojected.

   In reality, the resulting images now appear excessively noisy. A white-spectrum image noise, in fact, has a significant portion of power at higher frequencies, while an image has more low-frequency spectral content. Therefore, rather than a ramp filter, it is preferable to use one whose frequency response is more bandpass-like rather than a pure derivative.

Question D

1. Knowledge of the symbol x_k provides an *amount of information* defined as $I_k = I(x_k) = \log_2 \frac{1}{p_k} = -\log_2 p_k$ bits where p_k is its probability

2. The average information emitted by a discrete, memoryless random source is defined as

$$H = E\{I_k\} = \sum_{k=1}^{L} p_k I_k = \sum_{k=1}^{L} p_k \log_2 \frac{1}{p_k} = -\sum_{k=1}^{L} p_k \log_2 p_k \quad \text{bits/symbol}$$

   it is called Entropy, and it is bounded as $0 \leq H_s \leq \log_2 L$ in which L is the source alphabet size.

3. When the source symbols *are not* statistically independent, the expectation that defines the entropy is defined for a block of N symbols, its evaluation extended to all the possible values of the join pdf, and is indicated as H_N. As N increases H_N decreases, up to a point when all the memory of past events of the source has been *used*. If this happen, the source is called a Markov source.

4. To answer this question, you should have followed the link to the section of the book that discusses source coding, as we did in class, and possibly requested a translation. There, you'll find that it's impossible to build a binary encoder that uses an average number of binary symbols (called *binits* in this case) less than the entropy of the source - unless you accept a degradation of the message, such as missing parts of it.

5. Extending the concept of entropy to continuous sources leads to the definition of differential entropy, which diverges in terms of the r.v. dynamic, since $h(aX) = h(X) + \log|a|$, and consequently h can take on values that can be positive, negative or zero.

   This makes it impossible to use it as an absolute measure, but it is useful for comparing the average uncertainty between two continuous sources with the same variance. From this perspective, the Gaussian source has the maximum h, given the same variance.

Question E

1. I used the `entropy` function call (the code is below), that evaluates $\sum_k p_k * log\frac{1}{p_k}$, getting the result
   $H(X)$ = 0.8267463724926178 with a $p(x)$ = [0.74 0.26]
   $H(Y)$ = 0.950672092687066 with a $p(y)$ = [0.37 0.63]

2. Still using Python, I find
   $H(X,Y)$ = 1.4246582503299445 by using the joint $p(x,y)$ = [0.13 0.61 0.24 0.02]
   so that $\max\{H(X), H(Y)\} = 0.95 \leq \quad H(X,Y) = 1.42 \quad \leq H(X) + H(Y) = 1.776$ is verified

3. $H(Y/X)$ = 0.5979118778373267
   $H(X/Y)$ = 0.47398615764287855
   and therefore it is Y that provides the greatest reduction in uncertainty, since the residual uncertainty $H(X/Y) = H(X,Y) - H(Y)$ is smaller than $H(Y/X) = H(X,Y) - H(X)$. In fact, since $p(y)$ is more concentrated towards the value of 0.5 than $p(x)$, it turns out that $H(Y) > H(X)$

4. We can use the relationship $I(X;Y)$ = $H(X)$ + $H(Y)$ - $H(X,Y)$ and obtain
   $I(X;Y)$ = 0.3527602148497393
   If this were a communication channel, since $I(X;Y)$ is slightly larger than $\frac{1}{3}$, at least three binary symbols must be used to transmit one bit of information

5. Yes, it is possible to use the entropy function to calculate $I(X;Y)$ as well. Just call it by passing a second pdf $q(k)$ as an argument, so that the function calculates $\sum_k p_k * log\frac{p_k}{q_k}$ instead of $\sum_k p_k * log\frac{1}{p_k}$.
   Since $I(X;Y) = \sum_x \sum_y p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$ the $q(k)$ must be equal to the joint pdf evaluated as if X and Y would be statistically independent, or $q(k) = p(x)p(y)$ = [0.2738 0.4662 0.0962 0.1638]. By proceeding in this way, you get
   I(X;Y) = 0.3527602148497391
   which is equal (a part a final rounding) to the value obtained before

Python code used for Question E:

```
# adapted from https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.entropy.html
import numpy as np
```

```
from scipy.stats import entropy
base = 2 # work in units of bits
# marginal entropies
px = np.array([0.13+0.61, 0.24+0.02])
HX = entropy(px, base=base)
print ("H(X) = ", HX, " p(x) = ", px)
py = np.array([0.13+0.24, 0.61+0.02])
HY = entropy(py, base=base)
print ("H(Y) = ", HY, " p(y) = ", py)
# joint entropy
pxy = np.array([0.13, 0.61, 0.24, 0.02])
HXY = entropy(pxy, base=base)
print ("H(X,Y) = ", HXY, " p(x,y) = ", pxy)
# conditional entropies
print ("H(Y/X) = ", HXY - HX)
print ("H(X/Y) = ", HXY - HY)
# average mutual information
print ("I(X;Y) = ", HX + HY - HXY)
# Joint pdf for stat. indep. r.v.
pstind = np.array([px[0]*py[0], px[0]*py[1], px[1]*py[0], px[1]*py[1],])
# average mutual information, again
IXY = entropy(pxy, pstind, base=base)
print ("Evaluated I(X;Y) = ", IXY, " pstind = ", pstind)
```

Question F

1. The channel capacity C determines the maximum amount of information that can be transferred by a model known as a noisy channel. It is calculated as the maximum value that the average mutual information $I(X;Y)$ (between the messages emitted by the input source and those observed exiting the channel) can reach.

2. Since $I(X;Y)$ depends on both the forward conditional probabilities $p(y/x)$ of the channel and on the $p(x)$ describing the source, maximization is carried out by varying the latter. Therefore, capacity depends only on $p(y/x)$.

3. The Shannon's theorem for noisy channels asserts that information can (theoretically) be transmitted along a channel *without errors*, as long as the entropy of the source H_x at the channel input is less than the value of the channel capacity C
   If instead $H_x > C$, errors cannot be corrected, and their probability increases with H_x
   Since the theorem does not suggest how to find the optimal coding technique, nor does it distinguish between source and channel coding, it limits itself to indicating *the limit performances* obtainable with an *optimal* coding technique, that is, capable of reducing the probability of error p_e at will, provided that $H_x < C$