Assignment 1 (ML for TS) - MVA

Firstname Lastname youremail1@mail.com
Firstname Lastname youremail2@mail.com

November 4, 2025

1 Introduction

Objective. This assignment has three parts: questions about convolutional dictionary learning, spectral features, and a data study using the DTW.

Warning and advice.

- Use code from the tutorials as well as from other sources. Do not code yourself well-known procedures (e.g., cross-validation or k-means); use an existing implementation.
- The associated notebook contains some hints and several helper functions.
- Be concise. Answers are not expected to be longer than a few sentences (omitting calculations).

Instructions.

- Fill in your names and emails at the top of the document.
- Hand in your report (one per pair of students) by Sunday 9$^{\text{th}}$ November 23:59 PM.
- Rename your report and notebook as follows:
  `FirstnameLastname1_FirstnameLastname2.pdf` and
  `FirstnameLastname1_FirstnameLastname2.ipynb`.
  For instance, `LaurentOudre_ValerioGuerrini.pdf`.
- Upload your report (PDF file) and notebook (IPYNB file) using this link: LINK.

2 Convolution dictionary learning

Question 1

Consider the following Lasso regression:

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2} \|y - X\beta\|_2^2 \quad + \quad \lambda \|\beta\|_1 \tag{1}$$

where $y \in \mathbb{R}^n$ is the response vector, $X \in \mathbb{R}^{n \times p}$ the design matrix, $\beta \in \mathbb{R}^p$ the vector of regressors and $\lambda > 0$ the smoothing parameter.

Show that there exists $\lambda_{\max}$ such that the minimizer of (1) is $\mathbf{0}_p$ (a p-dimensional vector of zeros) for any $\lambda > \lambda_{\max}$.

Answer 1

1. The Primal Problem

The Lasso optimization problem is given by:

$$(P) \quad p^* = \min_{\beta \in \mathbb{R}^p} \quad \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

This is an unconstrained convex optimization problem.

2. Equivalent Constrained Formulation

To find the dual, we first introduce a new variable $z \in \mathbb{R}^n$ to represent the residual. We can rewrite the problem as an equivalent constrained optimization problem:

$$(P') \quad p^* = \min_{\beta \in \mathbb{R}^p, z \in \mathbb{R}^n} \quad \frac{1}{2}\|z\|_2^2 + \lambda\|\beta\|_1$$

$$\text{subject to} \quad z + X\beta = y$$

3. Strong Duality via Slater's Condition

This new problem (P') is a **convex optimization problem**:

- The objective function $f_0(\beta, z) = \frac{1}{2}\|z\|_2^2 + \lambda\|\beta\|_1$ is a sum of two convex functions (the squared L_2-norm and the L_1-norm, with $\lambda > 0$), and is therefore convex.
- The constraint $h(\beta, z) = -z - X\beta + y = 0$ is an **affine** constraint.

For a convex optimization problem with only affine constraints, a refinement of **Slater's condition** states that strong duality holds as long as the problem is feasible.

The problem is clearly feasible; for example, one can choose $\beta = 0_p$ (the p-dimensional zero vector) and $z = y$. This satisfies the constraint.

Because strong duality holds, the optimal value of the primal (p^*) is equal to the optimal value of the dual (d^*), and the KKT conditions are necessary and sufficient for optimality.

4. The Lagrangian and Dual Function

We introduce a Lagrange multiplier (dual variable) $\nu \in \mathbb{R}^n$ for the affine constraint. The Lagrangian L is:

$$L(\beta, z, \nu) = (\text{primal objective}) + \nu^T(\text{constraint})$$

$$L(\beta, z, \nu) = \frac{1}{2}\|z\|_2^2 + \lambda\|\beta\|_1 + \nu^T(y - X\beta - z)$$

The Lagrange dual function $g(\nu)$ is the infimum of the Lagrangian over the primal variables (β and z):

$$g(\nu) = \inf_{\beta, z} L(\beta, z, \nu)$$

We can group the terms for β and z and minimize them independently:

$$g(\nu) = \inf_{z \in \mathbb{R}^n} \left(\frac{1}{2}\|z\|_2^2 - \nu^T z\right) + \inf_{\beta \in \mathbb{R}^p} \left(\lambda\|\beta\|_1 - \nu^T X\beta\right) + y^T\nu$$

Minimization w.r.t. z: The function $L_z = \frac{1}{2}\|z\|_2^2 - \nu^T z$ is convex and differentiable. We find the minimum by setting its gradient to zero:

$$\nabla_z L_z = z - \nu = 0 \quad \Longrightarrow \quad z^* = \nu$$

The minimum value is $L_z(z^*) = \frac{1}{2}\|\nu\|_2^2 - \nu^T\nu = -\frac{1}{2}\|\nu\|_2^2$.

Minimization w.r.t. β: This is the conjugate function of the L_1-norm. Let $v = X^T\nu$.

$$\inf_{\beta\in\mathbb{R}^p}\left(\lambda\|\beta\|_1 - \nu^T X\beta\right) = -\sup_{\beta\in\mathbb{R}^p}\left(\left(\frac{1}{\lambda}X^T\nu\right)^T\lambda\beta - \|\lambda\beta\|_1\right)$$

Let's study the conjugate function of L_1-norm:

Let $x \in \mathbb{R}^n$. The ℓ_1 norm is

$$\|x\|_1 := \sum_{i=1}^n |x_i|.$$

For a function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ the *convex conjugate* (Fenchel conjugate) f^* is defined by

$$f^*(y) := \sup_{x\in\mathbb{R}^n}\{\langle y, x\rangle - f(x)\}, \qquad y \in \mathbb{R}^n,$$

where $\langle y, x\rangle = \sum_{i=1}^n y_i x_i$.

We will also use the dual norm fact: if $\|\cdot\|$ is a norm and $\|\cdot\|_*$ denotes its dual norm,

$$\|y\|_* = \sup_{\|x\|\le 1}\langle y, x\rangle,$$

then for the norm $f(x) = \|x\|$ one has

$$f^*(y) = \begin{cases} 0 & \text{if } \|y\|_\infty \le 1, \\ +\infty & \text{otherwise} \end{cases}$$

See apendix to see proof

If we inject our variable we get :

$$\inf_{\beta\in\mathbb{R}^p}\left(\lambda\|\beta\|_1 - \nu^T X\beta\right) = -f^*\left(\frac{X^T\nu}{\lambda}\right) = \begin{cases} 0 & \text{if } \|X^T\nu\|_\infty \le \lambda, \\ -\infty & \text{otherwise} \end{cases}$$

5. The Dual Problem

The dual problem (D) is to maximize $g(\nu)$. We can write this as a minimization problem by flipping the sign, which is a common convention:

$$(D) \quad d^* = \max_{\nu\in\mathbb{R}^n} \quad -\frac{1}{2}\|\nu\|_2^2 + y^T\nu \quad \text{subject to} \quad \|X^T\nu\|_\infty \le \lambda$$

6. Finding $\lambda_{\max}$

We want to find the value $\lambda_{\max}$ such that for any $\lambda > \lambda_{\max}$, the primal solution is $\beta^* = 0_p$.

If the primal solution is $\beta^* = 0_p$

Now we use the KKT optimality conditions. The stationarity condition (from minimizing the Lagrangian w.r.t. z) gave us the relationship:

$$z^* + \nu^* = 0 \quad \implies \quad \nu^* = -z^*$$

(Note: Using the other sign convention for the Lagrangian, $\nu^* = z^*$, leads to the same final result).

Substituting our primal solutions, the optimal dual variable must be:

$$\nu^* = -z^* = -y$$

For $\beta^* = 0_p$ to be the optimal solution, its corresponding dual partner $\nu^* = -y$ must be a **feasible solution for the dual problem (D)**.

The feasibility condition for the dual problem is its constraint:

$$\|X^T \nu^*\|_\infty \leq \lambda$$

We substitute $\nu^* = -y$ into this constraint:

$$\|X^T(-y)\|_\infty \leq \lambda$$

Since the L_∞-norm is absolute ($|-a| = |a|$), this is:

$$\|X^T y\|_\infty \leq \lambda$$

This is the condition on λ that ensures the solution is $\beta^* = 0_p$. If λ is greater than or equal to this value, the optimal solution will be the zero vector.

Therefore, the threshold $\lambda_{\max}$ is the smallest value for which this holds:

$$\lambda_{\max} = \|X^T y\|_\infty$$

This means $\lambda_{\max}$ is the largest (in absolute value) inner product between any single feature and the response vector y.

$$\lambda_{\max} = \dots \tag{2}$$

Question 2

For a univariate signal $\mathbf{x} \in \mathbb{R}^n$ with n samples, the convolutional dictionary learning task amounts to solving the following optimization problem:

$$\min_{(\mathbf{d}_k)_k, (\mathbf{z}_k)_k \|\mathbf{d}_k\|_2^2 \leq 1} \quad \left\| \mathbf{x} - \sum_{k=1}^{K} \mathbf{z}_k * \mathbf{d}_k \right\|_2^2 \quad + \quad \lambda \sum_{k=1}^{K} \|\mathbf{z}_k\|_1 \tag{3}$$

where $\mathbf{d}_k \in \mathbb{R}^L$ are the K dictionary atoms (patterns), $\mathbf{z}_k \in \mathbb{R}^{N-L+1}$ are activations signals, and $\lambda > 0$ is the smoothing parameter.

Show that

- for a fixed dictionary, the sparse coding problem is a lasso regression (explicit the response vector and the design matrix);
- for a fixed dictionary, there exists $\lambda_{\max}$ (which depends on the dictionary) such that the sparse codes are only 0 for any $\lambda > \lambda_{\max}$.

Answer 2

We fix all the dictionary atoms $\mathbf{d}_k \in \mathbb{R}^L$ for $k = 1, \dots, K$, and we note that $M = N - L + 1$ is the valid convolution length.

For a given pair $(\mathbf{z}_k, \mathbf{d}_k)$, the valid convolution $\mathbf{z}_k * \mathbf{d}_k \in \mathbb{R}^N$ can be written componentwise as

$$(\mathbf{z}_k * \mathbf{d}_k)[n] = \sum_{i=0}^{M-1} z_k[i]\, d_k[n-i], \quad \text{where } d_k[n-i] = 0 \text{ if } n - i < 0 \text{ or } n - i > L - 1.$$

This operation can be expressed as a matrix–vector product

$$\mathbf{y}_k = \mathbf{D}_k \mathbf{z}_k,$$

where $\mathbf{D}_k \in \mathbb{R}^{N \times M}$ has coefficients

$$\mathbf{D}_k[n, i] = \begin{cases} d_k[n-i], & \text{if } 0 \le n - i < L, \\ 0, & \text{otherwise.} \end{cases}$$

We define:

$$\mathbf{z} = \begin{bmatrix} \mathbf{z}_1 \\ \vdots \\ \mathbf{z}_K \end{bmatrix} \in \mathbb{R}^{K \cdot M}, \qquad \mathbf{D} = \begin{bmatrix} \mathbf{D}_1 & \mathbf{D}_2 & \cdots & \mathbf{D}_K \end{bmatrix} \in \mathbb{R}^{N \times K \cdot M}.$$

Hence we can then rewrite the sparse coding problem given the $(\mathbf{d}_k)$ in the LASSO form, with response vector $\mathbf{x} \in \mathbb{R}^N$, design matrix $\mathbf{D}$ and regressor vector $\mathbf{z}$:

$$\min_{\mathbf{z} \in \mathbb{R}^{K \cdot M}} \ \|\mathbf{x} - \mathbf{D}\mathbf{z}\|_2^2 \ + \ \lambda \|\mathbf{z}\|_1.$$

3 Spectral feature

Let X_n $(n = 0, \dots, N-1)$ be a weakly stationary random process with zero mean and autocovariance function $\gamma(\tau) := \mathbb{E}(X_n X_{n+\tau})$. Assume the autocovariances are absolutely summable, i.e. $\sum_{\tau \in \mathbb{Z}} |\gamma(\tau)| < \infty$, and square summable, i.e. $\sum_{\tau \in \mathbb{Z}} \gamma^2(\tau) < \infty$. Denote the sampling frequency by f_s, meaning that the index n corresponds to the time n/f_s. For simplicity, let N be even.

The *power spectrum* S of the stationary random process X is defined as the Fourier transform of the autocovariance function:

$$S(f) := \sum_{\tau=-\infty}^{+\infty} \gamma(\tau) e^{-2\pi f \tau / f_s}. \tag{4}$$

The power spectrum describes the distribution of power in the frequency space. Intuitively, large values of $S(f)$ indicate that the signal contains a sine wave at the frequency f. There are many

estimation procedures to determine this important quantity, which can then be used in a machine-learning pipeline. In the following, we discuss the large sample properties of simple estimation procedures and the relationship between the power spectrum and the autocorrelation.

(Hint: use the many results on quadratic forms of Gaussian random variables to limit the number of calculations.)

Question 3

In this question, let X_n $(n = 0, \dots, N-1)$ be a Gaussian white noise.

- Calculate the associated autocovariance function and power spectrum. (By analogy with the light, this process is called "white" because of the particular form of its power spectrum.)

Answer 3

Assume (X_n) is Gaussian white noise with variance σ^2, i.e. the X_n are i.i.d. $\mathcal{N}(0, \sigma^2)$.

Autocovariance

Independence gives for any integer lag τ:

$$\gamma(\tau) = \mathbb{E}[X_n X_{n+\tau}] = \begin{cases} \sigma^2, & \tau = 0, \\ 0, & \tau \neq 0. \end{cases}$$

Equivalently $\gamma(\tau) = \sigma^2 \mathbf{1}_{\{\tau=0\}}$ (Kronecker delta form).

Power spectrum

The power spectrum is the Fourier transform of $\gamma(\tau)$:

$$S(f) = \sum_{\tau=-\infty}^{+\infty} \gamma(\tau)\, e^{-2\pi i f \tau / f_s} = \sigma^2 \sum_{\tau=-\infty}^{+\infty} \mathbf{1}_{\{\tau=0\}} e^{-2\pi i f \tau / f_s} = \sigma^2,$$

i.e. the spectrum is *constant* (flat) across all frequencies. This constant spectrum is the reason the process is called "white" noise (analogy to white light which has equal power at all visible frequencies).

Question 4

A natural estimator for the autocorrelation function is the sample autocovariance

$$\hat{\gamma}(\tau) := (1/N) \sum_{n=0}^{N-\tau-1} X_n X_{n+\tau} \tag{5}$$

for $\tau = 0, 1, \dots, N-1$ and $\hat{\gamma}(\tau) := \hat{\gamma}(-\tau)$ for $\tau = -(N-1), \dots, -1$.

- Show that $\hat{\gamma}(\tau)$ is a biased estimator of $\gamma(\tau)$ but asymptotically unbiased. What would be a simple way to de-bias this estimator?

Answer 4

Bias of $\hat{\gamma}(\tau)$

Compute the expectation (for any stationary process with mean zero):

$$\mathbb{E}\big[\hat{\gamma}(\tau)\big] = \frac{1}{N}\sum_{n=0}^{N-\tau-1} \mathbb{E}[X_n X_{n+\tau}] = \frac{N-\tau}{N}\,\gamma(\tau).$$

Thus

$$\mathbb{E}\big[\hat{\gamma}(\tau)\big] = \Big(1 - \frac{\tau}{N}\Big)\,\gamma(\tau).$$

Therefore $\hat{\gamma}(\tau)$ is *biased* for finite N (unless $\gamma(\tau) = 0$ or $\tau = 0$). However, as $N \to \infty$ with fixed τ,

$$\mathbb{E}\big[\hat{\gamma}(\tau)\big] \longrightarrow \gamma(\tau),$$

so $\hat{\gamma}(\tau)$ is *asymptotically unbiased*.

A simple de-biasing (unbiased estimator)

A straightforward unbiased estimator is obtained by dividing by the number of terms $(N - \tau)$ instead of N:

$$\tilde{\gamma}(\tau) := \frac{1}{N-\tau}\sum_{n=0}^{N-\tau-1} X_n X_{n+\tau}.$$

Indeed,

$$\mathbb{E}\big[\tilde{\gamma}(\tau)\big] = \frac{1}{N-\tau}\sum_{n=0}^{N-\tau-1} \gamma(\tau) = \gamma(\tau).$$

So $\tilde{\gamma}(\tau)$ is unbiased for every finite N (for $0 \le \tau \le N-1$). This is the usual "$N - \tau$" normalization used when one wants an unbiased sample autocovariance.

Question 5

Define the discrete Fourier transform of the random process $\{X_n\}_n$ by

$$J(f) := (1/\sqrt{N})\sum_{n=0}^{N-1} X_n e^{-2\pi i f n/f_s} \tag{6}$$

The *periodogram* is the collection of values $|J(f_0)|^2, |J(f_1)|^2, \ldots, |J(f_{N/2})|^2$ where $f_k = f_s k/N$. (They can be efficiently computed using the Fast Fourier Transform.)

- Write $|J(f_k)|^2$ as a function of the sample autocovariances.
- For a frequency f, define $f^{(N)}$ the closest Fourier frequency f_k to f. Show that $|J(f^{(N)})|^2$ is an asymptotically unbiased estimator of $S(f)$ for $f > 0$.

Answer 5

$$|J(f_k)|^2 = J(f_k)\overline{J(f_k)} = (1/N)\sum_{n=0}^{N-1}\sum_{m=0}^{N-1} X_n X_m e^{-2\pi i f_k (n-m)/f_s}$$

Do the change of variable $\tau = m - n \in [-(N-1), N-1]$.

$$|J(f_k)|^2 = \frac{1}{N}\sum_{\tau=-(N-1)}^{N-1} e^{-2\pi \mathrm{i} f_k \tau/f_s} \sum_{n=0}^{N-1-|\tau|} X_n X_{n+|\tau|}$$

$$|J(f_k)|^2 = \sum_{\tau=-(N-1)}^{N-1} \hat{\gamma}(\tau)\, e^{-2\pi \mathrm{i} f_k \tau/f_s}.$$

For a frequency f, let $f^{(N)}$ be the closest Fourier frequency to f, with $k_n = \lfloor Nf/f_s \rfloor$.

$$\mathbb{E}[|J(f^{(N)})|^2] = \sum_{\tau=-(N-1)}^{N-1} \mathbb{E}[\hat{\gamma}(\tau)]\, e^{-2\pi \mathrm{i} f^{(N)} \tau/f_s}$$

$$\mathbb{E}[|J(f^{(N)})|^2] = \frac{N-|\tau|}{N}\sum_{\tau=-(N-1)}^{N-1} \gamma(\tau)\, e^{-2\pi \mathrm{i} f^{(N)} \tau/f_s}$$

TODO We assume the summability of the autocovariances and their square is sufficient to take the limit of N in the whole expression, to have

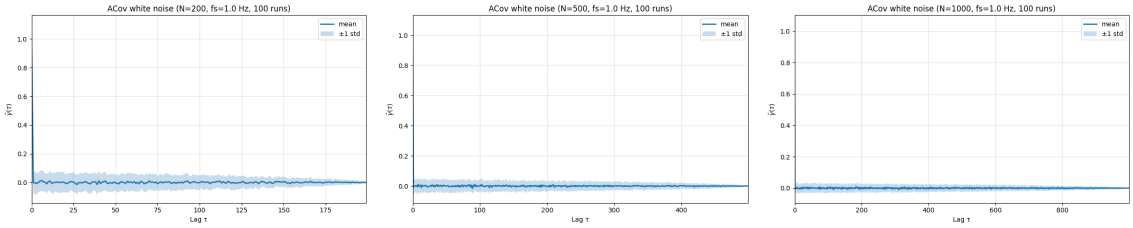
$$\mathbb{E}[|J(f^{(N)})|^2] \to \sum_{\tau=-\infty}^{+\infty} \gamma(\tau)\, e^{-2\pi \mathrm{i} f \tau/f_s} = S(f).$$

Question 6

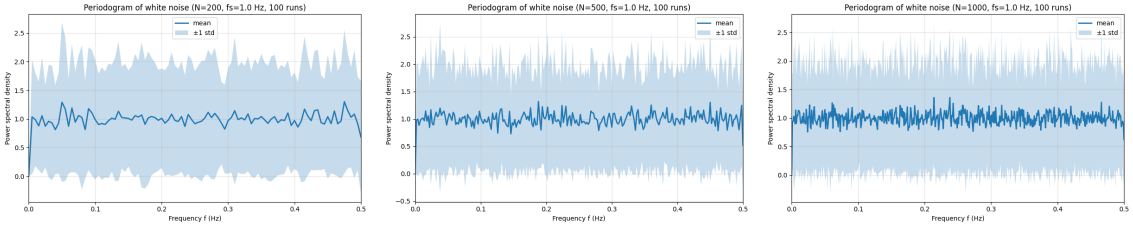
In this question, let X_n ($n = 0, \ldots, N-1$) be a Gaussian white noise with variance $\sigma^2 = 1$ and set the sampling frequency to $f_s = 1$ Hz

- For $N \in \{200, 500, 1000\}$, compute the *sample autocovariances* ($\hat{\gamma}(\tau)$ vs τ) for 100 simulations of X. Plot the average value as well as the average $\pm$, the standard deviation. What do you observe?
- For $N \in \{200, 500, 1000\}$, compute the *periodogram* ($|J(f_k)|^2$ vs f_k) for 100 simulations of X. Plot the average value as well as the average $\pm$, the standard deviation. What do you observe?

Add your plots to Figure 1.



Autocovariance ($N = 200$) Autocovariance ($N = 500$) Autocovariance ($N = 1000$)



Periodogram ($N = 200$) Periodogram ($N = 500$) Periodogram ($N = 1000$)

Figure 1: Autocovariances and periodograms of a Gaussian white noise (see Question 6).

Answer 6

- The theoretical autocovariance of white noise is σ^2 at lag=0 and 0 everywhere else. Here the sample autocovariance has the peak of 1 at lag=0 and abruptly decreases to low fluctuations around 0 with standard deviation of 0.1 for N=200. The standard deviation slowly decreases with the number of lags. That is not obvious from the expression derived further in Question 7. With bigger N, the standard deviation is lower for all lags, around 0.02 for N=500 at low lag.
- Theoretically, the periodogram of gaussian white noise is flat and equal to $\sigma^2/f_s = 1$ at all frequencies. Here we have a mean around 1 and a standard deviation of 1 too (we rescaled the autocovariances by 0.5 after computing the onesided periodogram) The standard deviation does not decrease with N.

Question 7

We want to show that the estimator $\hat{\gamma}(\tau)$ is consistent, i.e. it converges in probability when the number N of samples grows to ∞ to the true value $\gamma(\tau)$. In this question, assume that X is a wide-sense stationary *Gaussian* process.

- Show that for $\tau > 0$

$$\mathrm{var}(\hat{\gamma}(\tau)) = (1/N) \sum_{n=-(N-\tau-1)}^{n=N-\tau-1} \left(1 - \frac{\tau + |n|}{N}\right) \left[\gamma^2(n) + \gamma(n-\tau)\gamma(n+\tau)\right]. \tag{7}$$

  (Hint: if $\{Y_1, Y_2, Y_3, Y_4\}$ are four centered jointly Gaussian variables, then $\mathbb{E}[Y_1Y_2Y_3Y_4] = \mathbb{E}[Y_1Y_2]\mathbb{E}[Y_3Y_4] + \mathbb{E}[Y_1Y_3]\mathbb{E}[Y_2Y_4] + \mathbb{E}[Y_1Y_4]\mathbb{E}[Y_2Y_3]$.)

- Conclude that $\hat{\gamma}(\tau)$ is consistent.

Answer 7

This solution builds upon the results from Question 4. We assume X_n is a wide-sense stationary (WSS) Gaussian process with $\mathbb{E}[X_n] = 0$ (centered). The true autocovariance is $\gamma(\tau) = \mathbb{E}[X_n X_{n+\tau}]$. The estimator is $\hat{\gamma}(\tau) = \frac{1}{N}\sum_{n=0}^{N-\tau-1} X_n X_{n+\tau}$.

Part 1: Deriving the Variance

The variance of the estimator is defined as:

$$\text{var}(\hat{\gamma}(\tau)) = \mathbb{E}[\hat{\gamma}(\tau)^2] - (\mathbb{E}[\hat{\gamma}(\tau)])^2$$

From Question 4, we know the second term: $\mathbb{E}[\hat{\gamma}(\tau)] = \left(1 - \frac{\tau}{N}\right)\gamma(\tau)$. So, $(\mathbb{E}[\hat{\gamma}(\tau)])^2 = \left(1 - \frac{\tau}{N}\right)^2 \gamma^2(\tau)$.

The main task is to compute the first term, $\mathbb{E}[\hat{\gamma}(\tau)^2]$.

$$\mathbb{E}[\hat{\gamma}(\tau)^2] = \mathbb{E}\left[\left(\frac{1}{N}\sum_{n=0}^{N-\tau-1} X_n X_{n+\tau}\right)\left(\frac{1}{N}\sum_{m=0}^{N-\tau-1} X_m X_{m+\tau}\right)\right]$$

$$\mathbb{E}[\hat{\gamma}(\tau)^2] = \frac{1}{N^2}\sum_{n=0}^{N-\tau-1}\sum_{m=0}^{N-\tau-1} \mathbb{E}[X_n X_{n+\tau} X_m X_{m+\tau}]$$

Now we use the hint for the expectation of four centered Gaussian variables. Let $Y_1 = X_n, Y_2 = X_{n+\tau}, Y_3 = X_m, Y_4 = X_{m+\tau}$.

$$\mathbb{E}[Y_1 Y_2 Y_3 Y_4] = \mathbb{E}[Y_1 Y_2]\mathbb{E}[Y_3 Y_4] + \mathbb{E}[Y_1 Y_3]\mathbb{E}[Y_2 Y_4] + \mathbb{E}[Y_1 Y_4]\mathbb{E}[Y_2 Y_3]$$

Let's translate this back using the definition $\gamma(k) = \mathbb{E}[X_i X_{i+k}]$.

- $\mathbb{E}[Y_1 Y_2]\mathbb{E}[Y_3 Y_4] = \mathbb{E}[X_n X_{n+\tau}]\mathbb{E}[X_m X_{m+\tau}] = \gamma(\tau)\cdot\gamma(\tau) = \gamma^2(\tau)$
- $\mathbb{E}[Y_1 Y_3]\mathbb{E}[Y_2 Y_4] = \mathbb{E}[X_n X_m]\mathbb{E}[X_{n+\tau} X_{m+\tau}] = \gamma(m-n)\cdot\gamma((m+\tau)-(n+\tau)) = \gamma(m-n)\gamma(m-n) = \gamma^2(m-n)$
- $\mathbb{E}[Y_1 Y_4]\mathbb{E}[Y_2 Y_3] = \mathbb{E}[X_n X_{m+\tau}]\mathbb{E}[X_{n+\tau} X_m] = \gamma(m+\tau-n)\cdot\gamma(m-(n+\tau)) = \gamma(m-n+\tau)\gamma(m-n-\tau)$

Substituting this into the double sum for $\mathbb{E}[\hat{\gamma}(\tau)^2]$:

$$\mathbb{E}[\hat{\gamma}(\tau)^2] = \frac{1}{N^2}\sum_{n=0}^{N-\tau-1}\sum_{m=0}^{N-\tau-1} \left[\gamma^2(\tau) + \gamma^2(m-n) + \gamma(m-n+\tau)\gamma(m-n-\tau)\right]$$

We separate the $\gamma^2(\tau)$ term, which is constant in the sums:

$$\frac{1}{N^2}\sum_{n=0}^{N-\tau-1}\sum_{m=0}^{N-\tau-1} \gamma^2(\tau) = \frac{1}{N^2}(N-\tau)(N-\tau)\gamma^2(\tau) = \left(\frac{N-\tau}{N}\right)^2\gamma^2(\tau) = \left(1-\frac{\tau}{N}\right)^2\gamma^2(\tau)$$

This is exactly the term $(\mathbb{E}[\hat{\gamma}(\tau)])^2$.

Now, let's find the variance:

$$\text{var}(\hat{\gamma}(\tau)) = \mathbb{E}[\hat{\gamma}(\tau)^2] - (\mathbb{E}[\hat{\gamma}(\tau)])^2$$

$$\text{var}(\hat{\gamma}(\tau)) = \left[\left(1-\frac{\tau}{N}\right)^2\gamma^2(\tau)+\dots\right] - \left(1-\frac{\tau}{N}\right)^2\gamma^2(\tau)$$

The $\gamma^2(\tau)$ terms cancel out, leaving only the other two terms from the Gaussian expansion:

$$\text{var}(\hat{\gamma}(\tau)) = \frac{1}{N^2}\sum_{n=0}^{N-\tau-1}\sum_{m=0}^{N-\tau-1}\left[\gamma^2(m-n)+\gamma(m-n+\tau)\gamma(m-n-\tau)\right]$$

This is a double sum where the summand only depends on the difference $k = m-n$. We can convert this to a single sum over k. Let $A = N-\tau$. The sum is $\sum_{n=0}^{A-1}\sum_{m=0}^{A-1} f(m-n)$. The lag $k = m-n$ ranges from $-(A-1)$ to $A-1$. For any given k in this range, the number of (n,m) pairs that produce this lag is $(A-|k|)$, or $(N-\tau-|k|)$. So, we can rewrite the double sum as:

$$\text{var}(\hat{\gamma}(\tau)) = \frac{1}{N^2}\sum_{k=-(N-\tau-1)}^{N-\tau-1}(N-\tau-|k|)\left[\gamma^2(k)+\gamma(k+\tau)\gamma(k-\tau)\right]$$

To match the formula in the exercise, we factor out $1/N$:

$$\text{var}(\hat{\gamma}(\tau)) = \frac{1}{N}\sum_{k=-(N-\tau-1)}^{N-\tau-1}\frac{N-\tau-|k|}{N}\left[\gamma^2(k)+\gamma(k+\tau)\gamma(k-\tau)\right]$$

$$\text{var}(\hat{\gamma}(\tau)) = \frac{1}{N}\sum_{k=-(N-\tau-1)}^{N-\tau-1}\left(1-\frac{\tau+|k|}{N}\right)\left[\gamma^2(k)+\gamma(k+\tau)\gamma(k-\tau)\right]$$

This completes the proof of the first part.

Contrary to the correlogram, the periodogram is not consistent. It is one of the most well-known estimators that is asymptotically unbiased but not consistent. In the following question, this is proven for Gaussian white noise, but this holds for more general stationary processes.

Question 8

Assume that X is a Gaussian white noise (variance σ^2) and let $A(f) := \sum_{n=0}^{N-1} X_n \cos(-2\pi f n/f_s)$ and $B(f) := \sum_{n=0}^{N-1} X_n \sin(-2\pi f n/f_s$. Observe that $J(f) = (1/N)(A(f)+iB(f))$.

- Derive the mean and variance of $A(f)$ and $B(f)$ for $f = f_0, f_1, \dots, f_{N/2}$ where $f_k = f_s k/N$.
- What is the distribution of the periodogram values $|J(f_0)|^2, |J(f_1)|^2, \dots, |J(f_{N/2})|^2$.
- What is the variance of the $|J(f_k)|^2$? Conclude that the periodogram is not consistent.
- Explain the erratic behavior of the periodogram in Question 6 by looking at the covariance between the $|J(f_k)|^2$.

Answer 8

There is a small typo : $J(f) = (1/\sqrt{N})(A(f)+iB(f))$

1) $A(f)$ and $B(f)$ are linear combinations of centered independent Gaussian variables, hence they are also centered Gaussian:

$$\mathbb{E}[A(f)] = 0, \quad \mathbb{E}[B(f)] = 0.$$

For the variance, by bilinearity of covariances:

$$\begin{aligned}\operatorname{var}\big(A(f)\big) &= \operatorname{cov}\left(\sum_{n=0}^{N-1} X_n \cos(2\pi f n/f_s), \sum_{m=0}^{N-1} X_m \cos(2\pi f m/f_s)\right)\\ &= \sum_{n=0}^{N-1}\sum_{m=0}^{N-1} \operatorname{cov}(X_n, X_m)\, \cos(2\pi f n/f_s)\cos(2\pi f m/f_s)\\ &= \sigma^2 \sum_{n=0}^{N-1} \cos^2(2\pi f n/f_s).\end{aligned}$$

For $k \in \{1, \dots, N/2-1\}$, we have:

$$\operatorname{var}\big(A(f_k)\big) = \sigma^2 \sum_{n=0}^{N-1} \cos^2\left(\frac{2\pi k n}{N}\right) = \sigma^2 \sum_{n=0}^{N-1} \frac{1}{2}\left(1+\cos\left(\frac{4\pi k n}{N}\right)\right) = \frac{N\sigma^2}{2}.$$

Because for $k \neq 0$ and $k \neq N/2$, the sum of cosines is zero. The same result holds for $B(f_k)$.

For $k = 0$: $\cos^2(0) = 1$ and $\sin^2(0) = 0$. For $k = N/2$: $\cos^2(\pi n) = 1$ and $\sin^2(\pi n) = 0$,

To sum up:

$$\boxed{\operatorname{var}\big(A(f_k)\big) = \operatorname{var}\big(B(f_k)\big) = \tfrac{N}{2}\sigma^2.}$$

$$\boxed{\operatorname{var}\big(A(f_0)\big) = N\sigma^2, \quad \operatorname{var}\big(B(f_0)\big) = 0.}$$

$$\boxed{\operatorname{var}\big(A(f_{N/2})\big) = N\sigma^2, \quad \operatorname{var}\big(B(f_{N/2})\big) = 0.}$$

2) The periodogram writes:

$$|J(f_k)|^2 = \frac{1}{N}\big(A(f_k)^2 + B(f_k)^2\big).$$

Define the normalized variables Z_A and Z_B such that:

$$A(f_k) = \sqrt{(N/2)\sigma^2} Z_A(f_k), \quad B(f_k) = \sqrt{(N/2)\sigma^2} Z_B(f_k).$$

For $0 < k < N/2$:

$$|J(f_k)|^2 = \frac{1}{N}\left((N/2)\sigma^2 Z_A^2 + (N/2)\sigma^2 Z_B^2\right) = \frac{\sigma^2}{2}(Z_A^2 + Z_B^2),$$

For $k = 0$ or $k = N/2$, we have $B(f_k) = 0$ and $A(f_k) \sim \mathcal{N}(0, N\sigma^2)$, hence

$$|J(f_k)|^2 = \frac{1}{N}A(f_k)^2 = \frac{1}{N}N\sigma^2 Z_A^2 = \sigma^2 Z_A^2.$$

And using $2\cos(x)\sin(y) = \sin(2x)$, we have:

$$\begin{aligned}\operatorname{cov}(A(f_k), B(f_k)) &= \sum_{n=0}^{N-1}\sum_{m=0}^{N-1} \operatorname{cov}(X_n, X_m)\ \cos(2\pi f_k n/f_s)\sin(-2\pi f_k m/f_s) \\ &= \sigma^2 \sum_{n=0}^{N-1} \frac{1}{2}\sin(-4\pi f_k n/f_s) = 0.\end{aligned}$$

Hence $A(f_k)$, $B(f_k)$ are uncorrelated Gaussian variables, hence independent. Same for Z_A and Z_B. And the periodogram is then a sum of squares of two independent standard normal variables, that is it follows a χ^2 distribution:

For $0 < k < N/2$:

$$|J(f_k)|^2 = \frac{\sigma^2}{2}(Z_A^2 + Z_B^2) \sim \frac{\sigma^2}{2}\chi_2^2.$$

For $k = 0$ or $k = N/2$:

$$|J(f_k)|^2 = \sigma^2 Z_A^2 \sim \sigma^2 \chi_1^2.$$

3) For $0 < k < N/2$:

$$\operatorname{Var}[|J(f_k)|^2] = \operatorname{Var}\left[\frac{\sigma^2}{2}\chi_2^2\right] = \frac{\sigma^4}{4}\operatorname{Var}[\chi_2^2] = \frac{\sigma^4}{4}\cdot(2\cdot 2) = \sigma^4$$

And for $k = 0$ or $k = N/2$:

$$\operatorname{Var}[|J(f_k)|^2] = \operatorname{Var}[\sigma^2\chi_1^2] = \sigma^4 \operatorname{Var}[\chi_1^2] = \sigma^4 \cdot 2 = 2\sigma^4$$

For all k, the variance does not depend on N and does not go to 0 as $N \to \infty$. Without providing a strong argument of why it disproves consistency, we feel it must be the case and conclude that the periodogram is not consistent. Variance is a lower bound on mean squared error, but logical implication between MSE not converging to 0 and no consistency is not straightforward.

4) The covariance between two periodogram values at different frequencies f_k and f_l, with $k \neq l$, is 0. We don't prove it but it is similar calculation as for $A(f_k)$ and $B(f_k)$ not correlated and independent as gaussian, but here we take $A^2(f_k)$, $B^2(f_k)$, $A^2(f_l)$ and $B^2(f_l)$. Hence the periodogram value at one frequency is completely not correlated with the value at the next frequency. It explains the erratic behavior and noisy and spiky look, rather than a smooth curve.

Question 9

As seen in the previous question, the problem with the periodogram is the fact that its variance does not decrease with the sample size. A simple procedure to obtain a consistent estimate is to divide the signal into K sections of equal durations, compute a periodogram on each section, and average them. Provided the sections are independent, this has the effect of dividing the variance by K. This procedure is known as Bartlett's procedure.

- Rerun the experiment of Question 6, but replace the periodogram by Barlett's estimate (set $K = 5$). What do you observe?

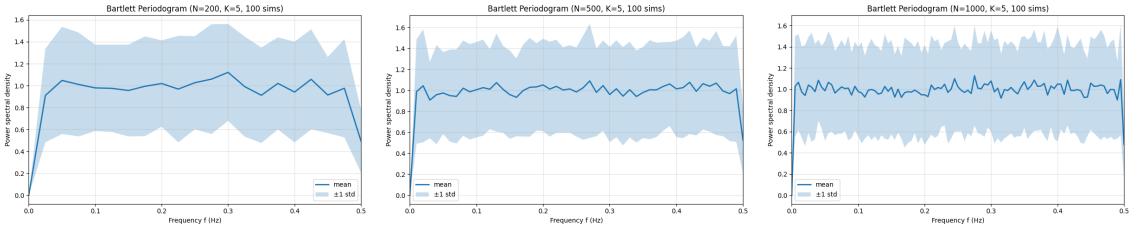
Add your plots to Figure 2.

Answer 9

The Bartlett method to compute periodogram reduces the variance by $1/K$:

$$\mathrm{Var}[\hat{S}_{\mathrm{Bartlett}}(f_k)] = \frac{1}{K}\mathrm{Var}[\hat{S}(f_k)]$$

As the periodograms on the non overlapping sections are independent. Hence it reduces the standard deviation by $1/\sqrt{K} \sim 0.45$, as we see in Figure 2,



Periodogram ($N = 200$) Periodogram ($N = 500$) Periodogram ($N = 1000$)

Figure 2: Barlett's periodograms of a Gaussian white noise (see Question 9).

4 Data study

4.1 General information

Context. The study of human gait is a central problem in medical research with far-reaching consequences in the public health domain. This complex mechanism can be altered by a wide range of pathologies (such as Parkinson's disease, arthritis, stroke,...), often resulting in a significant loss of autonomy and an increased risk of falls. Understanding the influence of such medical disorders on a subject's gait would greatly facilitate early detection and prevention of those possibly harmful situations. To address these issues, clinical and bio-mechanical researchers have worked to objectively quantify gait characteristics.

Among the gait features that have proved their relevance in a medical context, several are linked to the notion of step (step duration, variation in step length, etc.), which can be seen as the core atom of the locomotion process. Many algorithms have, therefore, been developed to automatically (or semi-automatically) detect gait events (such as heel-strikes, heel-off, etc.) from accelerometer and gyrometer signals.

Data. Data are described in the associated notebook.

4.2 Step classification with the dynamic time warping (DTW) distance

Task. The objective is to classify footsteps and then walk signals between healthy and non-healthy.

Performance metric. The performance of this binary classification task is measured by the F-score.

Question 10

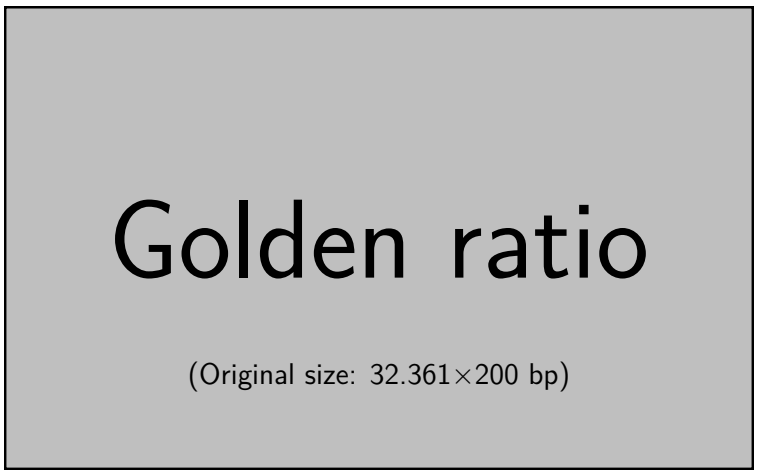
Combine the DTW and a k-neighbors classifier to classify each step. Find the optimal number of neighbors with 5-fold cross-validation and report the optimal number of neighbors and the associated F-score. Comment briefly.

Answer 10

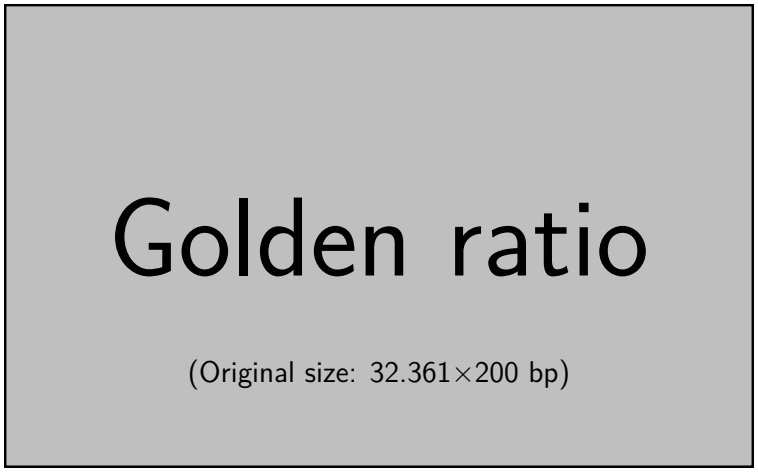
Question 11

Display on Figure 3 a badly classified step from each class (healthy/non-healthy).

Answer 11



Badly classified healthy step



Badly classified non-healthy step

Figure 3: Examples of badly classified steps (see Question 11).

A Conjugate of the vector ℓ_1 norm

Let $f(x) = \|x\|_1$. Compute

$$f^*(y) = \sup_{x \in \mathbb{R}^n} \{\langle y, x \rangle - \|x\|_1\}.$$

Case 1: $\|y\|_\infty \le 1$

Using the Hölder inequality (or the definition of the dual norm),

$$\langle y, x\rangle \le \|y\|_\infty \|x\|_1 \le \|x\|_1.$$

Thus for every x,

$$\langle y, x\rangle - \|x\|_1 \le 0.$$

Taking the supremum over x yields $f^*(y) \le 0$. The value 0 is attained in the limit at $x = 0$ (indeed $\langle y, 0\rangle - \|0\|_1 = 0$), hence

$$f^*(y) = 0 \quad \text{when } \|y\|_\infty \le 1.$$

Case 2: $\|y\|_\infty > 1$

There exists an index i with $|y_i| > 1$. Fix that index and consider vectors of the form $x = te_i \cdot \operatorname{sign}(y_i)$ for scalar $t > 0$, where e_i is the ith coordinate unit vector. Then

$$\langle y, x\rangle - \|x\|_1 = t|y_i| - t = t(|y_i| - 1).$$

Since $|y_i| - 1 > 0$, letting $t \to +\infty$ makes the expression tend to $+\infty$. Therefore $f^*(y) = +\infty$ for any y with $\|y\|_\infty > 1$.

Conclusion for vectors

Combining the two cases,

$$\boxed{(\|\cdot\|_1)^*(y) = \begin{cases} 0, & \|y\|_\infty \le 1, \\ +\infty, & \|y\|_\infty > 1, \end{cases}}$$

i.e. the conjugate is the *indicator function* of the ℓ_∞ unit ball:

$$(\|\cdot\|_1)^*(y) = \iota_{\{y:\|y\|_\infty \le 1\}}(y).$$

This matches the general fact that the conjugate of a (nonnegative, proper) norm is the indicator of the unit ball of its dual norm.

B Entrywise matrix ℓ_1 norm

Let $X \in \mathbb{R}^{m\times n}$ and define the entrywise ℓ_1 (sometimes called the elementwise ℓ_1) norm

$$\|X\|_1 := \sum_{i=1}^{m}\sum_{j=1}^{n} |X_{ij}|.$$

Identify matrices with vectors in $\mathbb{R}^{mn}$ by stacking entries. The dual norm (entrywise) is the elementwise ℓ_∞ norm

$$\|Y\|_{\infty,\text{entry}} := \max_{i,j} |Y_{ij}|.$$

Applying the same argument as for vectors (or the general dual-norm conjugate fact) we obtain

$$\boxed{(\|\cdot\|_1)^*(Y) = \begin{cases} 0, & \max_{i,j} |Y_{ij}| \le 1, \\ +\infty, & \max_{i,j} |Y_{ij}| > 1. \end{cases}}$$

Equivalently,

$$(\|\cdot\|_1)^*(Y) = \iota_{\{Y:\|Y\|_{\infty,\text{entry}}\le 1\}}(Y).$$

C Fenchel–Young equality and subgradient characterization

For a convex, proper f and points x, y the Fenchel–Young inequality states

$$\langle y, x\rangle \le f(x) + f^*(y).$$

Equality holds iff $y \in \partial f(x)$ (i.e. y is a subgradient of f at x). Specializing to $f(x) = \|x\|_1$ and using the conjugate above:

- If $\|y\|_\infty > 1$ then $f^*(y) = +\infty$ and Fenchel–Young gives no finite equality.
- If $\|y\|_\infty \le 1$ then $f^*(y) = 0$ and equality $\langle y, x\rangle = \|x\|_1$ holds exactly when $y \in \partial\|x\|_1$, i.e. componentwise

  $$y_i = \begin{cases} \operatorname{sign}(x_i) & \text{if } x_i \neq 0, \\ \theta_i \in [-1,1] & \text{if } x_i = 0, \end{cases} \qquad i = 1, \dots, n,$$

  and therefore $\|y\|_\infty \le 1$.

For matrices the subgradient description is the same applied entrywise.

D Remarks and related conjugates

- **Dual-norm viewpoint.** The result is a direct instance of the general fact: if $f(x) = \|x\|$ is a norm then $f^*(y) = \iota_{\{\|y\|_* \le 1\}}(y)$ where $\|\cdot\|_*$ is the dual norm.
- **Other matrix norms.** The entrywise ℓ_1 norm has dual the entrywise ℓ_∞ norm. By contrast, the nuclear (trace) norm $\|\cdot\|_*$(sum of singular values) is dual to the operator (spectral) norm $\|\cdot\|_{2\to 2}$; its conjugate is the indicator of the spectral-norm unit ball:

  $$(\|\cdot\|_*)^*(Y) = \iota_{\{\|Y\|_{2\to 2}\le 1\}}(Y).$$

  This is often used in low-rank matrix problems and is conceptually analogous.
- **Proximal operator.** The proximal operator of $\lambda\|x\|_1$ (soft-thresholding) can be derived from Moreau decomposition which uses the conjugate. Moreau identity:

  $$\operatorname{prox}_{\lambda f}(v) + \lambda \operatorname{prox}_{f^*/\lambda}\left(v/\lambda\right) = v.$$

  For $f = \|\cdot\|_1$ the projection onto the ℓ_∞ unit ball appears in the decomposition.

E Short examples

1. $x = (2, -3)$, $y = (1, 0.5)$: $\|y\|_\infty = 1 \le 1$, so $(\|\cdot\|_1)^*(y) = 0$ and $\langle y, x\rangle - \|x\|_1 = (1)(2) + (0.5)(-3) - (2+3) = 2 - 1.5 - 5 = -4.5 \le 0$.
2. $y = (1.2, 0)$: $\|y\|_\infty = 1.2 > 1$, hence $(\|\cdot\|_1)^*(y) = +\infty$ (supremum is unbounded).

This infimum is separable. For each component j:

- If $|v_j| > \lambda$, we can choose $\beta_j \to \mp\infty$ (opposite sign of v_j) to make the expression go to $-\infty$.
- If $|v_j| \leq \lambda$, the minimum is 0, achieved at $\beta_j = 0$.

Thus, the infimum is 0 if and only if $|v_j| \leq \lambda$ for all j, and $-\infty$ otherwise. This condition is $\|v\|_\infty \leq \lambda$, or:

$$\|X^T \nu\|_\infty \leq \lambda$$

Combining the results, the dual function is:

$$g(\nu) = \begin{cases} -\frac{1}{2}\|\nu\|_2^2 - \nu^T y & \text{if } \|X^T \nu\|_\infty \leq \lambda \\ -\infty & \text{if } \|X^T \nu\|_\infty > \lambda \end{cases}$$