UNIVERSITE DE GENEVE Section de Mathématiques
Faculté des sciences

Statistical Machine Learning I - 2025-26
TP 4

Exercise 1 (Universal Thresholding)**.** Consider the model

$$Y_i = \alpha_i + \epsilon_i, \quad i = 1, \dots, n,$$

where the noise terms ϵ_i are iid distributed as $\mathcal{N}(0, \sigma^2)$. We estimate the unknown signal coefficients $\boldsymbol{\alpha}$ using the hard-thresholding rule

$$\hat{\alpha}_i = \eta_\lambda(Y_i) \quad \text{with} \quad \eta_\lambda(y) = \begin{cases} y, & |y| > \lambda, \\ 0, & \text{otherwise.} \end{cases}$$

In this exercise, we analyze the *universal threshold* proposed by Donoho and Johnstone: $\lambda_n = \sigma\sqrt{2\log n}$.

1. Implement a function that generates data for different sparsity levels s, corresponding to the number of nonzero entries in the true vector $\boldsymbol{\alpha}$. The nonzero coefficients should have amplitude proportional to a fixed signal-to-noise ratio (SNR).

2. Using the generated data $\mathbf{Y}$, compute the hard-threshold estimator $\hat{\boldsymbol{\alpha}}$. Plot on the same figure the observed data $\mathbf{Y}$, the true coefficients $\boldsymbol{\alpha}$, and the estimated coefficients $\hat{\boldsymbol{\alpha}}$. Add horizontal lines indicating $\pm\lambda_n$, and repeat for several values of s to visualize the effect of sparsity.

3. Perform a Monte Carlo simulation to estimate

$$\mathbb{P}(\hat{\boldsymbol{\alpha}} = \mathbf{0})$$

under the null case $\boldsymbol{\alpha} = \mathbf{0}$, for increasing values of n. Show empirically that this probability tends to 1 as $n \to \infty$. Recall from TD3 that the theoretical asymptotic behaviour satisfies

$$1 - \mathbb{P}(\hat{\boldsymbol{\alpha}} = \mathbf{0}) \approx 1 - \frac{1}{\sqrt{\pi \log n}}$$

and plot both the simulated and theoretical curves on a log–log scale.

4. For fixed n, estimate the probability of exact support recovery (that is, the probability that $\hat{\boldsymbol{\alpha}}$ identifies exactly the same nonzero entries as $\boldsymbol{\alpha}$).Repeat the experiment for increasing values of s (the true sparsity) and for several n. Plot the resulting *phase transition curves* to illustrate how recovery performance degrades as s grows.

Exercise 2 (Hard vs. Soft Thresholding: a Bias–Variance Trade-Off)**.** In the previous exercise, we analyzed the ability of the universal threshold to recover the correct support (i.e., the set of nonzero coefficients) with high probability. In this exercise, we will focus on the *mean squared error (MSE)* criterion instead, and compare the performance of hard and soft thresholding.

1. Generate data using the same procedure as in Exercise 1, with a fixed number of nonzero coefficients s, and apply both hard and soft thresholding using the universal threshold $\lambda_n = \sigma\sqrt{2\log n}$. Compute the empirical MSE

$$\mathrm{MSE} = \frac{1}{n}\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|_2^2$$

   for each method.

2. Repeat the experiment for different values of the nonzero coefficient amplitude (e.g., SNR $\in \{1, 2, 4, 8\}$). For each SNR, estimate and plot the average MSE (over several Monte Carlo trials) for both hard and soft thresholding.

3. Discuss your results:

   - Do both estimators yield identical PESR? Why?
   - Which estimator achieves lower MSE at low SNR? Explain intuitively.
   - Which estimator becomes preferable as SNR increases? Why?