Home Work 2025-2026

Project: Nonparametric Learning under Multiplicative Label Corruption

1. To be returned no later than January 17 at 14 p.m to sana.louhichi@univ-grenoble-alpes.fr.
2. This project can be done alone or in pairs (Please indicate clearly your names in your file). **A single file to submit (in pdf).**
3. Don't forget to include the instructions for the used software.
4. The project may be complemented by an oral presentation or examination in case of doubts regarding the use of generative AI tools.

Context

We observe a dataset

$$\mathcal{D}_m = \{(X_i, Z_i)\}_{1 \le i \le m},$$

where the observations are stationary and distributed as a pair (X, Z), with $X \in \mathbb{R}^d$ and $Z \in \{-1, 1\}$. The observed label Z is a corrupted version of an unobserved true label $Y \in \{-1, 1\}$, according to an *instance-dependent classification noise model*.

Label corruption model

For each input point $X = x$, we associate a noise level $\eta(x) \in [0, 0.5[$. Conditionally on $X = x$ and Y, the observed label is defined by

$$Z = \begin{cases} -Y & \text{with probability } \eta(x), \\ Y & \text{with probability } 1 - \eta(x). \end{cases}$$

The function η is uniformly bounded: there exists a positive constant $\eta_{max} \in [0, 0.5[$ such that for any $x \in \mathbb{R}^d$,

$$\eta(x) \le \eta_{max}.$$

Interpretation. At each observation X, an adversary flips a biased coin with parameter $\eta(X)$. If the coin lands heads, the true label is flipped; otherwise, it is kept unchanged.

This mechanism is commonly referred to as *multiplicative label corruption* or *instance-dependent label noise*.

Project objectives

The goal of this project is to understand, model, and analyze a realistic statistical learning problem involving input-dependent label noise, with a level of mathematical and statistical rigor appropriate for Master's-level or engineering students.

Questions and expected outcomes

1. **Modeling.**
   - Propose a complete probabilistic model linking X, Y, Z, and the noise function $\eta(X)$.
   - Discuss identifiability issues: under which assumptions can η be estimated without observing Y?
2. **Estimation of the noise function.**
   - Propose a nonparametric estimator of the function η.
   - Provide statistical intuition and justification for the estimator.
3. **Numerical study.**
   - Simulate data according to the proposed model.
   - Implement the estimator of η that you proposed.
   - Illustrate empirically its performance (estimation error, influence of the sample size m, the dimension d, and the noise level).

Expected deliverables

- A concise report describing the model, the estimation procedure, and the numerical results.
- Reproducible code (Python or R) for data generation and simulations.
- Well-commented figures illustrating the observed behaviors.

Optional extension. Study the impact of label corruption on a nonparametric classifier, and propose a correction based on the estimated noise function η.