

# Optimization for Data Science

M2 MIAGE & MIAGE ID Apprentissage

*Project - 2025/2026*



- The last version of this document can be found at:  
<https://www.lamsade.dauphine.fr/~croyer/ensdocs/ODS/ProjODS.pdf>.
- Typos, questions, etc, can be sent to [clement.royer@lamsade.dauphine.fr](mailto:clement.royer@lamsade.dauphine.fr).
- Current version: January 26, 2026.
  - 2026.01.26: First version of the document.
  - 2026.01.19: Beta version.

## Assignment

- This project follows the template of the course notebooks by mixing optimization questions with implementation tasks.
- Students may discuss the project with their classmates, but they should submit the project in groups of  $n$  students with  $n \in \{1, 2\}$ .
- Students may submit their sources in either French or English.
- Students are expected to submit sources that include:
  - The companion notebook<sup>1</sup> filled out.
  - Their answers to the questions (either in a separate PDF or directly in the notebook).
- Please send your sources to [clement.royer@lamsade.dauphine.fr](mailto:clement.royer@lamsade.dauphine.fr) under the form of a compressed folder. Those sources must include the first and last name of every member of the group.
- The deadline to send the sources is **February 12, 2026 AOE** (Anywhere On Earth).

---

<sup>1</sup>Available at <https://www.lamsade.dauphine.fr/~croyer/ensdocs/ODS/SourcesProjODS.zip>

# Project: Robust regression

## 1 Robust nonconvex regression and gradient descent

In this part of the project, we consider data under the form of a matrix  $\mathbf{X} \in \mathbb{R}^{n \times d}$  and a vector  $\mathbf{y} \in \mathbb{R}^n$ . In order to compute a linear model that is robust to outliers in the data, we consider the use of the **smoothed biweight loss**

$$\begin{aligned} \phi: \mathbb{R} &\rightarrow \mathbb{R} \\ t &\mapsto \frac{t^2}{1+t^2}. \end{aligned} \quad (1)$$

We thus obtain the following optimization problem:

$$\underset{\mathbf{w} \in \mathbb{R}^d}{\text{minimize}} f(\mathbf{w}) := \frac{1}{2n} \sum_{i=1}^n \phi(\mathbf{x}_i^T \mathbf{w} - y_i). \quad (2)$$

The function  $f$  is nonconvex and  $\mathcal{C}_L^{2,1}$ , i.e. twice continuously differentiable with a Lipschitz continuous gradient with Lipschitz constant  $L = \frac{1}{n} \|\mathbf{X}^T\| \|\mathbf{X}\|$ . Moreover, for any  $\mathbf{w} \in \mathbb{R}^d$ , the gradient of  $f$  at  $\mathbf{w}$  is given by

$$\nabla f(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n \frac{\mathbf{x}_i^T \mathbf{w} - y_i}{(1 + (\mathbf{x}_i^T \mathbf{w} - y_i)^2)^2} \mathbf{x}_i = \frac{1}{n} \mathbf{X}^T \mathbf{z} \in \mathbb{R}^d, \quad \mathbf{z} := \left[ \frac{\mathbf{x}_i^T \mathbf{w} - y_i}{(1 + (\mathbf{x}_i^T \mathbf{w} - y_i)^2)^2} \right] \in \mathbb{R}^n. \quad (3)$$

Finally, the Hessian matrix of  $f$  at  $\mathbf{w}$  is given by

$$\nabla^2 f(\mathbf{w}) = \frac{1}{n} \mathbf{X}^T \text{diag} \left( \left\{ \frac{1 - 3(\mathbf{x}_i^T \mathbf{w} - y_i)^2}{(1 + (\mathbf{x}_i^T \mathbf{w} - y_i)^2)^3} \right\}_{i=1}^n \right) \mathbf{X} \in \mathbb{R}^{d \times d}. \quad (4)$$

**Question 1** Consider the toy problem corresponding to

$$\mathbf{X} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{y} = \begin{bmatrix} 1 \\ -1 \\ 1 \\ 1 \end{bmatrix} \quad (5)$$

a) Justify that the optimal value of this problem cannot be 0.

b) Show that the point  $\bar{\mathbf{w}} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$  is a saddle point of this problem.

**Question 2** Analyze the results obtained on gradient descent for the toy problem (5).

a) Why does the first run (starting from  $\mathbf{w}_0 = \begin{bmatrix} 0 \\ 1.4 \end{bmatrix}$ ) converge to the saddle point  $\bar{\mathbf{w}}$ ?

b) A research paper from 2015 showed that gradient descent escapes saddle points for almost initial point. How does this experiment confirm the results of the theorem?

**Question 3** Analyze the results obtained on gradient descent for the synthetic data proposed in the notebook.

- a) What stepsize strategy seems the most efficient? Justify your answer by taking into account the cost of both strategies.
- b) Do the gradient norms curves correspond to the theoretical convergence rate of gradient descent (in  $\mathcal{O}(1/\sqrt{K})$  for nonconvex functions, where  $K$  is the iteration index)? Is that surprising?

## 2 Stochastic gradient and robust regression

In this section, we consider problem (2) as a finite-sum problem. With this perspective, our problem can be written as

$$\underset{\mathbf{w} \in \mathbb{R}^d}{\text{minimize}} f(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n f_i(\mathbf{w}), \quad f_i(\mathbf{w}) := \frac{1}{2} \phi(\mathbf{x}_i^T \mathbf{w} - y_i), \quad (6)$$

where  $\phi$  is the biweight loss function defined in (1).

**Question 4** Consider one more time the toy problem (5).

- a) Consider a run of batch stochastic gradient such that  $\mathcal{S}_k = \{2, 4\}$ . Show that the resulting stochastic gradient is identical to that of a stochastic gradient iteration with  $i_k = 4$ .
- b) For this nonconvex problem, it can be shown that a stochastic gradient method has a convergence rate in  $\mathcal{O}(\frac{1}{K^{1/4}})$  where  $K$  is the iteration count. Is it better or worse than the rate of gradient descent? Is there a setting in which we can consider stochastic gradient to be more efficient?
- c) Do you expect stochastic gradient to outperform gradient descent on the toy problem (5)?

**Question 5** Analyze the performance of stochastic gradient methods on the robust regression problem.

- a) According to the plot, is stochastic gradient performing better than gradient descent?
- b) What value of the batch size appears to be the most efficient? How would you suggest a better value?

### 3 Regularized robust regression

In this last section, we consider a regularized version of problem (2), namely

$$\underset{\mathbf{w} \in \mathbb{R}^d}{\text{minimize}} f_2(\mathbf{w}) := f(\mathbf{w}) + \frac{\lambda}{2} \|\mathbf{w}\|^2, \quad (7)$$

where  $f$  is the objective function of problem (2).

**Question 6** Consider (for the last time) the toy dataset (5), now used in problem (2).

- a) Justify that the saddle point of question 1 is no longer a saddle point of (7) with  $\lambda > 0$ .
- b) Write down the iterations of gradient descent and proximal gradient descent applied to problem (7) with  $\mathbf{w}_0 = \bar{\mathbf{w}}$ . Compare the resulting value of  $\mathbf{w}_1$ .

**Question 7** Analyze the performance of the two approaches.

- a) According to the plot, is proximal gradient better than gradient descent? How can you explain it?
- b) Change the value of  $\lambda$  to a much larger value. What do you observe? Is it consistent with the properties of  $\ell_2$  regularization?