IASD/MASH project #1
Ridge Regression with Linear Model and Perceptrons

Gabriel Peyré

October 2023

You can use the code from the "Machine Learning" numerical tours (`https://www.numerical-tours.com/python/`) to help you.

Question 1 Load an original dataset $X \in \mathbb{R}^{N \times d}$ and $y \in \mathbb{R}^N$ with more data N than dimension of features d, and such that $\ker(X) = \{0\}$ (you should check this). Comment on this dataset (what are the features, the dimensions of the problem, how does the correlation matrix look like).

Question 2 We first want to solve regression

$$\min_w E(w) := \|Xw - y\|_2^2 = \sum_{i=1}^N \left(\langle w, x_i \rangle - y_i\right)^2. \tag{1}$$

Implement gradient descent

$$w_{k+1} = w_k - \tau \nabla E(w_k).$$

Display the convergence of w_k and $E(w_k)$ on the training loss for several fixed step sizes τ.

Question 3 Experimentally, what is the optimal step τ? How does this compare with the theory?

Question 4 Implement in Numpy a Multilayer Perceptron with 2 layers and q neurons, with parameter $\theta = (V, w)$

$$g(\theta, x) := \sum_{k=1}^q w_k \langle v_k, x \rangle$$

where $w = (w_k)_k \in \mathbb{R}^q$ are the outer weights and $V = (v_k)_k \in \mathbb{R}^{q \times d}$ are the inner weights (the v_k are the columns of V). This implementation should use matrix-vector multiplication, take as input X, and output the vector $(g(\theta, x_i))_i$ of all the output.

Question 5 We want to minimize

$$F(\theta) := \sum_i |g(\theta, x_i) - y_i|^2.$$

We start by fixing $V = V_0$ to be a realization of a random matrix using `np.random.randn`. Explain why the function $E(w) := F(V_0, w)$ is a regression problem of the same form as (1) but with a different matrix X that you should write. Is there always a unique solution to this problem? Implement gradient descent for a value of $q < N$ and a value of $q \gg N$. What does the theory tell us about the convergence? What is experimentally the optimal step τ in terms of convergence speed for w_k and for $E(w_k)$?

Question 6 Take an optimal $w^\star$ optimized using the previous question. What is the gradient of the function $G(V) := F(V, w^\star)$ (so we only optimize the inner weights)? Implement gradient descent, initializing V equal to V_0, and display the convergence for different values of τ.

Bonus question (not taken into account – not marked) Use a testing dataset X_{test} (different from the X you use for training) and compute the test error using the linear model and the MLP.