Exercise Collection – Regularization

Contents

Lecture exercises **1**

Exercise 1: Hypothesis Space, Capacity, Regularization . . . 1

Exercise 2: Lasso, Subdifferentials . . . 1

Lecture exercises

Exercise 1: Hypothesis Space, Capacity, Regularization

(a) Simulate a data set with $n = 100$ observations based on the relationship $Y = \sin(x_1) + \varepsilon$ with noise term ε following some distribution. Simulate $p = 100$ additional covariates $x_2, \ldots, x_{101}$ that are not related to Y.

(b) On this data set, use different models (and software packages) of your choice to demonstrate

- overfitting and underfitting;
- $L1$, $L2$ and elastic net regularization;
- the underdetermined problem;
- the bias-variance trade-off;
- early stopping (use a simple neural network as in Exercise 2).

Exercise 2: Lasso, Subdifferentials

Optimization routines for the Lasso use coordinate gradient descent, but instead of using gradients, they resort to subdifferentials. We now try to understand in more detail what subdifferentials are:

(a) Recall that the Taylor approximation of first order of a function $f(x)$ at point x_0 is

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0).$$

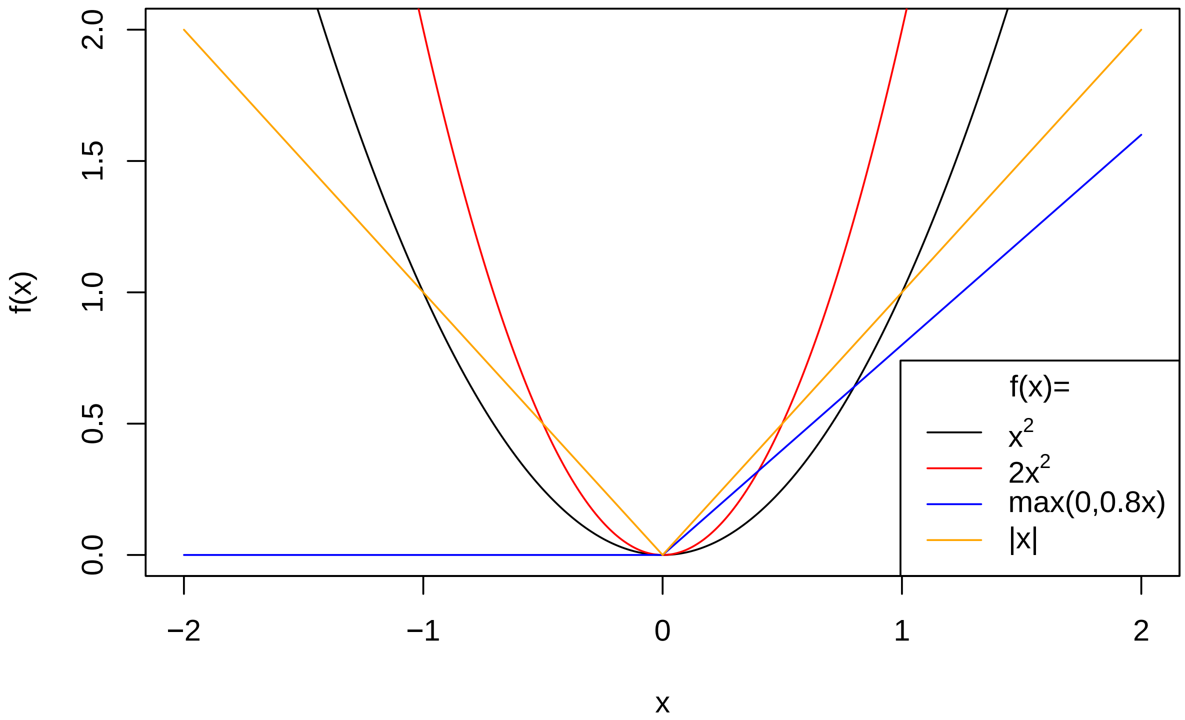
On the other hand, a differentiable function f is said to be convex on an interval $\mathcal{I}$ if and only if

$$f(x) \geq f(x_0) + f'(x_0)(x - x_0)$$

for all points $x, x_0 \in \mathcal{I}$.

(i) What conclusion can we therefore draw if we approximate a convex function with a Taylor approximation of first order?

(ii) Visualize such an approximation for different values x_0 for one of the following convex functions on $\mathcal{I} = [-2, 2]$.



(b) A subdifferential of f is a set of values $\breve{\nabla}_{x_0} f$ defined as

$$\breve{\nabla}_{x_0} f = \{g : f(x) \geq f(x_0) + g \cdot (x - x_0)\, \forall x \in \mathcal{I}\}.$$

Every scalar value $g \in \breve{\nabla}_{x_0}$ is said to be a subgradient of f at x_0. Does a subdifferential have any parallels to the previous question? How can we interpret g?

(c) We can make use of subdifferentials for convex but non-differentiable loss functions like the one induced by the Lasso. It holds that:

A point x_0 is the global minimum of a convex function $f \Leftrightarrow 0$ is contained in the subdifferential $\breve{\nabla}_{x_0} f$.

We can define a subdifferential at point x_0 also as a non-empty interval $[x_l, x_u]$ where the lower and upper limit is defined by

$$x_l = \lim_{x \to x_0^-} \frac{f(x) - f(x_0)}{x - x_0}, \quad x_u = \lim_{x \to x_0^+} \frac{f(x) - f(x_0)}{x - x_0}.$$

These resemble the limits of the derivative $\partial f / \partial x$ evaluated at a point very close to x_0 when coming from the left or right side, respectively.

(i) Derive $\breve{\nabla}_{x_0} f$ for $f(x) = |x|$ at $x_0 = 0$.

(ii) Is 0 a global minimum? Explain.

(iii) What is the subdifferential of the Lasso penalty $\lambda \sum_{j=1}^{p} |\theta_j|$? Hint: $\breve{\nabla}_{x_0}(f + g) = \breve{\nabla}_{x_0} f + \breve{\nabla}_{x_0} g$. Also, a subdifferential of a constant function is 0 and at any other differentiable point x_0, the subdifferential is equal to the gradient.

(d) Derive the subdifferential for the Lasso problem

$$\mathcal{R}_{reg} = n^{-1} \sum_{i=1}^{n} (y^{(i)} - x_1^{(i)} \theta_1 - x_2^{(i)} \theta_2)^2 + \lambda \sum_{j=1}^{2} |\theta_j|$$

w.r.t. θ_2, i.e., for an $L1$-regularized linear model with two linear features x_1 and x_2.