# Homework 11

Due May 10 at 11 pm

Yves Greatti - yg390

1. (Lasso and $\ell_0$) The file `X.txt` contains a $50 \times 300$ matrix $X$, and the file `y.txt` contains the $50 \times 1$ vector $y$. Each line of each file represents a row of the corresponding matrix, and the values on each line are space-delimited.

   (a) Consider the lasso problem

   $$\min_{\beta} \frac{1}{2n}\|X\beta - y\|^2 + \lambda\|\beta\|_1$$

   where $\lambda > 0$ is a parameter and $n = 50$. Construct a (semilogx) plot that draws a separate path for each coefficient value as a function of $\lambda$. Include values of $\lambda$ between $0.01$ and $2$ (you can include more if you want), and make your values spaced evenly on the log axis (e.g., `np.geomspace`). You can solve the lasso problem using whatever code/library you want.

   (b) Determine the minimizer of

   $$\begin{array}{ll} \text{minimize} & \|\beta\|_0 \\ \text{subject to} & X\beta = y. \end{array}$$

   Assume that the minimizer has small $\ell_0$ norm, i.e $\ell_0 \leq 2$. Explain your strategy and justify that it finds the minimizer. Report the nonzero coefficients of the minimizer, and their values. Remember that two floating point values may be different for numerical reasons even if they represent the same value.

   (c) Will your strategy in (b) always find the optimal minimizer of any least-squares problem with $\ell_0$ regularization?

2. (Proximal operator) The proximal operator of a function $f : \mathbb{R}^n \to \mathbb{R}$ is defined as

$$\text{prox}_f(y) := \arg\min_x f(x) + \frac{1}{2}||x - y||_2^2. \tag{1}$$

(a) Derive the proximal operator of the squared $\ell_2$ norm weighted by a constant $\alpha > 0$, i.e. $f(x) = \alpha ||x||_2^2$.

$$\text{prox}_f(y) := \arg\min_x \alpha ||x||_2^2 + \frac{1}{2}||x - y||_2^2$$

The two terms are quadratic, therefore differentiable, the gradient is

$$\nabla_x \text{prox}_f(y) = 2\alpha x + (x - y)$$

Setting the gradient to zero, yields:

$$\begin{aligned} 2\alpha x + (x - y) &= 0 \\ x &= \frac{1}{1 + 2\alpha} y \\ \text{prox}_f(y) &= \frac{1}{1 + 2\alpha} y, \alpha > 0 \end{aligned}$$

(b) Prove that the proximal operator of the $\ell_1$ norm weighted by a constant $\alpha > 0$ is a soft-thresholding operator,

$$\text{prox}_{\alpha||\cdot||_1}(y) = \mathcal{S}_\alpha(y), \tag{2}$$

where

$$\mathcal{S}_\alpha(y)[i] := \begin{cases} y[i] - \text{sign}(y[i])\alpha & \text{if } |y[i]| \geq \alpha, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

$$\text{prox}_{\alpha||\cdot||_1}(y) = \alpha\|x\|_1 + \frac{1}{2}||x - y||_2^2, \alpha > 0$$

And we are looking for:

$$\begin{aligned} 0 &\in \partial(\alpha\|x\|_1) + \nabla_x(\frac{1}{2}||x - y||_2^2) \\ 0 &\in \alpha\, \partial(\|x\|_1) + (x - y) \end{aligned}$$

We examine each component of $x$ and $y$ separately. Assume first that $x[i] \neq 0$ then $\partial(\|x\|_1) = \operatorname{sign}(x[i])$, setting the subgradient to 0, we have:

$$\begin{aligned} x[i] - y[i] + \alpha \operatorname{sign}(x[i]) &= 0 \\ x[i] &= y[i] - \alpha \operatorname{sign}(x[i]) \end{aligned}$$

Note that

$$\begin{aligned} x[i] < 0, \operatorname{sign}(x[i]) = -1 \rightarrow y[i] + \alpha < 0 \quad &\text{or } y[i] < -\alpha < 0 \\ x[i] > 0, \operatorname{sign}(x[i]) = 1 \rightarrow y[i] - \alpha > 0 \quad &\text{or } y[i] > \alpha > 0 \end{aligned}$$

thus in this case $\operatorname{sign}(x[i]) = \operatorname{sign}(y[i])$ and the optimal point is $y[i] - \alpha \operatorname{sign}(y[i])$. In the case where $x[i] = 0$, let $\gamma = \partial(\|x\|_1), |\gamma| \leq 1$ then it holds

$$\begin{aligned} x[i] - y[i] + \alpha\gamma = 0 \rightarrow y[i] - \alpha\gamma &= 0 \\ y[i] &= \gamma\alpha \\ |y[i]| &\leq \alpha \end{aligned}$$

Putting all together, we get

$$\operatorname{prox}_{\alpha ||\cdot||_1}(y) = \begin{cases} y[i] - \operatorname{sign}(y[i])\, \alpha & \text{if } |y[i]| \geq \alpha, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

(c) Prove that if $X \in \mathbb{R}^{p \times n}$ has orthonormal rows ($p \leq n$) and $y \in \mathbb{R}^n$, then for any function $f$

$$\arg\min_{\beta} \frac{1}{2} \left|\left|y - X^T\beta\right|\right|_2^2 + f(\beta) = \arg\min_{\beta} \frac{1}{2} ||Xy - \beta||_2^2 + f(\beta). \tag{5}$$

The two expressions for the same function $f$ differs on the first term, so we want to show that

$$\arg\min_{\beta} \frac{1}{2} \left|\left|y - X^T\beta\right|\right|_2^2 = \arg\min_{\beta} \frac{1}{2} ||Xy - \beta||_2^2$$

(d) Use the answers to the previous questions to compare the ridge-regression and lasso estimators for a regression problem where the features are orthonormal.

The use of $l_1$, $l_2$ norms gives rise to the problems, for $\lambda > 0$:

$$\frac{1}{2} \arg\min_{\beta} \|y - X^T\beta\|_2^2 + \lambda\|\beta\|_2^2 \text{ Ridge regression}$$

$$\frac{1}{2} \arg\min_{\beta} \|y - X^T\beta\|_2^2 + \lambda\|\beta\|_1 \text{ Lasso regression}$$

which is equivalent from part c) to

$$\arg\min_{\beta} \lambda\|\beta\|_2^2 + \frac{1}{2}\|\beta - Xy\|_2^2 \text{ Ridge regression}$$
$$\arg\min_{\beta} \lambda\|\beta\|_1 + \frac{1}{2}\|\beta - Xy\|_2^2 \text{ Lasso regression}$$

And for part a) and b), the solutions of these two problems are the proximal operators:

$$\beta_{\text{ridge}} = \frac{1}{1+2\lambda} Xy$$
$$\beta_{\text{lasso}} = \mathcal{S}_\lambda\left(Xy\right)$$

3. (Proximal gradient method)

   (a) The first-order approximation to a function $f : \mathbb{R}^p \to \mathbb{R}$ at $x \in \mathbb{R}^p$ equals

   $$f(x) + \nabla f(x)^T (y - x). \tag{6}$$

   We want to minimize this first-order approximation locally. To this end we fix a real constant $\alpha > 0$ and augment the approximation with an $\ell_2$-norm term that keeps us close to $x$,

   $$f_x(y) := f(x) + \nabla f(x)^T (y - x) + \frac{1}{2\,\alpha} ||y - x||_2^2. \tag{7}$$

   Prove that the minimum of $f_x$ is the gradient descent update $x - \alpha \nabla f(x)$.

   (b) Inspired by the previous question, how would you modify gradient descent to minimize a function of the form

   $$h(x) = f_1(x) + f_2(x), \tag{8}$$

   where $f_1$ is differentiable, and $f_2$ is nondifferentiable but has a proximal operator that is easy to compute?

   (c) Show that a vector $x^*$ is a solution to

   $$\text{minimize} \quad f_1(x) + f_2(x), \tag{9}$$

   where $f_1$ is differentiable, and $f_2$ is nondifferentiable, if and only if it is a fixed point of the iteration you proposed in the previous question for any $\alpha > 0$.

4. (Iterative shrinkage-thresholding algorithm)

   (a) What is the proximal gradient update corresponding to the lasso problem defined below? Your answer will involve a hyperparameter which we will call as $\alpha$.

   $$\frac{1}{2} ||y - X\beta||_2^2 + \lambda |\beta|_1$$

   (b) How would you check whether you have reached an optimum? How would you modify this to take into account possible numerical inaccuracies?

   (c) Implement the method and apply it to the problem in `pgd_lasso-question.ipynb`. You have to fill in blocks of code corresponds to the proximal gradient update step and termination condition. Report all the generated plots.