Exercise List: Proving convergence of the Gradient Descent Method on the Ridge Regression Problem.

Robert M. Gower and Francis Bach

September 14, 2019

1 Introduction

Ridge regression is perhaps the simplest example of a training problem in Machine Learning. Consider the task of learning a rule that maps the *feature vector* $x \in \mathbb{R}^d$ to outputs $y \in \mathbb{R}$. Furthermore you are given a set of labelled observations (x_i, y_i) for $i = 1, \ldots, n$. We restrict ourselves to linear mappings. That is, we need to find $w \in \mathbb{R}^d$ such that

$$x_i^\top w \approx y_i, \quad \text{for } i = 1, \ldots, n. \tag{1}$$

That is the *hypothesis function* is parametrized by w and is given by $h_w : x \mapsto w^\top x$.[1] To choose a w such that each $x_i^\top w$ is close to y_i, we use the squared loss $\ell(y) = y^2/2$ and the squared regularizor. That is, we minimize

$$w^* = \arg\min_w \frac{1}{n} \sum_{i=1}^n \frac{1}{2}(x_i^\top w - y_i)^2 + \frac{\lambda}{2}\|w\|_2^2, \tag{2}$$

where $\lambda > 0$ is the regularization parameter. We now have a complete training problem (2)[2].

With this simple ridge regression problem, we can illustrate many different techniques used in machine learning, such as using crossvalidation to select λ, dimension reduction tools, data scaling and stochastic optimization. In this exercise we will solve (2) using gradient descent, and we will establish how fast does gradient converge.

Using the matrix notation

$$X \overset{\text{def}}{=} [x_1, \ldots, x_n] \in \mathbb{R}^{d \times n}, \quad \text{and} \quad y = [y_1, \ldots, y_n] \in \mathbb{R}^n, \tag{3}$$

[1] We need only consider a linear mapping as opposed to the more general *affine* mapping $x_i \mapsto w^\top x_i + \beta$, because the zero order term $\beta \in \mathbb{R}$ can be incorporated by defining a new feature vectors $\hat{x}_i = [x_1, 1]$ and new variable $\hat{w} = [w, \beta]$ so that $\hat{x}_i^\top \hat{w} = x_i^\top w + \beta$

[2] Excluding the issue of selection λ using something like crossvalidation `https://en.wikipedia.org/wiki/Cross-validation_(statistics)`

we can re-write the objective function in (2) as

$$f(w) \overset{\text{def}}{=} \frac{1}{2n}\|X^\top w - y\|_2^2 + \frac{\lambda}{2}\|w\|_2^2. \tag{4}$$

First we introduce some necessary notation.

Notation: For every $x, w, \in \mathbb{R}^d$ let $\langle x, w\rangle \overset{\text{def}}{=} x^\top y$ and let $\|x\|_2 = \sqrt{\langle x, x\rangle}$. Let $A \in \mathbb{R}^{d\times d}$ be a matrix and let $\sigma_{\min}(A)$ and $\sigma_{\max}(A)$ be the smallest and largest singular values of A defined by

$$\sigma_{\min}(A) \overset{\text{def}}{=} \min_{x\in\mathbb{R}^d,\, x\neq 0} \frac{\|Ax\|_2}{\|x\|_2} \quad \text{and} \quad \sigma_{\max}(A) \overset{\text{def}}{=} \max_{x\in\mathbb{R}^d,\, x\neq 0} \frac{\|Ax\|_2}{\|x\|_2}. \tag{5}$$

Finally, a result you will need, if A is a symmetric positive semi-definite matrix the largest singular value of A can be defined instead as

$$\sigma_{\max}(A) = \max_{x\in\mathbb{R}^d,\, x\neq 0} \frac{\langle Ax, x\rangle_2}{\|x\|_2^2} = \max_{x\in\mathbb{R}^d,\, x\neq 0} \frac{\|Ax\|_2}{\|x\|_2}. \tag{6}$$

Therefore

$$\frac{\langle Ax, x\rangle_2}{\|x\|_2^2} \leq \sigma_{\max}(A), \quad \forall x \in \mathbb{R}^d \setminus \{0\}. \tag{7}$$

and

$$\frac{\|Ax\|_2}{\|x\|_2} \leq \sigma_{\max}(A), \quad \forall x \in \mathbb{R}^d \setminus \{0\}. \tag{8}$$

2 Gradient descent

We will now solve the following ridge regression problem

$$w^* = \arg\min_{w\in\mathbb{R}^d} \left(\frac{1}{2n}\|X^\top w - y\|_2^2 + \frac{\lambda}{2}\|w\|_2^2 \overset{\text{def}}{=} f(w)\right), \tag{9}$$

using gradient descent.

Ex. 1 — Consider the Gradient descent method

$$w^{t+1} = w^t - \alpha\nabla f(w^t), \tag{10}$$

where

$$\alpha = \frac{1}{\sigma_{\max}(A)}, \tag{11}$$

is a fixed stepsize and

$$A \overset{\text{def}}{=} \tfrac{1}{n}XX^\top + \lambda I. \tag{12}$$

Part I

Show that the gradient $\nabla f(x)$ of (9) is given by

$$\nabla f(w) = Aw - b = A(w - w^*),$$

where w^* is the solution to (9) and

$$b \overset{\text{def}}{=} \tfrac{1}{n} Xy.$$

Now that we have calculated the gradient, re-write the iterates (10) using this gradient.
Part II

Show or convince yourself that A as defined in (12) is positive semi-definite, that is

$$\langle Aw, w \rangle \geq 0, \quad \forall w \in \mathbb{R}^d, \tag{13}$$

and that

$$\sigma_{\max}(I - \alpha A) = 1 - \alpha\, \sigma_{\min}(A) = 1 - \frac{\sigma_{\min}(A)}{\sigma_{\max}(A)}. \tag{14}$$

Part III

Show that the iterates (10) converge to w^* according to

$$\|w^{t+1} - w^*\|_2 \leq \left(1 - \frac{\sigma_{\min}(A)}{\sigma_{\max}(A)}\right) \|w^t - w^*\|_2,$$

for all t. The number $(1 - \sigma_{\min}(A)/\sigma_{\max}(A))$ is known as the *rate of convergence.*
Hint 1: Subtract w^* from both sides of (10) and use the results from the previous two exercises.
Hint 2: Try and show that $b = Aw^*$!

Part IV

Let

$$\kappa(A) \overset{\text{def}}{=} \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)},$$

which is known as the condition number of A. What happens to κ as $\lambda \to \infty$ and $\lambda \to 0$, respectively? What does this imply about the speed at which gradient descent converges to the solution?

Part V

Let us consider the extreme case where $\lambda = 0$. Consider the coordinate change $\hat{w} = P^{-1}w$, where $P \in \mathbb{R}^{d\times d}$ is invertible. With this coordinate change we can solve the problem in $\hat{w}$ given by

$$\hat{w}^* = \arg\min_{\hat{w}\in\mathbb{R}^d} \left(\frac{1}{2n}\|X^\top P\hat{w} - y\|_2^2 + \frac{\lambda}{2}\|P\hat{w}\|_2^2 \right), \tag{15}$$

then switch back the coordinate system to get the solution in w^* given by

$$w^* = P\hat{w}^*. \tag{16}$$

If we use gradient descent to solve (15), at what rate does it converge? To get the fastest rate possible, what should P be? Does the choice

$$P = \operatorname{diag}(XX^\top)^{-1}, \tag{17}$$

make sense?

Extra question: Lookup and read about "batch normalization". Is it somehow related to preconditioning? Discuss with your colleagues.

Remark: The matrix P is known as the preconditioner and the particular choice given by (17) is a standard choice known as "feature scaling" and it is often used in machine learning.