Exercise List: Proving convergence of the (Stochastic)
Gradient Descent Method for the Least Squares Problem.

Robert M. Gower.
October 3, 2017

1 Introduction

This is an exercise in proving the convergence of iterative optimization methods. We will take a simple case study: solving the linear least squares problem, and prove the linear convergence of the gradient descent method and a variant of the stochastic gradient descent (SGD) method with importance sampling. This variant of SGD is also known as the randomized Kaczmarz method and the linear convergence we prove in Exe.2 was first established in [3].

First we introduce some necessary notation.

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

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

Thus clearly

\[
\frac{\|Ax\|_2}{\|x\|_2} \leq \sigma_{\text{max}}(A)^2, \quad \forall x \in \mathbb{R}^n.
\]

Let \(\|A\|_F^2 \overset{\text{def}}{=} \text{Tr}(A^\top A)\) denote the Frobenius norm of \(A\). Finally, a result you will need, is that for every symmetric positive semi-definite matrix \(G\) the \(L2\) induced matrix norm can be equivalently defined by

\[
\sigma_{\text{max}}(G) = \max_{x \in \mathbb{R}^n, x \neq 0} \frac{\langle Gx, x \rangle_2}{\|x\|_2} = \max_{x \in \mathbb{R}^n, x \neq 0} \frac{\|Gx\|_2}{\|x\|_2}.
\]
The Linear Least Squares Problem

Now consider the problem of solving the linear system

$$Ax = b,$$

where $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$. We assume that there exists a solution to (4). We also assume that $n \leq m$ and that $A$ has full column rank so that there is a unique solution $x^* \in \mathbb{R}^n$ to (4). We can recast (4) as the following Least Squares optimization problem

$$x^* = \arg \min_{x \in \mathbb{R}^n} \left( \frac{1}{2} \|Ax - b\|_2^2 \right) \overset{\text{def}}{=} f(x).$$

Exercises

Ex. 1 — Consider the Gradient descent method

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

where

$$\alpha = \frac{1}{\sigma_{\max}(A)^2},$$

is a fixed stepsize.

Part I

Show or convince yourself that

$$\sigma_{\max}(I - \alpha A^\top A) = 1 - \alpha \sigma_{\min}(A)^2 = 1 - \frac{\sigma_{\min}(A)^2}{\sigma_{\max}(A)^2}.$$
The least squares problem (5) can be re-written as

\[
\min_x \frac{1}{2} \|Ax - b\|_2^2 = \min_x \frac{1}{2} \sum_{i=1}^{m} (A_i x - b_i)^2 \overset{\text{def}}{=} \min_x \frac{1}{2} \sum_{i=1}^{m} f_i(x) \tag{11}
\]

where \( f_i(x) = (A_i x - b_i)^2 \), \( A_i \) denotes the \( i \)th row of \( A \) and \( b_i \) denotes the \( i \)th element of \( b \). Given this \textit{sum of terms} structure in (11) we can implement the stochastic gradient method as follows. From a given \( x^0 \in \mathbb{R}^n \), consider the iterates

\[
x^{t+1} = x^t - \alpha_j \nabla f_j(x^t), \tag{12}
\]

where

\[
\alpha_j = \frac{1}{\|A_j\|_2^2}, \tag{13}
\]

and \( j \) is a random index chosen from \{1, \ldots, m\} such that for every \( i \in \{1, \ldots, m\} \) the probability that \( j = i \) is given by \( \frac{\|A_i\|_2^2}{\|A\|_F^2} \). In other words, \( \mathbb{P}(j = i) = \frac{\|A_i\|_2^2}{\|A\|_F^2} \) for all \( i \in \{1, \ldots, m\} \).

\textbf{Part I}

Show that

\[
P_j \overset{\text{def}}{=} \alpha_j A_j^\top A_j = \frac{A_j^\top A_j}{\|A_j\|_2^2}, \tag{14}\]

is a projection operator which projects orthogonally onto \textbf{Range}(\( A_j \)). In other words, show that

\[
P_j P_j = P_j \quad \text{and} \quad (I - P_j)(I - P_j) = I - P_j. \tag{15}\]

Furthermore, verify that

\[
\mathbb{E}[P_j] = \sum_{i=1}^{m} \mathbb{P}(j = i) P_i = \frac{A^\top A}{\|A\|_F^2}. \tag{16}\]

\textbf{Part II}

Using analogous techniques from the previous exercise, show that the iterates (12) converge according to

\[
\mathbb{E}\left[\|x^{t+1} - x^*\|_2^2\right] \leq \left(1 - \frac{\sigma_{\min}(A)^2}{\|A\|_F^2}\right) \mathbb{E}\left[\|x^t - x^*\|_2^2\right]. \tag{17}\]

This is an amazing and recent result [3], since it shows that SGD converges exponentially fast despite the fact that the iterates (12) only require access to a single row of \( A \) at a time!
This result can be extended to any matrix $A$, including rank deficient matrices. Indeed, so long as there exists a solution to (4), the iterates (12) converge to the solution of least norm and at rate of 
\[ 1 - \frac{\sigma_{\min}^+(A)^2}{\|A\|_F^2} \] 
where $\sigma_{\min}^+(A)$ is the smallest nonzero singular value of $A$ [1]. Thus the assumption that $A$ has full column rank is not necessary. These results have also been extended to a general class of methods [2].

Part III

When is this stochastic gradient method (12) faster than the gradient descent method (6)? Note that the each iteration of SGD costs $O(n)$ floating point operations while an iteration of the GD method costs $O(nm)$ floating point operations. What happens if $m$ is very big? What if $\|A\|_F^2$ is very large? Discuss this.

References

