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

Robert M. Gower.

March 13, 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 $\langle x, y\rangle \stackrel{\text{def}}{=} x^\top y$ and let $\|x\|_2 = \sqrt{\langle x, x\rangle}$.

Let $\sigma_{\min}(A)$ and $\sigma_{\max}(A)$ be the smallest and largest singular values of A defined by

$$\sigma_{\min}(A) \stackrel{\text{def}}{=} \min_{x\in\mathbb{R}^n} \frac{\|Ax\|_2}{\|x\|_2} \quad \text{and} \quad \sigma_{\max}(A) \stackrel{\text{def}}{=} \max_{x\in\mathbb{R}^n} \frac{\|Ax\|_2}{\|x\|_2}. \tag{1}$$

Thus clearly

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

Let $\|A\|_F^2 \stackrel{\text{def}}{=} \mathrm{Tr}\left(A^\top A\right)$ 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

$$\sigma_{\max}(G) = \max_{x\in\mathbb{R}^n} \frac{\sqrt{\langle Gx, x\rangle_2}}{\|x\|_2} = \max_{x\in\mathbb{R}^n} \frac{\|Gx\|_2}{\|x\|_2}. \tag{3}$$

2 The Linear Least Squares Problem

Now consider the problem of solving the linear system

$$Ax = b, \tag{4}$$

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

$$x^* = \arg\min_{x\in\mathbb{R}^n} \left(\tfrac{1}{2}\|Ax-b\|_2^2 \overset{\text{def}}{=} f(x) \right). \tag{5}$$

3 Exercises

Ex. 1 — Consider the Gradient descent method

$$x^{t+1} = x^t - \alpha \nabla f(x^t), \tag{6}$$

where

$$\alpha = \frac{1}{\sigma_{\max}(A)^2}, \tag{7}$$

is a fixed stepsize.

Part I

Show or convince yourself that

$$\sigma_{\max}(I - \alpha A^\top A)^2 = 1 - \alpha\, \sigma_{\min}(A)^2 = 1 - \frac{\sigma_{\min}(A)^2}{\sigma_{\max}(A)^2}. \tag{8}$$

Part II

Calculate the gradient $\nabla f(x)$ of (5) and re-write the iterates (6) with this gradient.

Part III

Show that the iterates (6) converge to x^* according to

$$\|x^{t+1} - x^*\|_2^2 \leq \left(1 - \frac{\sigma_{\min}(A)^2}{\sigma_{\max}(A)^2}\right) \|x^t - x^*\|_2^2,$$

for all t.

Hint 1: Subtract x^* from both sides of (6) and use the results from the previous two exercises.

Hint 2: Remember that $b = Ax^*$!

Answer (Ex. I) — First note that

$$\begin{aligned}
\Big\langle (I-\alpha A^\top A)x, x\Big\rangle &= \|x\|_2^2 - \alpha\|Ax\|_2^2 \\
&\overset{(7)}{\geq} \|x\|_2^2 - \frac{\|Ax\|_2^2}{\sigma_{\max}(A)^2} \\
&\overset{(2)}{\geq} \|x\|_2^2 - \frac{\sigma_{\max}(A)^2\|x\|_2^2}{\sigma_{\max}(A)^2\|x\|_2^2} = 0,
\end{aligned}$$

thus the matrix $(I-\alpha A^\top A)$ is positive semi-definite and only has non-negative eigenvalues. Furthermore

$$\frac{\big\langle (I-\alpha A^\top A)x, x\big\rangle}{\|x\|_2^2} = 1-\alpha\frac{\big\langle A^\top Ax, x\big\rangle}{\|x\|_2^2}. \tag{9}$$

Since $(I-\alpha A^\top A)$ is symmetric positive semi-definite we can use (3) to calculate the induced norm, thus we have

$$\begin{aligned}
\sigma_{\max}(I-\alpha A^\top A)^2 &\overset{(3)+(9)}{=} \max_{x\in\mathbb{R}^n}\left(1-\alpha\frac{\big\langle A^\top Ax, x\big\rangle}{\|x\|_2^2}\right) \\
&= 1-\alpha\min_{x\in\mathbb{R}^n}\frac{\big\langle A^\top Ax, x\big\rangle}{\|x\|_2^2} \\
&= 1-\alpha\,\sigma_{\min}(A)^2.
\end{aligned}$$

Answer (Ex. II) — Differentiating we have

$$\nabla f(x) = A^\top(Ax-b) = A^\top A(x-x^*),$$

where the last equality follows since $Ax^* = b$. Consequently the gradient descent method (6) can be written as

$$x^{t+1} = x^t - \alpha A^\top A(x^t - x^*). \tag{10}$$

Answer (Ex. III) — Subtracting x^* from both sides of (10) gives

$$x^{t+1}-x^* = x^t - x^* - \alpha A^\top A(x^t-x^*) = (I-\alpha A^\top A)(x^t-x^*).$$

Taking norm squared in the above gives

$$\begin{aligned}
\|x^{t+1}-x^*\|_2^2 &\overset{(2)}{\leq} \sigma_{\max}\Big(I-\alpha A^\top A\Big)^2\|x^t-x^*\|_2^2 \\
&\overset{(8)}{=} (1-\alpha\,\sigma_{\min}(A)^2)\|x^t-x^*\|_2^2.
\end{aligned}$$

In particular for $\alpha = \frac{1}{\sigma_{\max}(A)^2}$ the above shows that

$$\|x^{t+1} - x^*\|_2^2 \quad \leq \quad \left(1 - \frac{\sigma_{\min}(A)^2}{\sigma_{\max}(A)^2}\right) \|x^t - x^*\|_2^2.$$

Ex. 2 — The least squares problem (5) can be re-written as

$$\min_x \tfrac{1}{2}\|Ax - b\|_2^2 \quad = \quad \min_x \tfrac{1}{2}\sum_{i=1}^m (A_{i:}x - b_i)^2 \quad \overset{\text{def}}{=} \quad \min_x \tfrac{1}{2}\sum_{i=1}^m f_i(x) \tag{11}$$

where $f_i(x) = (A_{i:}x - b_i)^2$, $A_{i:}$ denotes the ith row of A and b_i denotes the ith element of b. Given this *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 $i \in \{1, \ldots, m\}$.

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 $\mathbf{Range}\,(A_{j:})$. 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

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

Part II

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

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

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 $\left(1 - \frac{\sigma_{\min}^+(A)^2}{\|A\|_F^2}\right)$ 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.

Answer (Ex. I) — Verify by most all claim by direct computation. For instances

$$\mathbb{E}\left[P_j\right] = \sum_{i=1}^m \mathbb{P}(j=i) P_i = \sum_{i=1}^m \frac{\|A_{i:}\|_2^2}{\|A\|_F^2} \frac{A_{i:}^\top A_{i:}}{\|A_{i:}\|_2^2} = \sum_{i=1}^m \frac{A_{i:}^\top A_{i:}}{\|A\|_F^2} = \frac{A^\top A}{\|A\|_F^2}.$$

Answer (Ex. II) — First note that

$$\nabla f_j(x^t) = A_{j:}^\top (A_{j:} x - b_j) = A_{j:}^\top A_{j:} (x - x^*).$$

Using the above and subtracting x^* from both sides of (12) we have

$$\begin{aligned} x^{t+1} - x^* &= x^t - x^* - \alpha_j A_{j:}^\top A_{j:} (x^t - x^*) \\ &\overset{(13)}{=} \left(I - \frac{A_{j:}^\top A_{j:}}{\|A_{j:}\|_2^2}\right)(x^t - x^*). \end{aligned}$$

Taking norm squared in the above we have that

$$\begin{aligned} \|x^{t+1} - x^*\|_2^2 &= \|\left(I - \frac{A_{j:}^\top A_{j:}}{\|A_{j:}\|_2^2}\right)(x^t - x^*)\|_2^2 \\ &\overset{(15)}{=} \left\langle \left(I - \frac{A_{j:}^\top A_{j:}}{\|A_{j:}\|_2^2}\right)(x^t - x^*), x^t - x^* \right\rangle \\ &= \|x^t - x^*\|_2^2 - \left\langle \frac{A_{j:}^\top A_{j:}}{\|A_{j:}\|_2^2}(x^t - x^*), x^t - x^* \right\rangle. \end{aligned}$$

Taking expectation conditioned on x^t in the above gives

$$\begin{aligned}
\mathbb{E}\left[\|x^{t+1}-x^*\|_2^2 \,|\, x^t\right] &= \|x^t-x^*\|_2^2 - \left\langle \mathbb{E}\left[\frac{A_{j:}^\top A_{j:}}{\|A_{j:}\|_2^2}\right](x^t-x^*), x^t-x^* \right\rangle \\
&\overset{(16)}{=} \|x^t-x^*\|_2^2 - \frac{1}{\|A\|_F^2}\left\langle A^\top A(x^t-x^*), x^t-x^* \right\rangle \\
&\overset{(1)}{\leq} \|x^t-x^*\|_2^2 - \frac{\sigma_{\min}(A)^2}{\|A\|_F^2}\|x^t-x^*\|_2^2 \\
&= \left(1-\frac{\sigma_{\min}(A)^2}{\|A\|_F^2}\right)\|x^t-x^*\|_2^2.
\end{aligned}$$

It remains to take expectation in the above.

Answer (Ex. III) — ...

References

[1] R. M. Gower and P. Richtárik. “Stochastic Dual Ascent for Solving Linear Systems”. In: *arXiv:1512.06890* (2015).

[2] R. M. Gower and P. Richtárik. “Randomized Iterative Methods for Linear Systems”. In: *SIAM Journal on Matrix Analysis and Applications* 36.4 (2015), pp. 1660–1690.

[3] T. Strohmer and R. Vershynin. “A Randomized Kaczmarz Algorithm with Exponential Convergence”. In: *Journal of Fourier Analysis and Applications* 15.2 (2009), pp. 262–278.