Exercise List: Proving convergence of the Stochastic Gradient Descent and Coordinate Descent on the Ridge Regression Problem.

Robert M. Gower & Francis Bach & Nidham Gazagnadou

November 6, 2019

Introduction

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].

Using the matrix notation

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

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.

[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)`

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

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

and

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

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 stochastic gradient descent and stochastic coordinate descent.

Exercise 1 : Stochastic Gradient Descent (SGD)

Some more notation: Let $\|A\|_F^2 \overset{\text{def}}{=} \mathrm{Tr}\left(A^\top A\right)$ denote the Frobenius norm of A. Let

$$A \overset{\text{def}}{=} \tfrac{1}{n} X X^\top + \lambda I \in \mathbb{R}^{d \times d} \quad \text{and} \quad b \overset{\text{def}}{=} \tfrac{1}{n} X y. \tag{10}$$

We can exploit the separability of the objective function (2) to design a *stochastic* gradient method. For this, first we re-write the problem $Aw = b$ as different linear least squares problem

$$\hat{w}^* = \arg\min_w \tfrac{1}{2} \|Aw - b\|_2^2 \quad = \quad \arg\min_w \sum_{i=1}^d \tfrac{1}{2} (A_{i:} w - b_i)^2 \quad \overset{\text{def}}{=} \quad \arg\min_w \sum_{i=1}^d p_i f_i(w), \tag{11}$$

where $f_i(w) = \frac{1}{2p_i}(A_{i:} w - b_i)^2$, $A_{i:}$ denotes the ith row of A, b_i denotes the ith element of b and $p_i = \frac{\|A_{i:}\|_2^2}{\|A\|_F^2}$ for $i = 1, \ldots, d$. Note that $\sum_{i=1}^d p_i = 1$ thus the p_i's are probabilities.

From a given $w^0 \in \mathbb{R}^d$, consider the iterates

$$w^{t+1} = w^t - \alpha \nabla f_j(w^t), \tag{12}$$

where

$$\alpha = \frac{1}{\|A\|_F^2}, \tag{13}$$

and j is a random index chosen from $\{1, \dots, d\}$ sampled with probability p_j. In other words, $\mathbb{P}(j = i) = p_i = \frac{\|A_{i:}\|_2^2}{\|A\|_F^2}$ for all $i \in \{1, \dots, d\}$.

Ex. 1 — Show that the solution $\hat{w}^*$ to (11) and the solution to w^* to (9) are equal.

Answer (Ex. 1) — On the one hand, taking the gradient with respect to w in (9) leads to

$$\frac{1}{n} X(X^\top w - y) + \lambda w = \left(\frac{1}{n} X X^\top + \lambda I \right) w - \frac{1}{n} X y \stackrel{(10)}{=} Aw - b \ .$$

On the other hand, doing the same in (11), we get $A^\top(Aw - b) = A(Aw - b)$. So that, the argument of the minimum is given by the parameter for which the gradient is null:

$$w^* = \hat{w}^* = A^{-1} b \ .$$

Can we affirm that A is full rank because we added a regularization term so that solving $A(Aw - b) = 0$ leads to $(Aw - b) = 0$? ∎

Ex. 2 — Show that

$$\nabla f_j(w) = \frac{1}{p_j} A_{j:}^\top A_{j:}(w - w^*) \tag{14}$$

and that

$$\mathbb{E}_{j \sim p} \left[\nabla f_j(w) \right] \stackrel{\text{def}}{=} \sum_{i=1}^{d} p_i \nabla f_i(w) = A^\top A(w - w^*) \ ,$$

thus $\nabla f_j(w)$ is an unbiased estimator of the full gradient of the objective function in (11). This justifies applying the stochastic gradient method.

Answer (Ex. 2) — First note that

$$\nabla f_j(w) = \frac{1}{p_i} A_{j:}^\top (A_{j:} w - b_j) = \frac{1}{p_i} A_{j:}^\top A_{j:}(w - w^*) \ .$$

Taking expectation we have that

$$\mathbb{E}\left[\nabla f_j(w) \right] \quad = \quad \sum_{i=1}^{n} \frac{p_i}{p_i} A_{j:}^\top (A_{j:} w - b_j) = A^\top (Aw - b) = A^\top A(w - w^*) \ . \quad ∎$$

Ex. 3 — Let $\Pi_j \overset{\text{def}}{=} \frac{A_{j:}^\top A_{j:}}{\|A_{j:}\|_2^2}$, show that

$$\Pi_j \Pi_j = \Pi_j \ , \tag{15}$$

and

$$(I - \Pi_j)(I - \Pi_j) = I - \Pi_j. \tag{16}$$

In other words, Π_j is a projection operator which projects orthogonally onto $\mathbf{Range}\,(A_{j:})$. Furthermore, if $j \sim p_j$ verify that

$$\mathbb{E}\left[\Pi_j\right] = \sum_{i=1}^{d} p_i \Pi_i = \frac{A^\top A}{\|A\|_F^2}. \tag{17}$$

Answer (Ex. 3) — We check that Π_j is an orthogonal projector onto $\mathbf{Range}\,(A_{j:})$ by computing

$$\Pi_j \Pi_j = \frac{A_{j:}^\top A_{j:} A_{j:}^\top A_{j:}}{\|A_{j:}\|_2^2 \|A_{j:}\|_2^2} = \frac{A_{j:}^\top \|A_{j:}\|_2^2 A_{j:}}{\|A_{j:}\|_2^2 \|A_{j:}\|_2^2} = \frac{A_{j:}^\top A_{j:}}{\|A_{j:}\|_2^2} = \Pi_j \ ,$$

and

$$(I - \Pi_j)(I - \Pi_j) = I - 2\Pi_j + \Pi_j \Pi_j \overset{(15)}{=} I - \Pi_j \ .$$

Finally, we have

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

Ex. 4 — Show the following equality ruling the squared norm of the distance to the solution

$$\|w^{t+1} - w^*\|_2^2 \quad = \quad \|w^t - w^*\|_2^2 - \left\langle \frac{A_{j:}^\top A_{j:}}{\|A_{j:}\|_2^2}(w^t - w^*), w^t - w^* \right\rangle \ . \tag{18}$$

Answer (Ex. 4) — Using (14) and subtracting w^* from both sides of (12) we have

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

Taking norm squared in the above we have that

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

Ex. 5 — Using previous answer and analogous techniques from the course, show that the iterates (12) converge according to

$$
\mathbb{E}\left[\|w^{t+1}-w^*\|_2^2\right] \quad \leq \quad \left(1-\frac{\sigma_{\min}(A)^2}{\|A\|_F^2}\right)\mathbb{E}\left[\|w^t-w^*\|_2^2\right] \quad . \tag{19}
$$

Answer (Ex. 5) — Taking expectation conditioned on w^t in the above gives

$$
\begin{aligned}
\mathbb{E}\left[\|w^{t+1}-w^*\|_2^2 \,|\, w^t\right] &= \|w^t-w^*\|_2^2 - \left\langle \mathbb{E}\left[\frac{A_{j:}^\top A_{j:}}{\|A_{j:}\|_2^2}\right](w^t-w^*), w^t-w^* \right\rangle \\
&\overset{(17)}{=} \|w^t-w^*\|_2^2 - \frac{1}{\|A\|_F^2}\left\langle A^\top A(w^t-w^*), w^t-w^* \right\rangle \\
&\overset{(7)}{\leq} \|w^t-w^*\|_2^2 - \frac{\sigma_{\min}(A)^2}{\|A\|_F^2}\|w^t-w^*\|_2^2 \\
&= \left(1-\frac{\sigma_{\min}(A)^2}{\|A\|_F^2}\right)\|w^t-w^*\|_2^2.
\end{aligned}
$$

It remains to take expectation in the above. ∎

Remark: This is an amazing and recent result [2], since it shows that SGD converges exponentially fast despite the fact that the iterates (14) only require access to a single row of A at a time! This result can be extended to solving any linear system $Aw = b$, including the case where A rank deficient. Indeed, so long as there exists a solution to $Aw = b$, the iterates (14) 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 this method can solve any linear system.

BONUS

Exercise 2: Stochastic Coordinate Descent (CD)

Consider the minimization problem

$$w^* = \arg\min_{x\in\mathbb{R}^d} \left(f(w) \overset{\text{def}}{=} \frac{1}{2} w^\top A w - w^\top b \right), \tag{20}$$

where $A \in \mathbb{R}^{d\times d}$ is a symmetric positive definite matrix, and $w, b \in \mathbb{R}^d$.

Ex. 6 — First show that, using the notation (10), solving (20) is equivalent to solving (9).

Answer (Ex. 6) — Differentiating (20) or (9) in w gives

$$\nabla f(x) = Ax - b.$$

Consequently the unique solution w^* to both of these problems is given by $w^* = A^{-1}b$. ∎

Ex. 7 — Show that

$$\frac{\partial f(w)}{\partial w_i} = A_{i:} w - b_i \ , \tag{21}$$

where $A_{i:}$ is the ith row of A. Furthermore note that $w^* = A^{-1}b$, thus

$$\frac{\partial f(w)}{\partial w_i} = e_i^\top (Aw - b) = e_i^\top A(w - w^*) \ . \tag{22}$$

Answer (Ex. 7) — Follows immediately from $\nabla f(x) = Ax - b$ and $w^* = A^{-1}b$. ∎

Ex. 8 — **Question 2.3:** Consider a step of the stochastic coordinate descent method

$$w^{k+1} = w^k - \alpha_i \frac{\partial f(w^k)}{\partial x_i} e_i, \tag{23}$$

where $e_i \in \mathbb{R}^d$ is the ith unit coordinate vector, $\alpha_i = \dfrac{1}{A_{ii}}$, and $i \in \{1, \ldots, d\}$ is sampled i.i.d at each step according to $i \sim p_i$ where $p_i = \dfrac{A_{ii}}{\mathrm{Tr}\,(A)}$. Let $\|x\|_A^2 \overset{\text{def}}{=} x^\top A x$.

First, prove that

$$\|w^{k+1} - w^*\|_A^2 = \left\langle (I - \Pi_i^\top) A (I - \Pi_i)(w^k - w^*), w^k - w^* \right\rangle \ . \tag{24}$$

Answer (Ex. 8) — Subtracting w^* from both sides of (23) gives

$$\begin{aligned} w^{k+1} - w^* &\overset{(22)+(23)}{=} w^k - w^* - \alpha_i e_i^\top A(w^k - w^*) e_i \\ &= \left(I - \frac{e_i e_i^\top A}{A_{ii}}\right)(w^k - w^*). \end{aligned} \tag{25}$$

Let $\Pi_i = \frac{e_i e_i^\top A}{A_{ii}}$. Taking the squared norm $\|\cdot\|_A$ on both sides of (25) gives

$$\begin{aligned} \|w^{k+1} - w^*\|_A^2 &= \left\langle A(I - \Pi_i)(w^k - w^*), (I - \Pi_i)(w^k - w^*)\right\rangle \\ &= \left\langle (I - \Pi_i^\top) A (I - \Pi_i)(w^k - w^*), w^k - w^* \right\rangle. \end{aligned}$$

∎

Ex. 9 — **Question 2.4:** Let $r^k \overset{\text{def}}{=} A^{1/2}(w^k - w^*)$. Deduce from (24) that

$$\|r^{k+1}\|_2^2 = \|r^k\|_2^2 - \left\langle \frac{A^{1/2} e_i e_i^\top A^{1/2}}{A_{ii}} r^k, r^k \right\rangle . \tag{26}$$

Answer (Ex. 9) — Let $r^k = A^{1/2}(w^k - w^*)$ and note that

$$(I - \Pi_i^\top) A (I - \Pi_i) = A - 2A\Pi_i + \Pi_i^\top A \Pi_i = A - \frac{A e_i e_i^\top A}{A_{ii}}.$$

Using this we have from (24) that

$$\begin{aligned} \|r^{k+1}\|_2^2 &= \left\langle \left(A - \frac{A e_i e_i^\top A}{A_{ii}}\right)(w^k - w^*), w^k - w^* \right\rangle \\ &= \|r^k\|_2^2 - \left\langle \frac{A e_i e_i^\top A}{A_{ii}}(w^k - w^*), w^k - w^* \right\rangle \\ &= \|r^k\|_2^2 - \left\langle \frac{A^{1/2} e_i e_i^\top A^{1/2}}{A_{ii}} r^k, r^k \right\rangle. \quad \blacksquare \end{aligned} \tag{27}$$

Ex. 10 — Finally, prove the convergence of the iterates of CD (23) converge according to

$$\mathbb{E}\left[\|w^{k+1} - w^*\|_A^2\right] \leq \left(1 - \frac{\lambda_{\min}(A)}{\mathrm{Tr}\,(A)}\right) \mathbb{E}\left[\|w^k - w^*\|_A^2\right] \tag{28}$$

thus (23) converges to the solution.
Hint: Since A is symmetric positive definite you can use that

$$\lambda_{\min}(A) = \inf_{x \in \mathbb{R}^d, x \neq 0} \frac{x^\top A x}{\|x\|_2^2}.$$

You will need to use that $x^\top A x \geq \lambda_{\min}(A)\|x\|_2^2$ at some point.

Answer (Ex. 10) — Taking expectation conditioned on r^k over the second term in (27) gives

$$\begin{aligned}
\mathbb{E}\left[\left\langle \frac{A^{1/2}e_i e_i^\top A^{1/2}}{A_{ii}} r^k, r^k \right\rangle \mid r^k\right] &= \sum_{j=1}^{n} \frac{A_{jj}}{\mathrm{Tr}\,(A)} \left\langle \frac{A^{1/2}e_j e_j^\top A^{1/2}}{A_{jj}} r^k, r^k \right\rangle \\
&= \frac{1}{\mathrm{Tr}\,(A)} \left\langle A^{1/2} \sum_{j=1}^{n} e_j e_j^\top A^{1/2} r^k, r^k \right\rangle \\
&= \frac{1}{\mathrm{Tr}\,(A)} \left\langle A r^k, r^k \right\rangle \\
&\geq \frac{\lambda_{\min}(A)}{\mathrm{Tr}\,(A)} \|r^k\|_2^2.
\end{aligned}$$

Consequently taking expectation conditioned on r^k in (26) gives

$$\mathbb{E}\left[\|r^{k+1}\|_2^2 \,|\, r^k\right] \leq \left(1 - \frac{\lambda_{\min}(A)}{\mathrm{Tr}\,(A)}\right) \|r^k\|_2^2. \tag{29}$$

It now remains to take expectation and re-write $\|r^k\|_2^2 = \|w^k - w^*\|_A^2$. ∎

Ex. 11 — **Question 2.6:** When is this stochastic coordinate descent method *faster* than the stochastic gradient method (14) or gradient descent? Note that each iteration of SGD and CD costs $O(d)$ floating point operations while an iteration of the GD method costs $O(d^2)$ floating point operations (assuming that A has been previously calculated and stored). What happens if d is very big? What if $\mathrm{Tr}\,(A)$ is very large? Discuss this.

Answer (Ex. 11) — Let

$$\kappa_{SGD} \stackrel{\text{def}}{=} \frac{\|A\|_F^2}{\sigma_{\min}^2(A)} = \frac{\mathrm{Tr}\left(A^\top A\right)}{\sigma_{\min}^2(A)} = \sum_{i=1}^{d} \frac{\sigma_i^2(A)}{\sigma_{\min}^2(A)},$$

be the complexity constant of SGD and let

$$\kappa_{CD} \quad \overset{\text{def}}{=} \quad \frac{\operatorname{Tr}(A)}{\lambda_{\min}(A)} = \sum_{i=1}^{d} \frac{\sigma_i(A)}{\sigma_{\min}(A)},$$

be the complexity constant of CD, where we used that A is positive semi-definite so that $\lambda_i(A) = \sigma_i(A)$.
Consider the extreme case where $\sigma_i(A) = \sigma_j(A)$ for every $i, j \in \{1, \ldots, d\}$. In this case $\kappa_{SGD} = d = \kappa_{CD}$.
Now consider the case that the singular values are evenly spread out with $\sigma_i(A) = i \times \tau$ where $\tau > 0$. In this case

$$kappa_{SGD} = \sum_{i=1}^{d} \frac{i^2 \times \tau^2}{\tau^2} = O(d^3)$$

and

$$kappa_{CD} = \sum_{i=1}^{d} \frac{i \times \tau}{\tau} = O(d^2).$$

Essentially, the complexity of coordinated descent κ_{CD} is far less sensitive to *ill-conditioned* data, that is, data where the smallest and the largest singular values are far apart.

References

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

[2] 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.