EECS 225A Statistical Signal Processing Lecture 16 - 3/17/2020

Lecture 16: Gradient Descent and Least Mean Squares Algorithm

Lecturer: Jiantao Jiao *Scribe: Aviral Pandey*

In this note we will discuss the gradient descent (GD) algorithm and the Least-Mean-Squares (LMS) algorithm, where we will interpret the LMS algorithm as a special instance of stochastic gradient descent (SGD). In this lecture everything is real-valued.

Recall the setting of least squares below. We would like to find a coefficient w such that $y_i \approx x_i^T w$ for each $1 \leq i \leq m$. It gives rise to the least squares formulation:

$$\min_w \sum_{i=1}^m (y_i - x_i^T w)^2 \iff \min_w \|y - Xw\|_2^2, \tag{1}$$

where

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}, X = \begin{bmatrix} - & x_1^T & - \\ - & x_2^T & - \\ & \vdots & \\ - & x_m^T & - \end{bmatrix}$$

Assuming X is full column rank (which requires $m \geq n$), we have the optimal solution given by

$$w_* = (X^T X)^{-1} X^T y \tag{2}$$

However, computing this optimal solution requires $O(mn^2)$ running time. The recursive least squares algorithm computes it in an recursive fashion in m steps where each step requires complexity $O(n^2)$. We want to address the following two questions in this lecture:

1. Suppose we are fine with $O(n^2)$ per-step computational complexity, but want to generalize RLS to general problems. How can we do that?
2. Suppose $O(n^2)$ is too high complexity even for a single step and we can only do $O(n)$. Is there anything we can still do?

There are many answers to these two questions, and our answers in this lecture are GD and SGD, respectively.

1 Gradient Descent

We would like to introduce the general framework of gradient descent in convex optimization. For consistency the optimization variable will be denoted as x rather than w.

Say we want to minimize the function $f(x)$ where $x \in \mathbb{R}^n$.

1.1 Convex functions

We assume that the function f satisfies the following properties:

1. f is convex;
2. f has Lipschitz gradient:

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|. \tag{3}$$

There are some interesting consequences of the Lipschitz gradient assumptions. In particular:

Lemma 1. *[1, Theorem 2.1.5] Suppose (3) holds. Then,*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|x - y\|^2 \tag{4}$$

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2 \tag{5}$$

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2 \tag{6}$$

The gradient descent algorithm starts with an initial point $x_0 \in \mathbb{R}^n$ and for each $k \geq 0$ computes the iterates

$$x_{k+1} = x_k - h_k \nabla f(x_k). \tag{7}$$

For simplicity we assume that $h_k \equiv h > 0$. Denote by x_* an arbitrary optimal point of our problem and let $f_* = f(x_*)$.

The following theorem characterizes the performance of gradient descent.

Theorem 2. *[1, Theorem 2.1.14] Let f be convex with Lipschitz gradient with constant L, and $0 < h < \frac{2}{L}$. Then, the gradient method generates a sequence of points $\{x_k\}$ such that for any $k \geq 0$,*

$$f(x_k) - f_* \leq \frac{2(f(x_0) - f_*)\|x_0 - x_*\|^2}{2\|x_0 - x_*\|^2 + kh(2 - Lh)(f(x_0) - f_*)}. \tag{8}$$

Proof Let $r_k = \|x_k - x_*\|$. Then,

$$r_{k+1}^2 = \|x_k - x_* - h\nabla f(x_k)\|^2 \tag{9}$$

$$= r_k^2 - 2h\langle \nabla f(x_k), x_k - x_* \rangle + h^2\|\nabla f(x_k)\|^2. \tag{10}$$

It follows from (6) that

$$\langle \nabla f(x_k), x_k - x_* \rangle \geq \frac{1}{L}\|\nabla f(x_k)\|^2 \tag{11}$$

since $\nabla f(x_*) = 0$. Hence, we have

$$r_{k+1}^2 \leq r_k^2 - \frac{2h}{L}\|\nabla f(x_k)\|^2 + h^2\|\nabla f(x_k)\|^2 \tag{12}$$

$$= r_k^2 - h(2/L - h)\|\nabla f(x_k)\|^2. \tag{13}$$

Since $0 < h < \frac{2}{L}$, we have $r_k^2 \leq r_0^2$.

We also have

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \tag{14}$$

$$= f(x_k) + \langle \nabla f(x_k), -h\nabla f(x_k) \rangle + \frac{L}{2}\| - h\nabla f(x_k)\|^2 \tag{15}$$

$$= f(x_k) - \omega\|\nabla f(x_k)\|^2, \tag{16}$$

where $\omega = h(1 - Lh/2)$ and we used (4). Denote $\Delta_k = f(x_k) - f_*$. We also have

$$\Delta_k = f(x_k) - f_* \tag{17}$$

$$\leq \langle \nabla f(x_k), x_k - x_* \rangle \tag{18}$$

$$\leq r_0\|\nabla f(x_k)\|, \tag{19}$$

where the first inequality is due to the convexity of f and the second inequality follows from Cauchy–Schwarz. Therefore,

$$\Delta_{k+1} = f(x_{k+1}) - f_* \tag{20}$$
$$\leq f(x_k) - f_* - \omega \|\nabla f(x_k)\|^2 \tag{21}$$
$$\leq f(x_k) - f_* - \omega \frac{\Delta_k^2}{r_0^2} \tag{22}$$
$$= \Delta_k - \frac{\omega}{r_0^2}\Delta_k^2, \tag{23}$$

where the first step follows from (16) and the second step follows from (19).

Dividing both sides by $\Delta_k \Delta_{k+1}$, we have

$$\frac{1}{\Delta_{k+1}} \geq \frac{1}{\Delta_k} + \frac{\omega}{r_0^2}\frac{\Delta_k}{\Delta_{k+1}} \tag{24}$$
$$\geq \frac{1}{\Delta_k} + \frac{\omega}{r_0^2}. \tag{25}$$

Summing up the inequalities, we have

$$\frac{1}{\Delta_{k+1}} \geq \frac{1}{\Delta_0} + \frac{\omega}{r_0^2}(k+1). \tag{26}$$

□

In order to choose the optimal step size, we need to maximize the function $h(2 - Lh)$ w.r.t h, which is given by $h = \frac{1}{L}$. In this case, the final bound reads

$$f(x_k) - f(x_*) \leq \frac{2L(f(x_0) - f_*)\|x_0 - x_*\|^2}{2L\|x_0 - x_*\|^2 + k(f(x_0) - f_*)}. \tag{27}$$

We can also upper bound $f(x_0) - f_*$ using $\|x_0 - x_*\|^2$. Indeed, we have

$$f(x_0) - f_* \leq \langle \nabla f(x_*), x_0 - x_* \rangle + \frac{L}{2}\|x_0 - x_*\|^2 \leq \frac{L}{2}\|x_0 - x_*\|^2. \tag{28}$$

Since the RHS of (27) is increasing in $f(x_0) - f_*$, we have the following bound when $h = \frac{1}{L}$:

$$f(x_k) - f_* \leq \frac{2L\|x_0 - x_*\|^2}{k+4}. \tag{29}$$

Let's interpret this result in the case of quadratic function *without* assuming that X has full column rank.

$$f(w) = \|y - Xw\|_2^2, w \in \mathbb{R}^{n\times 1}, X \in \mathbb{R}^{m\times n}$$
$$\nabla f(w) = -2X^T(y - Xw)$$

The gradient $\nabla f(w)$ is Lipschitz with constant $L = 2\lambda_{\max}(X^T X)$. The previous general theorem tells us that we need to guarantee that $h < \frac{2}{L} = \frac{1}{\lambda_{\max}(X^T X)}$. We will now show through the special case of quadratic functions that this requirement is tight.

Indeed, with the gradient descent formula

$$w_t = w_{t-1} - h\nabla f(w_{t-1}), \tag{30}$$

applied to this case, we have

$$\begin{aligned} w_t &= w_{t-1} + 2hX^T(y - Xw_{t-1}) \\ &= w_{t-1} + 2h(X^Ty - X^TXw_{t-1}) \\ &= w_{t-1} - 2hX^TX(w_{t-1} - w_*), \end{aligned}$$

where w_* satisfies $X^Ty = X^TXw_*$. In other words, w_* is *any* solution to the normal equation. Let $v_t \triangleq w_t - w_*$, then

$$v_t = v_{t-1} - 2h(X^TX)v_{t-1} = (I - 2hX^TX)v_{t-1}, \tag{31}$$

which implies that

$$v_t = (I - 2hX^TX)^t v_0. \tag{32}$$

We would like $\|v_t\|_2 \to 0$ for any arbitrary v_0.

If X has full column rank, then it is possible since we only need to guarantee that all the eigenvalues of $(I - 2hX^TX) \in (-1, 1)$. Its eigenvalues are certainly strictly less than one (full column rank assumption), and to ensure that it is strictly better than -1 we need

$$h < \frac{1}{\lambda_{\max}(X^TX)}. \tag{33}$$

However, if X does not have full column rank it is impossible to drive $\|v_t\|$ zero for arbitrary v_0. Indeed, if we pick v_0 such that $Xv_0 = 0, v_0 \neq 0$, then $v_t \equiv v_0$.

This example explains why Theorem 2 does not claim $\|x_k - x_*\| \to 0$ since it is in general impossible to guarantee, and even poorly defined: the minimizer of a general convex function may not be a point. It may be a set. The minimum is always attained if we consider convex functions in a compact set.

Theorem 2 guarantees that $f(w_t) - f(w_*)$ converges to zero. When we specialize to this setting, we can show that [1]

$$f(w_t) - f(w_*) = \|X(w_t - w_*)\|^2, \tag{34}$$

and $X(w_t - w_*) = Xv_t$ satisfies

$$Xv_t = X(I - 2hX^TX)^t v_0, \tag{35}$$

which annihilates any v_0 satisfying $Xv_0 = 0$.

As we will see, if we assume the function f is strongly convex, we can indeed guarantee that $\|x_k - x_*\| \to 0$.

1.2 Strongly convex functions

In additional to the Lipschitz gradient assumption in (3), we assume that the function f is strongly convex:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|x - y\|^2. \tag{36}$$

By symmetry we also have

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2}\|x - y\|^2. \tag{37}$$

Summing up these two inequalities, we have

$$0 \geq \langle \nabla f(x) - \nabla f(y), y - x \rangle + \mu\|x - y\|^2, \tag{38}$$

[1] Hint: apply the normal equation.

which is the same as

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2. \tag{39}$$

It turns out that for strongly convex functions with Lipschitz gradient, we can combine it with (6) to get a *weighted* lower bound on $\langle \nabla f(x) - \nabla f(y), x - y \rangle$:

Lemma 3. *[1, Theorem 2.1.12] Suppose f is strongly convex (36) with Lipschitz gradient (3). Then for any $x, y \in \mathbb{R}^n$ we have*

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{\mu L}{\mu + L} \|x - y\|^2 + \frac{1}{\mu + L} \|\nabla f(x) - \nabla f(y)\|^2. \tag{40}$$

If f is second-order differentiable, then we can obtain another equivalent condition on the Hessian of f summarizing the Lipschitz gradient assumptions and the strong convexity assumption:

$$\mu I \preceq \nabla^2 f(x) \preceq LI \tag{41}$$

We now show that the convergence rate of gradient descent for strongly convex functions.

Theorem 4. *[1, Theorem 2.1.15] Suppose $0 < h \leq \frac{2}{\mu + L}$ and f is strongly convex with Lipschitz gradient. Then the gradient descent algorithm satisfies*

$$\|x_k - x_*\|^2 \leq \left(1 - \frac{2h\mu L}{\mu + L}\right)^k \|x_0 - x_*\|^2. \tag{42}$$

If $h = \frac{2}{\mu + L}$, then

$$\|x_k - x_*\| \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^k \|x_0 - x_*\|, \tag{43}$$

$$f(x_k) - f_* \leq \frac{L}{2} \left(\frac{\kappa - 1}{\kappa + 1}\right)^{2k} \|x_0 - x_*\|^2, \tag{44}$$

where $\kappa = \frac{L}{\mu}$ is the condition number.

Proof Let $r_k = \|x_k - x_*\|$. Then,

$$r_{k+1}^2 = \|x_k - x_* - h\nabla f(x_k)\|^2 \tag{45}$$

$$= r_k^2 - 2h\langle \nabla f(x_k), x_k - x_* \rangle + h^2 \|\nabla f(x_k)\|^2. \tag{46}$$

Applying Lemma 3, we have

$$\langle \nabla f(x_k), x_k - x_* \rangle = \langle \nabla f(x_k) - \nabla f(x_*), x_k - x_* \rangle \tag{47}$$

$$\geq \frac{\mu L}{\mu + L} \|x_k - x_*\|^2 + \frac{1}{\mu + L} \|\nabla f(x_k)\|^2. \tag{48}$$

Hence,

$$r_{k+1}^2 \leq r_k^2 - \frac{2h\mu L}{\mu + L} \|x_k - x_*\|^2 - \frac{2h}{\mu + L} \|\nabla f(x_k)\|^2 + h^2 \|\nabla f(x_k)\|^2 \tag{49}$$

$$= \left(1 - \frac{2h\mu L}{\mu + L}\right) r_k^2 + h \left(h - \frac{2}{\mu + L}\right) \|\nabla f(x_k)\|^2. \tag{50}$$

For all $h \in (0, 2/(\mu + L)]$, the term $h\left(h - \frac{2}{\mu+L}\right) \leq 0$. Hence we have

$$\|x_k - x_*\|^2 \leq \left(1 - \frac{2h\mu L}{\mu + L}\right)^k \|x_0 - x_*\|^2. \tag{51}$$

The bound on $f(x_k) - f_*$ follows from (28) and existing results on $\|x_k - x_*\|^2$. □

For all $h \in (0, 2/(\mu + L)]$, the best exponent $\left(1 - \frac{2h\mu L}{\mu+L}\right)$ is achieved when $h = \frac{2}{\mu+L}$, and the best exponent is

$$1 - \frac{2\mu L}{\mu + L}\frac{2}{\mu + L} = \frac{(\mu + L)^2 - 4\mu L}{(\mu + L)^2} = \frac{(\mu - L)^2}{(\mu + L)^2}. \tag{52}$$

Again, it turns out to be the best choice even for quadratic functions with known μ, L as we show below.

Recall (32). Suppose we sort all of the eigenvalues of $X^T X$ as

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_n$$

with corresponding unit norm eigenvectors $\{u_i\}$, we have

$$v_t = \sum_{i=1}^{n} (1 - 2h\lambda_i)^t u_i (u_i^T v_0)$$

The convergence of v_t depends on the magnitude of $(1 - 2h\lambda_i)$. We can formulate finding the optimal h as the following problem:

$$\arg\min_h \max_i \{|1 - 2h\lambda_i|\} = \arg\min_h \max\{|1 - 2h\lambda_1|, |1 - 2h\lambda_n|\} \tag{53}$$

$$= \frac{1}{\lambda_1 + \lambda_n}, \tag{54}$$

where in the first step we used the fact that $|1 - hx|$ is a convex function of x and attains its maximum only at extremes, and for the second step we solved this convex optimization problem for h and the solution is the one achieving $|1 - 2h\lambda_1| = |1 - 2h\lambda_n|$. For this problem, $\mu = 2\lambda_n, L = 2\lambda_1$, and this answer to h is exactly $\frac{2}{\mu+L}$.

There exists another method to prove Theorem 4 that is a precise generalization of the argument we just did for the quadratic function case. For the sake of completeness, we present it here.

$$\begin{aligned}
\|x_k - h\nabla f(x_k) - x_*\| &= \|x_k - h\nabla f(x_k) - (x_* - h\nabla f(x_*))\| && \nabla f(x_*) = 0 \\
&= \left\| \int_0^1 (I - h\nabla^2 f(x_* + t(x_k - x_*)))(x_k - x_*) dt \right\| && \text{Mean value theorem} \\
&\leq \sup_t \left\| I - h\nabla^2 f(x_* + t(x_k - x_*)) \right\| \|x_k - x_*\| \\
&\leq \max\{|1 - h\mu|, |1 - hL|\} \|x_k - x_*\| && \text{Lipschitz gradient and strong convexity}
\end{aligned}$$

which results in the recursion

$$\|x_{k+1} - x_*\|^2 \leq \max\{|1 - h\mu|^2, |1 - hL|^2\} \|x_k - x_*\|^2 .$$

We can then choose

$$h = \arg\min_h \max\{|1 - h\mu|^2, |1 - hL|^2\} = \frac{2}{\mu + L}$$

for optimal rate of convergence.

2 Least Mean Squares (LMS)

Let's go back to the least squares problem:

$$\frac{1}{m}\|y - Xw\|_2^2 = \frac{1}{m}\sum_{i=1}^{m}(y_i - x_i^T w)^2$$

The LMS algorithm initiates at some point w_1 and at each step $i \geq 1$, it picks the function $(y_i - x_i^T w)^2$ and performs gradient descent on this function.

$$\begin{aligned}\frac{\partial(y_i - x_i^T w)^2}{\partial w} &= -2x_i(y_i - x_i^T w)\\ w_{i+1} &= w_i - h\frac{\partial(y - x_i^T w)^2}{\partial w}|_{w=w_i}\\ &= w_i + 2hx_i(y_i - x_i^T w_i)\end{aligned}$$

Generalizing the above, consider the optimization problem

$$\min_w \frac{1}{m}\sum_{t=1}^{m} f_t(w)$$

Then the generalized LMS update is:

$$w_{t+1} = w_t - h_t \nabla f_t(w_t)$$

The following theorem gives an *regret* interpretation of the updates above. In the next lecture, we will present additional conditions to guarantee that this procedure indeed is solving the least squares problem in expectation.

Theorem 5. *Suppose $\|\nabla f_t(w_t)\| \leq G, \forall t$, using the above algorithm and picking $h = \frac{1}{\sqrt{m}}$, then*

$$\frac{1}{m}\sum_{t=1}^{m}\langle \nabla f_t(w_t), w_t - w_*\rangle \leq \frac{\|w_1 - w_*\|_2^2 + G^2}{2\sqrt{m}}$$

Furthermore, if f_t is convex, then

$$\frac{1}{m}\sum_{t=1}^{m}(f_t(w_t) - f_t(w_*)) \leq \frac{\|w_1 - w_*\|_2^2 + G^2}{2\sqrt{m}}$$

Proof

$$\begin{aligned}\|w_{t+1} - w_*\|^2 &= \|w_t - h_t\nabla f_t(w_t) - w_*\|_2^2\\ &= \|w_t - w_*\|^2 + h_t^2\|\nabla f_t(w_t)\|_2^2 - 2h_t\langle \nabla f_t(w_t), w_t - w_*\rangle\\ \Rightarrow \langle \nabla f_t(w_t), w_t - w_*\rangle &\leq \frac{\|w_t - w_*\|_2^2 - \|w_{t+1} - w_*\|_2^2}{2h_t} + \frac{h_t}{2}G^2\end{aligned}$$

Summing over t,

$$\begin{aligned}\sum_{t=1}^{m}\langle \nabla f_t(w_t), w_t - w_*\rangle &\leq \frac{\|w_1 - w_*\|_2^2}{2h_1} + \sum_{t=2}^{m}\|w_t - w_*\|_2^2\left(\frac{1}{2h_t} - \frac{1}{2h_{t-1}}\right) + \frac{G^2}{2}\sum_{t=1}^{m}h_t\\ &= \frac{\|w_1 - w_*\|_2^2}{\frac{2}{\sqrt{m}}} + \frac{G^2}{2}\frac{m}{\sqrt{m}}\\ &= \left(\|w_1 - w_*\|_2^2 + G^2\right)\frac{\sqrt{m}}{2}\end{aligned}$$

The final conclusion follows from dividing both sides by m. The last statement of the theorem follows from the definition of convexity $f_t(w_t) - f_t(w_*) \leq \langle \nabla f_t(w_t), w_t - w_* \rangle$. $\square$

References

[1] Y. Nesterov, *Lectures on convex optimization*. Springer, 2018, vol. 137.