The University of Texas at Austin
Department of Electrical and Computer Engineering

EE381K: Convex Optimization — Fall 2019

Lecture 26

Aryan Mokhtari Wednesday, December 4, 2019.

Goal: In this lecture we talk about the convergence rate of gradient descent method when we use Exact Line Search or Backtracking Line Search and the objective function is smooth and strongly-convex. We also study its convergence rate when we minimize a convex and smooth function.

1 Exact Line Search

We've proven that using $\eta = \frac{1}{M}$ as the step size for gradient descent always provides an acceptable convergence. The problem with this method is that M is usually not known. There are several other methods for choosing the step size, including step sizes that vary for each iteration. Of the latter, the most straightforward method is exact line search, which calculates the optimal step size for every iteration. This results in each iteration having an optimization problem of its own.
Algorithm (Gradient descent with exact line search)

1. Set iteration counter $k = 0$, and make an initial guess $\mathbf{x}_0$ for the minimum
2. Compute $\nabla f(\mathbf{x}_k)$
3. Choose $\eta_k = \arg\min_{\eta}\{f\left(\mathbf{x}_k - \eta \nabla f(\mathbf{x}_k)\right)\}$
4. Update $\mathbf{x}_{k+1} = \mathbf{x}_k - \eta_k \nabla f(\mathbf{x}_k)$ and $k = k + 1$.
5. Go to 2 until $\|\nabla f(\mathbf{x}_k)\| < \epsilon$

Exact line search is used when the cost of the minimization problem with one variable is low compared to the cost of computing the search direction.

1.1 Rate of Convergence for Exact Line Search

We can determine the rate of convergence by comparing exact line search to the previous fixed-step descent. If we follow an update of $\mathbf{x}^+ = \mathbf{x} - \eta \nabla f(\mathbf{x})$ then we have

$$f(\mathbf{x} - \eta \nabla f(\mathbf{x})) \leq f(\mathbf{x}) - \nabla f(\mathbf{x})^\top (-\eta \nabla f(\mathbf{x})) + \frac{M}{2} \| - \eta \nabla f(\mathbf{x})\|^2$$

which is equivalent to

$$f(\mathbf{x} - \eta \nabla f(\mathbf{x})) \leq f(\mathbf{x}) - \left(\eta - \frac{M\eta^2}{2}\right) \|\nabla f(\mathbf{x})\|^2$$

If we minimize the left hand side with respect to η then we get the iterate obtained by following the exact line-search method and if we minimize the right hand side with respect to η we obtain that $\eta = 1/M$. Hence we can write

$$f(\hat{\mathbf{x}}_{k+1}) \leq f(\mathbf{x}_k) - \frac{1}{2M} \|f(\mathbf{x}_k)\|^2$$

where $\hat{\mathbf{x}}_{k+1} = \mathbf{x}_k - \eta_{ELS}\nabla f(\mathbf{x}_k)$
$f(\hat{\mathbf{x}}_{k+1})$ thus fulfills step 3 of the proof for the theorem in the previous lecture, the rest of the proof can be followed identically for this case. Therefore exact line search also converges linearly with rate $(1 - \frac{m}{M})$.

2 Gradient Descent using Backtracking Line Search

Backtracking line search is another way to compute the step size to be taken in each iteration of the *gradient descent* method. The best step size to choose at each iteration can be obtained through *exact line search* which involves solving a one dimensional optimization problem. However, it may be computationally inefficient to solve an optimization problem at each iteration. Thus, a natural approximation is to choose a step size which is not necessarily the best possible, but one that is reasonably good, i.e., produce 'enough' decrease in the objective f at each iteration. One such way of ensuring that the decrease in the function value is 'sufficient' in successive iterations without having to solve an optimization problem is by the method of *backtracking line search*.

Backtracking Line Search (BTLS) is defined by two parameters α, β with $0 < \alpha < 0.5$ and $0 < \beta < 1$. A step size η is defined to be '*good*' at a point x if

$$f(\mathbf{x} - \eta\nabla f(\mathbf{x})) \leq f(\mathbf{x}) - \alpha\eta||\nabla f(\mathbf{x})||^2 \tag{1}$$

The way to find such a good step size η in any iteration is by first starting at large initial test step size η_0 (usually taken 1) and check if this step size satisfies (1). If it is satisfied, then one proceeds with this step size, else reduces the step size by a factor of β and tests it again. The algorithm is formally described below.

Algorithm 1 Gradient Descent with Backtracking Line Search

```
x ← 0
η_0 ← 1
iterations ← 1
N ← number of iterations
for iterations ≤ N do
    η ← η_0
    while f(x − η∇f(x)) > f(x) − αη||∇f(x)||^2 do
        η ← βη
    end while
    x ← x − η∇f(x)
    iterations ← iterations +1
end for
return x and f(x)
```

For BTLS to be a viable option, we need to ensure that

1. The while loop in Algorithm 1 terminates, i.e., there always exists a good $\eta > 0$ for every x.
2. The final step size η at each iteration should not be too small to have a significant descent.

It is shown that BTLS ensures the above conditions in the following section.

2.1 Analysis of Gradient Descent using BTLS

For the analysis, we assume that the function f is strictly convex and f is differentiable at every point.

Claim: For any strictly convex function f, $\eta \leq \frac{1}{M}$ satisfies (1) for all x.

Proof. Since f is strictly convex, for all x, y

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

Substituting $y = x - \eta \nabla f(\mathbf{x})$ in (2), we get

$$f(x - \eta \nabla f(\mathbf{x})) \leq f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), -\eta \nabla f(\mathbf{x}) \rangle + \frac{M}{2} ||\eta \nabla f(\mathbf{x})||^2 \tag{3}$$

$$\leq f(\mathbf{x}) - \eta \left(1 - \frac{\eta M}{2}\right) ||\nabla f(\mathbf{x})||^2 \tag{4}$$

Thus, one can see that, for any $\eta \leq \frac{1}{M}$,

$$f(x - \eta \nabla f(\mathbf{x})) \leq f(\mathbf{x}) - \frac{\eta}{2} ||\nabla f(\mathbf{x})||^2 \leq f(\mathbf{x}) - \eta \alpha ||\nabla f(\mathbf{x})||^2 \tag{5}$$

for any $\alpha < 0.5$. □

This result answers both points we needed to ensure the viability of running BTLS. All values of $\eta \in \left(0, \frac{1}{M}\right]$ necessarily satisfy (1), which addresses the first point. Furthermore, it is easy to see that the step size in line 9 of Algorithm 1 is lower bounded by $\frac{\beta}{M}$. Hence, the step size used in any iteration can be bounded away from zero as $\eta \geq \min(1, \frac{\beta}{M})$.

Theorem 1. *If f is m-strongly convex and M-smooth, then successive iterations of gradient descent with BTLS satisfy*

$$f(\mathbf{x}_{k+1}) - f^* \leq c(f(\mathbf{x}_k) - f^*)$$

with $c = \left(1 - \min\left\{2m\alpha, \frac{2\beta\alpha m}{M}\right\}\right) \in (0, 1)$,

Proof. From the condition on the while loop in Algorithm 1,

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \alpha\eta ||\nabla f(\mathbf{x}_k)||^2 \tag{6}$$

Since we know η is lower bounded by $\min(1, \frac{\beta}{M})$,

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \alpha \min(1, \frac{\beta}{M}) ||\nabla f(\mathbf{x}_k)||^2 \tag{7}$$

$$f(\mathbf{x}_{k+1}) - f^* \leq f(\mathbf{x}_k) - f^* - \alpha \min(1, \frac{\beta}{M}) ||\nabla f(\mathbf{x}_k)||^2 \tag{8}$$

Recall that from strong convexity

$$||\nabla f(\mathbf{x})||^2 \geq 2m(f(\mathbf{x}) - f^*) \tag{9}$$

Substituting (9) in (8), the claim follows. □

The above theorem guarantees a linear convergence with rate c.

2.2 Comparison of BTLS and Exact Line Search

We numerically compare the performance of BTLS and Exact Line search. The objective function to minimize is chosen to be

$$f(\mathbf{x}) = e^{(a^T x - 0.5)} + e^{(b^T x - 0.1)} + e^{(c^T x - 0.1)} \tag{10}$$

where $a^T = [1\ 2], b^T = [1\ -3]$ and $c^T = [-1\ 0]$. The optimal f^* (see Fig. 1) for this function and the optimal point x^* are computed separately using CVX. The initial starting point for both the algorithms was set to $x_0^T = [2\ 1]$.

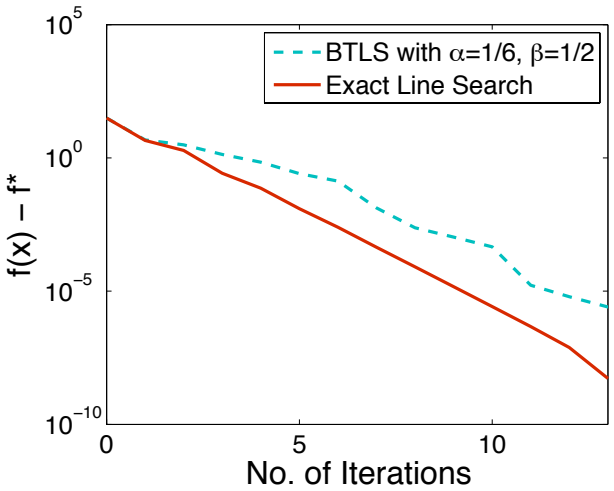

Figure 1: The plot comparing the error with iteration for BTLS and Exact Line Search

3 Convergence of GD for smooth and convex functions

In this section, we study the convergence rate of GD for solving the optimization problem

$$\min_{\mathbf{x}} \ f(\mathbf{x})$$

when the objective function f is smooth and convex (may not be strongly convex). In this case, the error of gradient descent approaches zero at a sublinear rate of $\mathcal{O}(1/k)$.

Theorem 2. *If f is convex and M-smooth, then successive iterations of gradient descent with stepsize $\eta \in (0, \frac{2}{M})$ satisfy*

$$f(\mathbf{x}_k) - f^* \leq \frac{2(f(\mathbf{x}_0) - f^*)\|\mathbf{x}_0 - \mathbf{x}^*\|^2}{2\|\mathbf{x}_0 - \mathbf{x}^*\|^2 + k\eta(2 - \eta M)(f(\mathbf{x}_0) - f^*))}$$

for $k \geq 0$. Note that here $\mathbf{x}^*$ is an arbitrary optimal solution.

Proof. First we show that the distance to an optimal solution $\mathbf{x}^*$ is non-increasing. To prove this part we will use the following inequality which is a side result of M-smoothness:

$$(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^\top (\mathbf{x} - \mathbf{y}) \geq \frac{1}{M}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2. \tag{11}$$

Part I: First, note that

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 = \|\mathbf{x}_k - \mathbf{x}^* - \eta \nabla f(\mathbf{x}_k)\|^2 \tag{12}$$
$$= \|\mathbf{x}_k - \mathbf{x}^*\|^2 - 2\eta \nabla f(\mathbf{x}_k)^\top (\mathbf{x}_k - \mathbf{x}^*) + \eta^2 \|\nabla f(\mathbf{x}_k)\|^2 \tag{13}$$
$$= \|\mathbf{x}_k - \mathbf{x}^*\|^2 - 2\eta (\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}^*))^\top (\mathbf{x}_k - \mathbf{x}^*) + \eta^2 \|\nabla f(\mathbf{x}_k)\|^2 \tag{14}$$
$$\leq \|\mathbf{x}_k - \mathbf{x}^*\|^2 - 2\eta (\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}^*))^\top (\mathbf{x}_k - \mathbf{x}^*) + \eta^2 \|\nabla f(\mathbf{x}_k)\|^2 \tag{15}$$

Now using the inequality in (11), we can show that

$$(\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}^*))^\top (\mathbf{x}_k - \mathbf{x}^*) \geq \frac{1}{M} \|\nabla f(\mathbf{x}_k)\|^2.$$

Applying this inequality into the inequality in (15) leads to

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 \leq \|\mathbf{x}_k - \mathbf{x}^*\|^2 - \frac{2\eta}{M} \|\nabla f(\mathbf{x}_k)\|^2 + \eta^2 \|\nabla f(\mathbf{x}_k)\|^2 \tag{16}$$
$$\leq \|\mathbf{x}_k - \mathbf{x}^*\|^2 - \eta \left(\frac{2}{M} - \eta \right) \|\nabla f(\mathbf{x}_k)\|^2 \tag{17}$$
$$\leq \|\mathbf{x}_k - \mathbf{x}^*\|^2 \tag{18}$$

Using this argument we can show that $\|\mathbf{x}_k - \mathbf{x}^*\|^2 \leq \|\mathbf{x}_0 - \mathbf{x}^*\|^2$ for any $k \geq 0$.

Part II: Now we proceed to the second part of the proof that show decrease in objective function value. Using M-smoothness we can write

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^\top (\mathbf{x}_{k+1} - \mathbf{x}_k) + \frac{M}{2} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \tag{19}$$
$$= f(\mathbf{x}_k) - \eta \|\nabla f(\mathbf{x}_k)\|^2 + \frac{M\eta^2}{2} \|\nabla f(\mathbf{x}_k)\|^2 \tag{20}$$
$$= f(\mathbf{x}_k) - \eta \left(1 - \frac{M\eta}{2} \right) \|\nabla f(\mathbf{x}_k)\|^2 \tag{21}$$

Hence, we have

$$f(\mathbf{x}_{k+1}) - f^* \leq f(\mathbf{x}_k) - f^* - \eta \left(1 - \frac{M\eta}{2} \right) \|\nabla f(\mathbf{x}_k)\|^2 \tag{22}$$

Now note that due to convexity of f we know that

$$f^* \geq f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^\top (\mathbf{x}^* - \mathbf{x}_k) \tag{23}$$

which implies that

$$f(\mathbf{x}_k) - f^* \leq \nabla f(\mathbf{x}_k)^\top (\mathbf{x}_k - \mathbf{x}^*) \tag{24}$$

Further, we can derive an upper bound for $\nabla f(\mathbf{x}_k)^\top (\mathbf{x}_k - \mathbf{x}^*)$ and show that

$$f(\mathbf{x}_k) - f^* \leq \|\nabla f(\mathbf{x}_k)\| \|\mathbf{x}_k - \mathbf{x}^*\| \leq \|\nabla f(\mathbf{x}_k)\| \|\mathbf{x}_0 - \mathbf{x}^*\| \tag{25}$$

Use this inequality to simplify (22) as

$$f(\mathbf{x}_{k+1}) - f^* \leq f(\mathbf{x}_k) - f^* - \eta \left(1 - \frac{M\eta}{2} \right) \left(\frac{(f(\mathbf{x}_k) - f^*)^2}{\|\mathbf{x}_0 - \mathbf{x}^*\|^2} \right) \tag{26}$$

Part III: Now we show that if a sequence of positive values a_k satisfy

$$a_{k+1} \leq a_k - ca_k^2$$

where $c_k > 0$ then we have

$$\frac{1}{a_{k+1}} \geq \frac{1}{a_k} + c$$

Note that if we multiply both sides by $1 + ca_k$ then we have

$$\begin{aligned}(1 + ca_k)a_{k+1} &\leq (1 + ca_k)(1 - ca_k)a_k &(27)\\ &\leq (1 - c^2a_k^2)a_k &(28)\\ &\leq a_k &(29)\end{aligned}$$

Therefore,

$$\frac{1}{a_{k+1}} \geq \frac{(1 + ca_k)}{a_k} \geq \frac{1}{a_k} + c \tag{30}$$

Now by using the result in part III, the inequality in (26) implies that

$$\frac{1}{f(\mathbf{x}_{k+1}) - f^*} \geq \frac{1}{f(\mathbf{x}_k) - f^*} + \frac{\eta\left(1 - \frac{M\eta}{2}\right)}{\|\mathbf{x}_0 - \mathbf{x}^*\|^2}$$

Now by summing this expression from 0 to $k - 1$ we obtain that

$$\frac{1}{f(\mathbf{x}_k) - f^*} \geq \frac{1}{f(\mathbf{x}_0) - f^*} + \frac{k\eta\left(1 - \frac{M\eta}{2}\right)}{\|\mathbf{x}_0 - \mathbf{x}^*\|^2}$$

which implies that

$$f(\mathbf{x}_k) - f^* \leq \frac{2(f(\mathbf{x}_0) - f^*)\|\mathbf{x}_0 - \mathbf{x}^*\|^2}{2\|\mathbf{x}_0 - \mathbf{x}^*\|^2 + k\eta(2 - \eta M)(f(\mathbf{x}_0) - f^*))}$$

□

Corollary 1. *If we choose the stepsize to be $\eta = 1/M$ then we have*

$$f(\mathbf{x}_k) - f^* \leq \frac{2M(f(\mathbf{x}_0) - f^*)\|\mathbf{x}_0 - \mathbf{x}^*\|^2}{2M\|\mathbf{x}_0 - \mathbf{x}^*\|^2 + k(f(\mathbf{x}_0) - f^*))}$$

By using the following inequality

$$f(\mathbf{x}_0) \leq f^* + \nabla f(\mathbf{x}^*)^\top(\mathbf{x}_0 - \mathbf{x}^*) + \frac{M}{2}\|\mathbf{x}_0 - \mathbf{x}^*\|^2 = f^* + \frac{M}{2}\|\mathbf{x}_0 - \mathbf{x}^*\|^2$$

we obtain the following result.

Corollary 2. *If we choose the stepsize to be $\eta = 1/M$ then we have*

$$f(\mathbf{x}_k) - f^* \leq \frac{2M\|\mathbf{x}_0 - \mathbf{x}^*\|^2}{k + 4}$$