EE C227C Convex Optimization and Approximation Lecture 3 - 01/25/22

Lecture 3: Gradient Descent

Lecturer: Jiantao Jiao *Scribe: Aekus Bhathal, Jensen Gao, Seunghoon Paik, Tianjun Zhang*

In this lecture, we prove bounds on the convergence rates of gradient descent for convex and strongly convex functions with Lipschitz continuous gradients.

1 Recap

We motivate gradient descent by first trying to optimize the family of convex functions with Lipschitz continuous gradients. These properties give rise to three useful inequalities.

$\forall x, y \in \mathcal{D}(f)$:

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

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

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2 \tag{3}$$

We now define gradient descent (GD)

Algorithm 1 Gradient Descent

```
choose $x_0$
$k \leftarrow 1$
while end condition not met do
    $x_k \leftarrow x_{k-1} - h_k \nabla f(x_{k-1})$
    $k \leftarrow k + 1$
end while
return $x_k$
```

2 Analysis

We focus on the case of constant learning rates, although adaptive learning rates have widespread analysis and utility.

Theorem 1. *For any convex function f with Lipschitz constant L of the gradient of f, choose a learning rate $0 \leq h \leq \frac{2}{L}$. Then, the iterates of gradient descent $\{x_k\}$ satisfy*

$$f(x_k) - f(x^*) \leq \frac{2\big(f(x_0) - f(x^*)\big)\|x_0 - x^*\|^2}{2\|x_0 - x^*\|^2 + kh\big(2 - Lh\big)\big(f(x_0) - f(x^*)\big)} \tag{4}$$

where x^* is an optimal point for f such that $\forall x \in \mathcal{D}(f)$, $f(x^*) \leq f(x)$.

Before beginning the proof, we make a few observations about this bound. First, the suboptimality gap at any timestep k is dependent on the initial suboptimality gap and distance from the initial iterate and one of the optimizers. If we consider the initial iterate constant, then the suboptimality gap at timestep k

is purely determined by the learning rate. We see that the tightest bound is provided by $h = \frac{1}{L}$, simplifying the bound to

$$f(x_k) - f(x^*) \leq \frac{2L\big(f(x_0) - f(x^*)\big)\|x_0 - x^*\|^2}{2L\|x_0 - x^*\|^2 + k\big(f(x_0) - f(x^*)\big)}. \tag{5}$$

Proof Let $r_k = \|x_k - x^*\|$. We can write r_{k+1} recursively in terms of r_k.

$$\begin{aligned} r_{k+1}^2 &= \|x_{k+1} - x^*\|^2 \\ &= \|x_k - h\nabla f(x_k) - x^*\|^2 \\ &= \|x_k - x^*\|^2 - 2h\langle x_k - x^*, \nabla f(x_k)\rangle + h^2\|\nabla f(x_k)\|^2 \\ &\leq r_k^2 - 2h\frac{1}{L}\|\nabla f(x_k)\|^2 + h^2\|\nabla f(x_k)\|^2 \\ &= r_k^2 - h(\frac{2}{L} - h)\|\nabla f(x_k)\|^2 \end{aligned}$$

where the inequality is because f has Lipschitz gradient with constant L. Because $h(\frac{2}{L} - h)$ is a non-negative constant when $0 \leq h \leq \frac{2}{L}$, this shows that r_k is a monotonically decreasing sequence.
Also, by the convexity and L-Lipschitz gradient property of f,

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k\rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\ &= f(x_k) + \langle \nabla f(x_k), -h\nabla f(x_k)\rangle + \frac{L}{2}\| - h\nabla f(x_k)\|^2 \\ &= f(x_k) - \omega\|\nabla f(x_k)\|^2 \end{aligned}$$

where $\omega = h(1 - \frac{Lh}{2})$, a positive constant if $0 \leq h \leq \frac{2}{L}$. Therefore, gradient descent is a descent method if f possesses a Lipschitz gradient.
Now, we can analyze $\Delta_k = f(x_k) - f^*$.

$$\begin{aligned} \Delta_k &= f(x_k) - f^* \\ &\leq \langle \nabla f(x_k), x_k - x^*\rangle \\ &\leq \|\nabla f(x_k)\|\|x_k - x^*\| \\ &\leq r_0\|\nabla f(x_k)\| \\ &\Rightarrow \|\nabla f(x_k)\| \geq \frac{\Delta_k}{r_0} \end{aligned}$$

where the first inequality follows from the convexity of f, the second inequality follows from Cauchy-Schwartz, and the third inequality follows from the monotonicity of r_k. Then,

$$\begin{aligned} \Delta_{k+1} &= f(x_{k+1}) - f^* \\ &\leq f(x_k) - f^* - \omega\|\nabla f(x_k)\|^2 \\ &\leq \Delta_k - \frac{\omega}{r_0^2}\Delta_k^2. \end{aligned}$$

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

$$\begin{aligned} \frac{1}{\Delta_{k+1}} &\geq \frac{1}{\Delta_k} + \frac{\omega}{r_0^2}\frac{\Delta_k}{\Delta_{k+1}} \\ &\geq \frac{1}{\Delta_k} + \frac{\omega}{r_0^2} * 1 \\ &\Rightarrow \frac{1}{\Delta_{k+1}} \geq \frac{1}{\Delta_0} + \frac{\omega}{r_0^2}(k+1) \end{aligned}$$

which implies the statement of the theorem. □

Corollary 2. *Choose $h = 1/L$, then we have*

$$f(x_k) - f(x^*) \le \frac{2L\|x_0 - x^*\|^2}{k+4}. \tag{6}$$

Proof

$$f(x_k) - f(x^*) \le \frac{2L(f(x_0) - f(x^*))\|x_0 - x^*\|^2}{2L\|x_0 - x^*\|^2 + k(f(x_0) - f(x^*))} \le \frac{2L\|x_0 - x^*\|^2}{k+4}.$$

The last inequality holds since $f(x_0) - f(x^*) \le \langle \nabla f(x^*), x_0 - x^* \rangle + \frac{L}{2}\|x_0 - x^*\|^2 = \frac{L}{2}\|x_0 - x^*\|^2$. □

Suppose the objective function is convex with Lipschitz gradient. Naturally, one may be interested in knowing if there exists another method that speeds up convergence. Surprisingly, there is such a method made by Nesterov (1983). Before Nesterov's result, it was proven that an accelerated $1/\epsilon^2$ convergence speed is possible for quadratic functions, but difficult to generalize for all convex functions. Nesterov's method does not accelerate the quadratic function case that fast; however, it accelerates all convex functions.

Definition 3 (Dimension free)**.** *The method we have observed is **dimension free**. It means that the dimension does not explicitly appear in the convergence speed. That is, the number of the iterations to achieve the desired error tolerance is independent of the dimension. However in the computation of the gradient, dimension plays an important role.*

3 Strong Convexity

Definition 4 (Strongly convex functions)**.** *A differentiable function f is strongly convex if for some $\mu > 0$, $f(y) \ge f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2$. That is, the function has a local quadratic lower bound.*

Observation 5. *Suppose f is strongly convex. Then, $\langle \nabla f(x) - \nabla f(y), x - y \rangle \ge \mu\|x - y\|^2$.*

Proof We have two inequalities: (i) $f(y) \ge f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}|y - x\|^2$; (ii)$f(x) \ge f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2}\|y - x\|^2$. Adding these two inequalities gives $\langle \nabla f(x) - \nabla f(y), x - y \rangle \ge \mu\|x - y\|^2$. □

Lemma 6 (Weighted lower bound)**.** *Suppose f is strongly convex with Lipschitz gradient. Then,*

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \ge \frac{\mu L}{\mu + L}\|x - y\|^2 + \frac{1}{\mu + L}\|\nabla f(x) - \nabla f(y)\|^2. \tag{7}$$

Proposition 7 (The method for certifying strong convexity)**.** *Suppose a function f is twice differentiable. Then,*

- *f has a Lipschitz gradient $\Leftrightarrow \nabla^2 f(x) \preceq LI_d$*
- *f is strongly convex $\Leftrightarrow \nabla^2 f(x) \succeq \mu I_d$*

Theorem 8. *Suppose $0 < h \le \frac{2}{\mu+L}$ and a function f is strongly convex with Lipschitz gradient. Then the GD method satisfies*

$$\|x_k - x^*\| \le \left(1 - \frac{2h\mu L}{\mu + L}\right)^k \|x_0 - x^*\|^2. \tag{8}$$

Choose $h = \frac{2}{\mu+L}$. We have

$$\|x_k - x^*\| \leq \left(\frac{Q-1}{Q+1}\right)^k \|x_0 - x^*\| \tag{9}$$

$$f(x_k) - f^* \leq \frac{L}{2}\left(\frac{Q-1}{Q+1}\right)^{2k} \|x_0 - x^*\|^2 \tag{10}$$

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

We see that strongly convex functions converge geometrically with gradient descent in contrast with convex functions that converge with rate $\propto \frac{1}{k}$. Another observation is that shrinking Q shrinks the geometric ratio and thus faster convergence. This intuitively makes sense as smaller Q with Lipschitz constant L, implies μ and the gradients are larger.

Proof Using the same definition $r_k = \|x_k - x^*\|$, we have

$$r_{k+1}^2 = r_k^2 + h^2\|\nabla f(x)\|^2 - 2h\langle \nabla f(x), x_k - x^*\rangle.$$

Apply the bound in (6),

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

Hence,

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

Selecting $0 < h \leq \frac{2}{\mu+L}$, means the second term on the right side is non-positive. Hence, the bound becomes

$$r_{k+1}^2 \leq (1 - \frac{2h\mu L}{\mu + L})r_k^2.$$

Here, the optimal h is $\frac{2}{\mu+L}$, giving us the bound

$$r_{k+1} \leq \Big(\frac{Q-1}{Q+1}\Big)r_k.$$

which naturally leads to the general formula in theorem 8. The bound on the suboptimality gap follows since

$$f(x_k) - f^* \leq \langle \nabla f(x^*), x_k - x^*\rangle + \frac{L}{2}\|x_k - x^*\|^2.$$

The first term is zero, so

$$\begin{aligned} f(x_k) - f^* &\leq \frac{L}{2}\|x_k - x^*\|^2 \\ &\leq \frac{L}{2}\left(\frac{Q-1}{Q+1}\right)^{2k} \|x_0 - x^*\|^2. \end{aligned}$$

□

Remark Nesterov's acceleration method applied to strongly convex function replaces Q with $\sqrt{Q}$ in theorem 8, which is provably the optimal rate.