Columbia University in the City of New York
Optimization Methods for Machine Learning
Instructors: Satyen Kale
Authors: Chao Qin
Email: cq2199@columbia.edu

SCRIBE

6

Gradient Descent

1. Convexity

This section covers a short recap of the convex optimization problem and introduces some relevant results.

Definition 6.1 (Convex optimization problem) A convex optimization problem is an optimization problem of the form

$$\min f(x)$$
$$\text{s.t. } x \in K$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is convex and K is a convex set of $\mathbb{R}^d$.

Theorem 6.1 Let $K = \mathbb{R}^d$ and x^* be a minimizer of f on $\mathbb{R}^d$. If f is differentiable at x^*, then $\nabla f(x^*) = 0$.

The following is the proof sketch of this result.

Proof. For and x and δ,

$$f(x+\delta) \approx f(x) + \nabla f(x)^\top \delta.$$

Since x^* is a minimzer, we have $\nabla f(x^*)^\top \delta \geq 0$ for small $\delta \in \mathbb{R}^d$. If we choose $\delta = -\epsilon \nabla f(x^*)^\top$, for some small $\epsilon > 0$ so that the approximation above is valid, we have $-\epsilon \nabla f(x^*)^\top \nabla f(x^*) \geq 0$, which leads to $\nabla f(x^*) = 0$. □

In the above result, $K = \mathbb{R}^d$. In fact, we have the following result for general convex set K.

Theorem 6.2 Let f be differentiable at $x^* \in K$. If x^* is a minimizer of f on K, then

$$\nabla f(x^*)^\top (y - x^*) \geq 0, \quad \forall y \in K. \tag{1}$$

Theorem 6.1 is a special case of Theorem 6.2. If $K = \mathbb{R}^d$, we can take $y = x^* - \nabla f(x^*)$, and then Equation (1) implies $\nabla f(x^*) = 0$, which is Theorem 6.1. Theorem 6.2 can be proved similarly to Theorem 6.1. The following result shows that the inverse of Theorem 6.2 is also true:

Theorem 6.3 Let f be differentiable at $x^* \in K$. If Equation (1) holds, then x^* is a minimizer.

2. Gradient Descent

In this section, we first introduce the oracle complexity and gradient descent algorithm, and then we analyze the complexity of this algorithm.

The following oracles are widely seen in the literature.

- Zero-order oracle: given x, returns $f(x)$
- First-order oracle: given x, returns $\nabla f(x)$
- Second-order oracle: given x, returns $\nabla^2 f(x)$

Based on the definition of first-order oracle, we can define the first-order complexity of an optimization algorithm.

Definition 6.2 (First-Order Complexity) First-order oracle of an optimization algorithm is the number of calls it needs to make a first-order oracle to compute x s.t.

$$f(x) \leq \min_{x' \in K} f(x') + \epsilon$$

for a given $\epsilon > 0$.

Typical result could be $O(1/\epsilon^2), O(1/\epsilon), O(\log\left(1/\epsilon\right))$.

Now we introduce the gradient descent algorithm. For simplicity, we assume $K = \mathbb{R}^d$.

```
Param: $\eta > 0$, which is the stepsize;
Init: $x_0 \in K$ arbitrary;
for $t = 0, 1, 2, \ldots$ do
    $x_{t+1} = x_t - \eta \nabla f(x_t)$
end
return $x_T$ (or some combination of $x_0, \ldots, x_T$)
```

Algorithm 1: Gradient Descent

It is easy to see that the first order complexity of gradient descent is T. Next we provide the analysis of this algorithm. We consider the potential function $\|x_t - x^*\|^2$. We have

$$\|x_{t+1} - x^*\|^2 = \|x_t - \eta \nabla f(x_t) - x^*\|^2 = \|x_t - x^*\|^2 + \eta^2 \|\nabla f(x_t)\|^2 - 2\eta \nabla f(x_t)^\top (x_t - x^*),$$

which leads to

$$\nabla f(x_t)^\top (x_t - x^*) = \frac{1}{2\eta}\left(\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2\right) + \frac{\eta}{2}\|\nabla f(x_t)\|^2.$$

By convexity, we have

$$f(x_t) - f(x^*) \leq \frac{1}{2\eta}\left(\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2\right) + \frac{\eta}{2}\|\nabla f(x_t)\|^2.$$

Taking the summation of t from 0 to $T-1$, we have

$$\sum_{t=0}^{T-1} f(x_t) - f(x^*) \leq \frac{1}{2\eta}\left(\|x_0 - x^*\|^2 - \|x_T - x^*\|^2\right) + \frac{\eta}{2}\sum_{t=0}^{T-1}\|\nabla f(x_t)\|^2.$$

At this point, to get a meaningful bound, we need to control $\sum_{t=0}^{T-1}\|\nabla f(x_t)\|^2$ somehow. One way to do this is to assume that f is L-Lipschitz for some $L \geq 0$, i.e. for all $x, y \in \mathbb{R}^d$, we have $|f(x) - f(y)| \leq L\|x - y\|$. This turns out to be equivalent to the following:

$$\|\nabla f(x)\| \leq L, \quad \forall x.$$

Under this assumption,

$$\text{RHS} \leq \frac{1}{2\eta}\|x_0 - x^*\|^2 + \frac{\eta}{2}L^2 T.$$

By Jensen's inequality,

$$\text{LHS} = \sum_{t=0}^{T-1} f(x_t) - f(x^*) \geq T\left(f(\bar{x}) - f(x^*)\right)$$

where $\bar{x} = \frac{1}{T}\sum_{t=0}^{T-1} x_t$. Combining the above two inequalities, we have

$$f(\bar{x}) \leq f(x^*) + \frac{1}{2\eta T}\|x_0 - x^*\|^2 + \frac{\eta}{2}L^2.$$

We pick

$$\eta = \frac{\|x_0 - x^*\|}{L\sqrt{T}},$$

and then

$$f(\bar{x}) \leq f(x^*) + \frac{L\|x_0 - x^*\|}{\sqrt{T}}.$$

Thus to achieve a given sub-optimality gap $\epsilon > 0$, we need

$$\frac{L\|x_0 - x^*\|}{\sqrt{T}} \leq \epsilon \implies T \geq \frac{L^2\|x_0 - x^*\|^2}{\epsilon^2}.$$

In fact, x^* is unknown, and thus the η that we picked is unknown. Usually, people make the assumption that $\|x^*\| \leq D$. Then we have

$$\|x_0 - x^*\| \leq \|x_0\| + D$$

by triangle inequality. Now we can pick

$$\eta = \frac{\|x_0\| + D}{L\sqrt{T}},$$

and correspondingly,

$$T \geq \frac{L^2\left(\|x_0\| + D\right)^2}{\epsilon^2}.$$

A good choice for x_0 is $x_0 = 0$, in which case the above bound becomes

$$T \geq \frac{L^2 D^2}{\epsilon^2}.$$