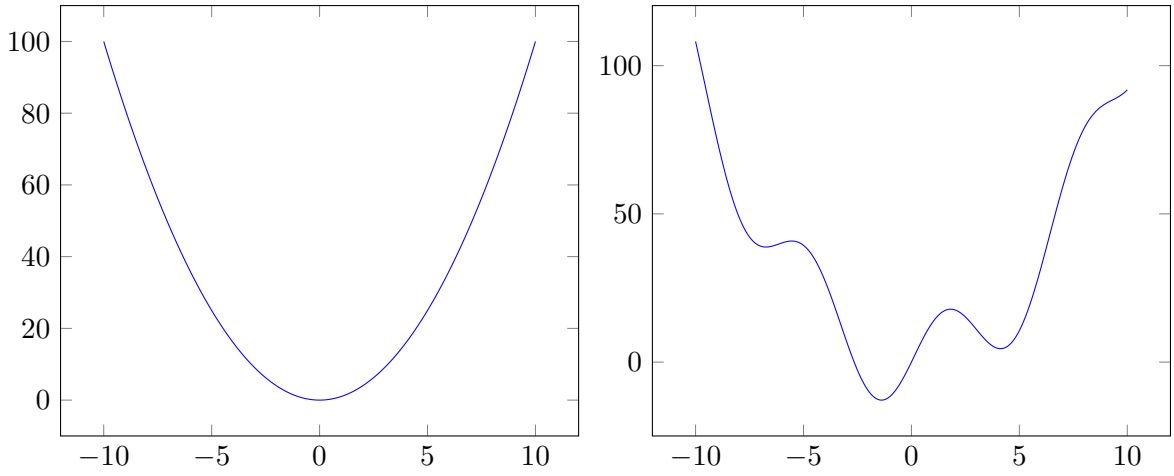
| **CS787: Advanced Algorithms** | **Scribe:** Jialu Bao **Lecture 17:** Continuous Optimization: |
|---|---|
| Gradient Descent | **Date:** Oct. 31st, 2019 |

## 17.1 Introduction

**Definition 17.1.1 (Convex function)** *A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is **convex** if for any $x, y \in \mathbb{R}^n$, any $t \in [0, 1]$,*

$$f(tx + (1-t)y) \leq t \cdot f(x) + (1-t) \cdot f(y)$$

Consider the following problem: given function $f : \mathbb{R}^n \to \mathbb{R}$, minimize $f(x)$. If $f$ is convex, then any local minimal is the global minimal, as shown in the left; if $f$ is not convex, then as shown in right diagram, local minimal is not necessarily global minimal.




For convex functions, a widely-used method to minimize their values is gradient descent. The algorithm is simple:

**Algorithm 1** Gradient Descent

**Given:** Convex function $f : \mathbb{R}^n \to \mathbb{R}$, a randomly chosen $x_0$ in the domain of $f$, some positive step size $\eta$

1: **for** $t = 1$ to $T$ **do**
2: $\quad x_t \leftarrow x_{t-1} - \eta \nabla f(x_{t-1})$

In the following, we will explain some preliminaries from calculus and bound how fast it converges.

## 17.2 Preliminaries

First, recall theorems from calculus.

**Theorem 17.2.1 (Taylor's Theorem with remainder)** *If $f : \mathbb{R}^n \to \mathbb{R}$ is at least twice differentiable at $x$, then*

$$f(y) = f(x) + \nabla f(x)^T (y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(x)(y - x) + o((y - x)^T (y - x))$$

*where $\nabla f(x)$ is the gradient of $f(x)$, and $\nabla^2 f(x)$ is the Hessian matrix, i.e.,*

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \vdots \\ \frac{\partial f}{\partial x_N}(x) \end{bmatrix} \qquad \nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(x) & \dots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(x) \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(x) & \dots & \frac{\partial^2 f}{\partial x_n \partial x_n}(x) \end{bmatrix}$$

Denote $f(y) - f(x) - \nabla f(x)(y-x) - \frac{1}{2}\nabla^2 f(x)(y-x)^2$ as $R$. By the theorem, $R$ is in $o\left((y - x)^T (y - x)\right)$, i.e.

$$\lim_{y \to x} \frac{R}{(y - x)^2} = 0$$

This implies that $R$ is also $O((y - x)^T (y - x))$.

**Theorem 17.2.2 (Lagrange remainder)** *For $f : \mathbb{R}^n \to \mathbb{R}$ that is at least twice differentiable at $x$, then there exists $x' \in [x, y]$ such that*

$$f(y) = f(x) + \nabla f(x)^T (y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(x')(y - x)$$

**Proof:** For simplicity, we only prove it for univariate function $f$. Let $F(\alpha) := f(\alpha) + f'(\alpha)(y - \alpha)$. Then $F(y) = f(y)$, $F(x) = f(x) + f'(x)(y - x)$.

Recall the generalized mean value theorem that, for any $F, G$ continuous on $[a, b]$, differentiable on $(a, b)$, and $G'(x) \neq 0$ for any $x \in (a, b)$, there exists $c \in (a, b)$ such that

$$\frac{F(b) - F(a)}{G(b) - G(a)} = \frac{F'(c)}{G'(c)}$$

Substitute in $F$ defined above and $G(\alpha) := (y - \alpha)^2$ on interval between $x, y$,

$$\begin{aligned} F(y) - F(x) &= \frac{F'(c)}{G'(c)} \cdot (G(y) - G(x)) \\ &= \frac{f''(c)(y - c)}{-2(y - c)} \cdot (0 - (y - x)^2) \\ &= \frac{1}{2} f''(c)(x - y)^2 \end{aligned}$$

Thus,

$$\begin{aligned} f(y) = F(y) &= F(x) + \frac{1}{2} f''(c)(x - y)^2 \\ &= f(x) + f'(x)(y - x) + \frac{1}{2} f''(c)(x - y)^2 \end{aligned}$$

■

When $f$ is convex, Hessian matrix $\nabla^2 f(x)$ is positive-semi-definite, so $(y-x)^T \nabla^2 f(x)(y-x) \geq 0$, which together with Lagrange remainder theorem implies that

$$f(y) \geq f(x) + \nabla f(x)^T (y-x)$$

So if $x_t = x_{t-1} - \eta \nabla f(x_{t-1})$,

$$\begin{aligned} & f(x_{t-1}) \geq f(x_t) + \nabla f(x_t)^T \cdot \eta \cdot \nabla f(x_{t-1}) \\ \implies & f(x_t) \leq f(x_{t-1}) - \nabla f(x_t)^T \cdot \eta \cdot \nabla f(x_{t-1}) \end{aligned}$$

For convex function $f$ and $x_t, x_{t-1}$ that are not minimal, $f(x_t), f(x_{t-1}$ are positive, so $f(x_t)$ in gradient descent will decrease strictly monotonically until it reaches the global minimal. However, this doesn't give us any guarantee of how fast the gradient descent converges.

## 17.3 Convergence of gradient descent

Notation: $||\cdot||$ are all L-2 norms .

### 17.3.1 Assumptions

Besides $f$ is convex, we in addition assume that $f$ is

1. $\beta$-smoothness, i.e., the eigenvalues of $\nabla^2 f(x)$ is at most $\beta$. Equivalently, for any $x, y$ in the domain,
$$||\nabla f(y) - \nabla f(x)|| \leq \beta ||y - x||$$
2. $\alpha$-strongly convex, i.e., the eigenvalues of $\nabla^2 f(x)$ is at least $\alpha$. Equivalently, for any $x, y$ in the domain,
$$f(y) \geq f(x) + \nabla f(x)^T (y-x) + \frac{\alpha}{2} \cdot ||y-x||^2$$

### 17.3.2 Bounding the convergence

**Theorem 17.3.1** *If $\eta \leq \frac{1}{\beta}$ we have that*

$$||x_t - x^*||^2 \leq (1 - \frac{\eta \cdot \alpha}{2})^t \cdot ||x_0 - x^*||^2,$$

*i.e., $x_t$ converges to the minimum $x^*$ quickly. In addition*

$$f(x_t) - f(x^*) \leq \beta (1 - \frac{\eta \cdot \alpha}{2})^t \cdot ||x_0 - x^*||^2$$

*i.e., the distance between $f(x^t)$ and $f(x^*)$ drops exponentially.*

This shows that the gap between $f(x^*)$ and $f(x_t)$ in gradient descent decreases exponentially, and thus indicating gradient descent converges very fast. To prove the theorem, we first use the following lemma

**Lemma 17.3.2** *If $f$ is $\beta$-smooth and $\alpha$-strongly convex, then*

$$\nabla f(x_t)^T(x_t - x^*) \geq \frac{\alpha}{4}||x_t - x^*||^2 + \frac{1}{2\beta}||\nabla f(x_t)||^2$$

**Proof:** We first prove the theorem 17.3.2 assuming the lemma 17.3.2. First, since $x_t := x_{t-1} + \eta\nabla f(x_{t-1})$,

$$\begin{aligned}||x_t - x^*||^2 &= ||x_{t-1} - x^* - \eta\nabla f(x_{t-1})||^2 \\ &= ||x_{t-1} - x^*||^2 + \eta^2 \cdot ||\nabla f(x_{t-1})||^2 - 2\eta\nabla f(x_{t-1})^T(x_{t-1} - x^*)\end{aligned}$$

By the lemma above, this is at most

$$\begin{aligned}&||x_{t-1} - x^*||^2 + \eta^2 \cdot ||\nabla f(x_{t-1})||^2 - 2\eta\left(\frac{\alpha}{4}||x_t - x^*||^2 + \frac{1}{2\beta}||\nabla f(x_t)||^2\right) \\ =&(1 - \frac{\alpha\cdot\eta}{2})||x_{t-1} - x^*||^2 + (\eta^2 - 2\eta\cdot\frac{1}{2\beta})||\nabla f(x_t)||^2\end{aligned}$$

Recall that we assumed $\eta \leq \frac{1}{\beta}$, so $0 \leq \eta^2 \leq \eta\cdot\frac{1}{\beta}$, and $\eta^2 - 2\eta\cdot\frac{1}{2\beta} \leq 0$. So it follows that

$$\begin{aligned}||x_t - x^*||^2 &\leq (1 - \frac{\alpha\cdot\eta}{2})||x_{t-1} - x^*||^2 + (\eta^2 - 2\eta\cdot\frac{1}{2\beta})||\nabla f(x_t)||^2 \\ &\leq (1 - \frac{\alpha\cdot\eta}{2})||x_{t-1} - x^*||^2 \qquad (17.3.1)\end{aligned}$$

For the convergence of function output, theorem 17.2.2 implies that

$$\begin{aligned}&f(x^*) \geq f(x_t) + \nabla f(x_t)^T(x_t - x^*) \\ \implies &f(x_t) - f(x^*) \leq \nabla f(x_t)^T(x_t - x)\end{aligned}$$

Since $x^*$ is a local minimal, $\nabla f(x^*)^T = \vec{0}$, so the inequality above can be rewritten as

$$f(x_t) - f(x^*) \leq \left(\nabla(f(x_t)^T - \nabla f(x^*)^T\right)(x_t - x)$$

By $\beta$-smoothness,

$$||f(x_t)^T - \nabla f(x^*)^T|| \leq \beta||x^t - x*||$$

so combined,

$$f(x_t) - f(x^*) \leq ||\left(\nabla(f(x_t)^T - \nabla f(x^*)^T\right)(x_t - x)|| \leq \beta||x_t - x*||^2 \qquad (17.3.2)$$

Thus, we have the theorem from inequalities 17.3.1 and 17.3.2. ■

Now we prove the 17.3.2. **Proof:** Note that if we can prove both of the followings, then the lemma follows from adding up two inequalities and divide by two

1. $\nabla f(x_t)(x_t - x^*) \geq \frac{\alpha}{2}||x_t - x^*||^2$

2. $\nabla f(x_t)(x_t - x^*) \geq \frac{1}{\beta}||\nabla f(x_t)||^2$

To prove (1), we use $\alpha$-strong convex,

$$f(x^*) \geq f(x_t) + \nabla f(x_t)^T(x^* - x_t) + \frac{\alpha}{2} \cdot ||x^* - x_t||^2$$
$$\implies \nabla f(x_t)^T(x_t - x^*) \geq \frac{\alpha}{2} \cdot ||x^* - x_t||^2 + f(x_t) - f(x^*)$$

Since $f$ achieves the local minimal at $x^*$, $f(x_t) - f(x^*) \geq 0$ for any $x_t$, so it follows that

$$\nabla f(x_t)^T(x_t - x^*) \geq \frac{\alpha}{2} \cdot ||x^* - x_t||$$

To prove (2), we recall that by Lagrange Remainder theorem, there exists some $x'$ between $x_t, x^*$ such that,

$$\nabla f(x_t) = \nabla f(x^*) + \nabla^2 f(x')(x_t - x^*)$$

$\nabla f(x^*) = 0$ as $x^*$ is the global minimal, and it follows that

$$\begin{aligned}
\nabla f(x_t) &= 0 + \nabla^2 f(x')(x_t - x^*) \\
\implies \nabla^2 f(x')^{-1} \cdot \nabla f(x_t) &= x_t - x^* \\
\implies \nabla f(x_t)^T \cdot \nabla^2 f(x')^{-1} \cdot \nabla f(x_t) &= \nabla f(x_t)^T \cdot (x_t - x^*) \\
\implies \nabla f(x_t)^T \cdot (x_t - x^*) &= \nabla f(x_t)^T \cdot \nabla^2 f(x')^{-1} \cdot \nabla f(x_t) \\
&\geq \text{min eigenvalue of } \nabla^2 f(x')^{-1} \cdot ||\nabla f(x_t)||^2
\end{aligned}$$

The last inequality follows from the fact that for any symmetric matrix $M$, the minimum eigenvalue

$$\lambda_{min} = \min_{x \neq 0} \frac{x^T A x}{x^T x},$$

which follows from Courant-Fischer theorem

Also, for any square non-singular matrix $M$, the inverse of any eigenvalue of $M$ is an eigenvalue of $M^{-1}$. Hessian matrix is symmetric and positive semi-definite, so all of its eigenvalues are non-negative real numbers. Thus, the minimal eigenvalue of $\nabla^2 f(x')^{-1}$ is the inverse of the max eigenvalue of the $\nabla^2 f(x')$, and

$$\text{min eigenvalue of } \nabla^2 f(x')^{-1} = \frac{1}{\text{max eigenvalue of } \nabla^2 f(x')}$$

Combined with the inequality above, we have

$$\nabla f(x_t)^T \cdot (x_t - x^*) \geq \frac{1}{\text{max eigenvalue of } \nabla^2 f(x')} \cdot ||\nabla f(x_t)||^2$$

By $\beta$-smoothness, max eigenvalue of $\nabla^2 f(x') \leq \beta$, so

$$\nabla f(x_t)^T \cdot (x_t - x^*) \geq \frac{1}{\beta} \cdot ||\nabla f(x_t)||^2$$

■

## 17.4 Applications

There are many applications of gradient descent. One example is the interior point algorithm that solves LP problems. More details can be find in this note.

## 17.5 References

1. http://mitliagkas.github.io/ift6085/ift-6085-lecture-3-notes.pdf