Introduction to Optimization

Note 6

Jun Moon

junmoon@hanyang.ac.kr

Outline

Last time

- Existence, convex sets and functions, convex optimization
- Constrained optimization and duality

This chapter

- Optimization algorithms: Gradient method for unconstrained optimization

Unconstrained Minimization Problem

Consider: $f : \mathbb{R}^n \to \mathbb{R}$

$$\min_x f(x)$$

where f is twice differentiable and convex. Assume the existence of the optimal solution. Then the necessary and sufficient condition for the optimality is

$$\nabla f(x^*) = 0$$

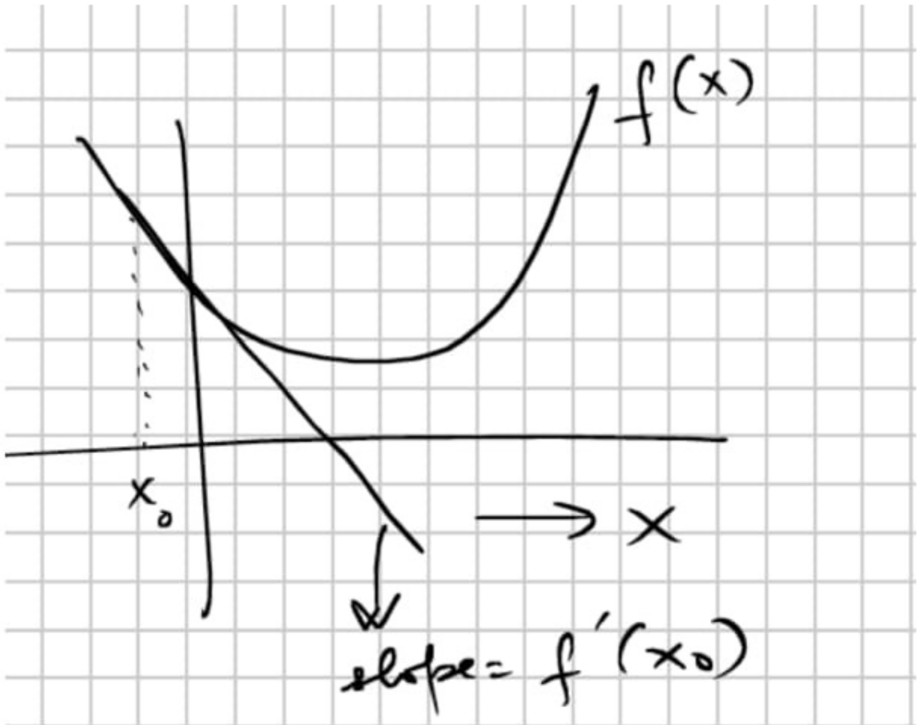
Unconstrained Minimization Problem

Optimality condition:

$$\nabla f(x^*) = 0$$

- Solving the unconstrained problem is equivalent to solving $\nabla f(x^*) = 0$ (note that $\nabla f(x)$ is a n-dimensional vector)
- Sometimes it can be solved analytically, but mostly it is not. Therefore, we will learn the algorithm how to reach to $\nabla f(x^*) = 0$
- Here, the algorithm means that for a sequence of $\{x_k\}$, we have $\lim_{k \to \infty} f(x_k) = p^*$, where $p^* = f(x^*)$
- We want $f(x_{k+1}) \leq f(x_k)$
- Note that f is continuous; therefore, we need to compute either p^* or x^*
- The algorithm terminates when $f(x_k) \leq p^* + \epsilon$ for some ϵ

Unconstrained Minimization Problem



Unconstrained Minimization Problem

Consider the scalar problem

$$\min_x f(x)$$

Assume that you have an initial guess x_0. You want to go to x^* from x_0. That is you want to design a function f such that $x_{k+1} = g(x_k)$ and $\lim_{k\to\infty} g(x_k) = x^*$. One way to do is to compute the gradient of f

Unconstrained Minimization Problem

- If $f'(x_0) < 0$, then $f(x)$ is decreasing at x_0; therefore,

$$x_1 = x_0 - t_0 f'(x_0)$$

- If $f'(x_0) > 0$, then $f(x)$ is increasing at x_0;

$$x_1 = x_0 - t_0 f'(x_0)$$

- t_0; step size (determines how much we want to move from x_0 to x_1
- Hence, the gradient descent is

$$x_{k+1} = x_k - t_k f'(x_k)$$
$$x_0 \text{ initial condition}$$

Unconstrained Minimization Problem

Vector case

$$f(x_{k+1}) \approx f(x_0) + \nabla f(x_k)^T (x_{k+1} - x_k)$$

We want $f(x_{k+1}) \leq f(x_0)$, which holds if $\nabla f(x_k)^T (x_{k+1} - x_k) \leq 0$
Hence,

$$x_{k+1} = x_k - t_k \nabla f(x_k)$$

- How to choose t_k?
- If t_k is very large, then you may not reach to optimal point
- If t_k is very small, its convergence is very slow

We will first consider the case when $t_k = t$ for all k (t_k is a constant)

Unconstrained Minimization

$$\min_{x \in \mathbb{R}^n} f(x)$$

- Assumption
  - f is convex and continuously differentiable
  - The optimal value $f^* = \inf_{x \in \mathbb{R}^n} f(x)$ is finite
- Minimization methods
  - Iterative methods for the form

$$x_{k+1} = x_k + t_k d_k, \; x_0 = x \in \mathbb{R}^n$$

    t_k: step size
    d_k: direction of the iterative algorithm
  - Generate a sequence of points $\{x_k\}$ such that $f(x_k) \to f^*$ as $k \to \infty$
  - Can be interpreted as iterative methods for solving the system of equations to satisfy the necessary and sufficient optimality condition

$$\nabla f(x^*) = 0$$

Unconstrained Minimization: Gradient Descent Method

- Gradient descent algorithm

$$x_{k+1} = x_k - t_k \nabla f(x_k), \ x_0 = x$$

- t_k: stepsize
  - Constant: $t_k = t$
  - Diminishing: $t_k \to 0$ with $\sum_{k=1}^{\infty} t_k = \infty$
  - Linear search types (optimal stepsize, hard to find) Steepest gradient descent
    - Exact line search: $t_k = \arg\min_{t>0} f(x_k + td_k)$
    - Backtracking line search

Gradient Descent with Bounded Gradients

Theorem

Suppose that the gradient is bounded, that is, for some $L > 0$

$$\|\nabla f(x)\| \leq L \quad \forall x, y \in \mathbb{R}^n.$$

Let the stepsize be constant, i.e., $\alpha_k = \alpha$. Then the gradient descent algorithm generates the sequence $\{x_k\}$ such that

$$\liminf_{k \to \infty} f(x_k) \leq f^* + \frac{\alpha L}{2}$$

Also, when diminishing step size is used, i.e., $\sum_{k=0}^{\infty} \alpha_k = \infty$, $\lim_{k \to \infty} f(x_k) = f^*$.

Gradient Descent with Bounded Gradients

Proof: Assume that the result does not hold, i.e., for some $\hat{y}$ with $f(\hat{y}) = f^* + \epsilon$,

$$f(x_k) - f(\hat{y}) \geq \frac{\alpha L}{2} + \epsilon, \ \forall k.$$

Then we can show that $\|x_k - \hat{y}\|^2 \leq \|x_0 - \hat{y}\|^2 - 2k\alpha\epsilon$, which fails to hold when k is sufficiently large.

Gradient Descent with a Fixed Step Size

We assume that

$$\|\nabla f(x) - \nabla f(y)\| \leq M\|x - y\|$$

It means that the first derivative of f is Lipschitz continuous.

Gradient Descent with a Fixed Step Size

Theorem: Under the Lipschitz gradient assumption, and if $\inf f(x) = f^* > -\infty$, then the gradient algorithm with $t < \frac{2}{M}$ has the following property

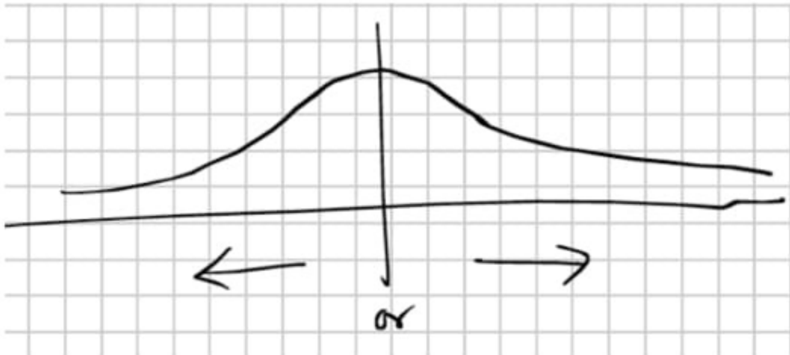
$$\lim_{k \to \infty} \nabla f(x_k) = 0$$

- f needs not be convex
- This also implies $\lim_{k \to \infty} \|\nabla f(x_k)\| = 0$

Gradient Descent with a Fixed Step Size

For the nonconvex case, the result does not imply that $\lim_{k\to\infty} x_k = x^*$, i.e., does not imply the convergence to the optimal point

Example: $f(x) = \frac{1}{1+x^2}$. Note that f is not convex



We can show that $f''(x)$ is bounded, which implies $f'(x)$ is Lipschitz

Gradient Descent with a Fixed Step Size

- We can show that $f''(x)$ is bounded, which implies $f'(x)$ is Lipschitz
- $f'(x) = \frac{-2x}{(1+x^2)^2}$
- The gradient algorithm is

$$x_{k+1} = x_k + t\frac{2x}{(1+x^2)^2}$$

- The algorithm diverges

Gradient Descent with a Fixed Step Size

Proof of the theorem: Let

$$g(t) = f(x + t(y - x))$$

Then

$$g(1) = g(0) + \int_0^1 g'(t)dt$$

We have

$$\begin{aligned} f(y) &= f(x) + \int_0^1 \nabla f(x + t(y - x))^T (y - x) dt \\ &= f(x) + \nabla f(x)^T (y - x) + \int_0^1 \Big(\nabla f(x + t(y - x)) - \nabla f(x) \Big)^T (y - x) dt \end{aligned}$$

Gradient Descent with a Fixed Step Size

By C-S inequality, we have

$$
\begin{aligned}
&\leq f(x) + \nabla f(x)^T (y-x) + M \int_0^1 t\|y-x\|\|y-x\| dt \\
&= f(x) + \nabla f(x)^T (y-x) + \frac{M}{2}\|y-x\|^2
\end{aligned}
$$

This implies

$$
\begin{aligned}
f(x_{k+1}) &\leq f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \frac{M}{2}\|x_{k+1} - x_k\|^2 \\
&= f(x_k) + \nabla f(x_k)^T(-t\nabla f(x_k)) + \frac{M}{2} t^2 \|\nabla f(x_k)\|^2 \\
&= f(x_k) - t(1 - \frac{M}{2}t)\|\nabla f(x_k)\|^2 \\
&\Leftrightarrow \;\; t(1 - \frac{M}{2}t)\|\nabla f(x_k)\|^2 \leq f(x_k) - f(x_{k+1})
\end{aligned}
$$

Gradient Descent with a Fixed Step Size

$$\Leftrightarrow \quad t(1 - \frac{M}{2}t)\|\nabla f(x_k)\|^2 \leq f(x_k) - f(x_{k+1})$$

$$\Leftrightarrow \quad t(1 - \frac{M}{2}t)\sum_{k=1}^{n}\|\nabla f(x_k)\|^2 \leq f(x_k) - f(x_{n+1}) \leq f(x_k) - f^*$$

Since f^* is finite and note that $(1 - \frac{M}{2}t) > 0$, we must have

$$\limsup_{n\to\infty}\sum_{k=1}^{n}\|\nabla f(x_k)\|^2 < \infty$$

Therefore

$$\lim_{n\to\infty}\|\nabla f(x_k)\|^2 = 0$$

Gradient Descent with a Fixed Step Size

If f is convex, then the gradient algorithm converges to the optimal point.

Theorem: Suppose that f is convex, and f satisfies all the statements in the previous theorem. Then the gradient descent algorithms converges to the optimal point, that is, $\lim_{k\to\infty} x_k = x^*$, where

$$x^* \in X_{opt} = \{\bar{x} \ : \ f(\bar{x}) = \inf_x f(x)\}$$

Note that $t < \frac{2}{M}$

- We note that for the nonconvex optimization problem, the gradient algorithm converges when $t < \frac{1}{M}$. But this guarantees only the local optimality.

Gradient Descent with a Fixed Step Size

Let $\tilde{x} \in X_{opt}$. Then

$$x_{k+1} - \tilde{x} = x_k - \tilde{x} - t\nabla f(x_k)$$

We have

$$\begin{aligned} \|x_{k+1} - \tilde{x}\|^2 &= \|x_k - \tilde{x}\|^2 + t^2\|\nabla f(x_k)\|^2 - 2t\nabla f(x_k)^T(x_k - \tilde{x}) \\ &= \|x_k - \tilde{x}\|^2 + t^2\|\nabla f(x_k)\|^2 + 2t\nabla f(x_k)^T(\tilde{x} - x_k) \end{aligned}$$

Gradient Descent with a Fixed Step Size

Since f is convex

$$f(x_k) + \nabla f(x_k)^T (\tilde{x} - x_k) \leq f(\tilde{x})$$

We have

$$\begin{aligned} \|x_{k+1} - \tilde{x}\|^2 &\leq \|x_k - \tilde{x}\|^2 + t^2 \|\nabla f(x_k)\|^2 + 2t(f(\tilde{x}) - f(x_k)) \\ &\leq \|x_k - \tilde{x}\|^2 + t^2 \|\nabla f(x_k)\|^2 \end{aligned}$$

since $(f(\tilde{x}) - f(x_k)) \leq 0$

Gradient Descent with a Fixed Step Size

Adding $t^2 \sum_{i=k+1}^{\infty} \|\nabla f(x_i)\|^2$ to both sides, we have

$$\|x_{k+1} - \tilde{x}\|^2 + t^2 \sum_{i=k+1}^{\infty} \|\nabla f(x_i)\|^2$$
$$\leq \underbrace{\|x_k - \tilde{x}\|^2 + t^2 \sum_{i=k}^{\infty} \|\nabla f(x_i)\|^2}_{u_k}$$

u_k is a non-increasing sequence, and is lower bounded by zero. Hence, there exists u^* such that

$$\lim_{k\to\infty} u_k = u^* < \infty$$

Gradient Descent with a Fixed Step Size

Since $\lim_{k\to\infty} \|\nabla f(x_k)\| = 0$,

$$\lim_{k\to\infty} \|x_k - \tilde{x}\|^2$$

exists and is finite for any $\tilde{x} \in X_{opt}$. This means that $\{x_k\}$ is bounded. Then there exists a subsequence $\{x_{n_k}\}$ such that

$$\lim_{k\to\infty} x_{n_k} = \bar{x}$$

for some $\bar{x}$. Note that until now, $\bar{x}$ needs not be in X_{opt}. But by continuity of f and the previous theorem, we have

$$\lim_{k\to\infty} \nabla f(x_{n_k}) = \nabla f(\bar{x}) = 0$$

Hence, $\bar{x} \in X_{opt}$.

Gradient Descent with a Fixed Step Size

We know that

$$\lim_{k\to\infty} \|x_k - \bar{x}\|^2$$

exists and is finite. Note that this hold for every subsequence, therefore, we must have

$$\lim_{k\to\infty} \|x_k - \bar{x}\|^2 = 0$$

Hence, $\lim_{k\to\infty} x_k = \bar{x} \in X_{opt}$

- The convexity of f allows the gradient descent algorithm converges to some element of X_{opt}. As we have seen, without the convexity, the algorithm may not converge to the optimal point.

Gradient Descent with a Fixed Step Size

- The convexity of f allows the gradient descent algorithm converges to some element of X_{opt}. As we have seen, without the convexity, the algorithm may not converge to the optimal point.
- We want to know how fast the gradient descent algorithm converges to the optimal point.
- To do this, we need an additional condition
- Strong convexity

$$\nabla^2 f(x) \geq mI$$

By definition, strong convexity implies strict convexity. And we have

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{1}{2} m \|y - x\|^2$$

- The Lipschitz condition implies

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{1}{2} M \|y - x\|^2$$

Gradient Descent with a Fixed Step Size

Implication of the strong convexity

$$
\begin{aligned}
f(y) &\geq f(x) + \nabla f(x)^T(y-x) + \frac{1}{2}m\|y-x\|^2 \\
&\geq f(x) + \nabla f(x)^T(\bar{y}-x) + \frac{1}{2}m\|\bar{y}-x\|^2, \quad \bar{y} = x - (1/m)\nabla f(x) \\
&= f(x) - \frac{1}{2m}\|\nabla f(x)\|^2
\end{aligned}
$$

Hence,

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

This implies

$$
\|\nabla f(x)\| \leq (2m\epsilon)^{1/2} \Rightarrow f(x) \leq f^* + \epsilon
$$

Gradient Descent with a Fixed Step Size

Implication of the strong convexity

$$
\begin{aligned}
f^* &\geq f(x) + \nabla f(x)^T (x^* - x) \frac{m}{2} \|x^* - x\|^2 \\
&\geq f(x) - \|\nabla f(x)\| \|(x^* - x)\| + \frac{m}{2} \|x^* - x\|^2
\end{aligned}
$$

This implies

$$
-\|\nabla f(x)\| \|(x^* - x)\| + \frac{m}{2} \|x^* - x\|^2 \leq 0
$$

And we have

$$
\|x^* - x\| \leq \frac{2}{m} \|\nabla f(x)\|
$$

Gradient Descent with a Fixed Step Size

Note that the strong convexity implies the strict convexity; hence, X_{opt} is singleton, that is, x^* is unique. We have

$$\begin{aligned}
\|x_{k+1} - x^*\|^2 &= \|x_k - t\nabla f(x_k) - x^* + t\nabla f(x^*)\|^2 \\
&= \|(x_k - x^*) - t(\nabla f(x_k) - \nabla f(x^*))\|^2 \\
&= \|x_k - x^*\|^2 + t^2\|\nabla f(x_k) - \nabla f(x^*)\|^2 \\
&\quad + 2t(x_k - x^*)^T(\nabla f(x_k) - \nabla f(x^*)) \\
&\leq \|x_k - x^*\|^2 + t^2 M^2\|x_k - x^*\|^2 \\
&\quad + 2t(x^* - x_k)^T\nabla f(x_k) + 2t(x_k - x^*)^T\nabla f(x^*) \\
&= \|x_k - x^*\|^2 + t^2 M^2\|x_k - x^*\|^2 \\
&\quad - 2tf(x^*) + 2t(x^* - x_k)^T\nabla f(x_k) \\
&\quad + 2tf(x^*) + 2t(x_k - x^*)^T\nabla f(x^*)
\end{aligned}$$

Gradient Descent with a Fixed Step Size

By the first order optimality condition, we have

$$\begin{aligned}
\|x_{k+1} - x^*\|^2 &\le \|x_k - x^*\|^2 + t^2 M^2 \|x_k - x^*\|^2 \\
&\quad - 2tf(x^*) + 2t(x^* - x_k)^T \nabla f(x_k) + 2tf(x_k) \\
&\le \|x_k - x^*\|^2 + t^2 M^2 \|x_k - x^*\|^2 - \frac{1}{2} m \|x_k - x^*\|^2 2t \\
&= (1 - mt + M^2 t^2) \|x_k - x^*\|^2
\end{aligned}$$

This implies

$$\|x_k - x^*\|^2 \le (1 - mt + M^2 t^2)^k \|x_0 - x^*\|^2$$

Hence, if $|1 - mt + M^2 t^2| < 1$, then then it converges exponentially fast

Gradient Descent with a Fixed Step Size

Theorem

Suppose that f is strongly convex, i.e., $mI \leq \nabla^2 f(x) \leq MI$ for all x. Then with the constant stepsize $t < \frac{\min(2,m)}{M}$,

$$\|x_k - x^*\| \leq cq^k \quad 0 < q < 1$$

That is, the gradient descent algorithm has the geometric convergence rate.

Remark: Using the strong convexity assumption,

$$\|x_{k+1} - x^*\|^2 \leq (1 - mt + t^2 M)^{k+1} \|x_0 - x^*\|^2$$

- Geometric convergence is not bad, but there are not many functions that satisfy the strong convexity assumption
- Note that with the strongly convexity, x^* is unique

Gradient Descent with a Variable Step Size

Now, we consider the case when t_k is a function of k (varying with respect to k)

Recall that

$$x_{k+1} = x_k - t_k \nabla f(x_k)$$

Two approaches

- Exact line search: impractical

$$t = \arg\min_{s \geq 0} f(x + s\Delta x)$$

  Find t such that it descent direction minimizes the objective function

$$f(x + t\Delta x) \leq f(x + s\Delta x), \ \forall s \geq 0$$

- Backtracking search (Armijo's rule): quite practical

Gradient Descent with a Variable Step Size

Algorithm 9.3 *Gradient descent method.*

given a starting point $x \in \textbf{dom}\, f$.

repeat

1. $\Delta x := -\nabla f(x)$.
2. *Line search.* Choose step size t via exact or backtracking line search.
3. *Update.* $x := x + t\Delta x$.

until stopping criterion is satisfied.

Algorithm 9.2 *Backtracking line search.*

given a descent direction Δx for f at $x \in \textbf{dom}\, f$, $\alpha \in (0, 0.5)$, $\beta \in (0, 1)$.

$t := 1$.

while $f(x + t\Delta x) > f(x) + \alpha t \nabla f(x)^T \Delta x, \quad t := \beta t$.

Gradient Descent with a Variable Step Size

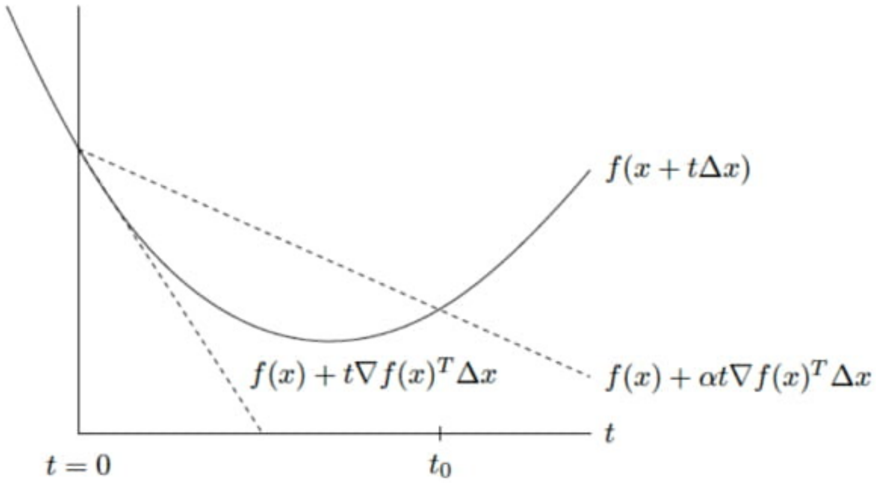


Figure 9.1 *Backtracking line search.* The curve shows f, restricted to the line over which we search. The lower dashed line shows the linear extrapolation of f, and the upper dashed line has a slope a factor of α smaller. The backtracking condition is that f lies below the upper dashed line, *i.e.*, $0 \leq t \leq t_0$.

Unconstrained Minimization Problem

Consider: $f : \mathbb{R}^n \to \mathbb{R}$

$$\min_x f(x)$$

where f is twice differentiable and convex.

- Gradient method with the fixed and variable step sizes: geometric convergence
- We want a faster algorithm: Newton method

Summary of the gradient descent method

Unconstrained Minimization

$$\min_{x \in \mathbb{R}^n} f(x)$$

- Assumption
  - f is convex and continuously differentiable
  - The optimal value $f^* = \inf_{x \in \mathbb{R}^n} f(x)$ is finite
- Minimization methods
  - Iterative methods for the form

$$x_{k+1} = x_k + \alpha_k d_k, \; x_0 = x \in \mathbb{R}^n$$

    α_k: step size
    d_k: direction of the iterative algorithm
  - Generate a sequence of points $\{x_k\}$ such that $f(x_k) \to f^*$ as $k \to \infty$
  - Can be interpreted as iterative methods for solving the system of equations to satisfy the necessary and sufficient optimality condition

$$\nabla f(x^*) = 0$$

Unconstrained Minimization: Gradient Descent Method

- Gradient descent algorithm

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k), \ x_0 = x$$

- α_k: stepsize
  - Constant
  - Diminishing: $\alpha_k \to 0$ with $\sum_{k=1}^{\infty} \alpha_k = \infty$
  - Linear search types (optimal stepsize, hard to find)
    - Exact line search: $\alpha_k = \arg\min_{\alpha > 0} f(x_k + \alpha d_k)$
    - Backtracking line search

Gradient Descent with Bounded Gradients

Theorem

Suppose that the gradient is bounded, that is, for some $L > 0$

$$\|\nabla f(x)\| \leq L \quad \forall x, y \in \mathbb{R}^n.$$

Let the stepsize be constant, i.e., $\alpha_k = \alpha$. Then the gradient descent algorithm generates the sequence $\{x_k\}$ such that

$$\liminf_{k\to\infty} f(x_k) \leq f^* + \frac{\alpha L}{2}$$

Also, when diminishing step size is used, i.e., $\sum_{k=0}^{\infty} \alpha_k = \infty$, $\lim_{k\to\infty} f(x_k) = f^*$.

Proof: Assume that the result does not hold, i.e., for some $\hat{y}$ with $f(\hat{y}) = f^* + \epsilon$,

$$f(x_k) - f(\hat{y}) \geq \frac{\alpha L}{2} + \epsilon, \ \forall k.$$

Then we can show that $\|x_k - \hat{y}\|^2 \leq \|x_0 - \hat{y}\|^2 - 2k\alpha\epsilon$, which fails to hold when k is sufficiently large.

Gradient Descent with the Lipschitz Gradient

Theorem

Suppose that the gradient of f is Lipschitz continuous, i.e., for some $M > 0$

$$\|\nabla f(x) - \nabla f(y)\| \leq M\|x - y\| \quad \forall x, y \in \mathbb{R}^n$$

Then for the constant stepsize $\alpha \leq \frac{2}{M}$, we have

$$\lim_{k\to\infty} \|\nabla f(x_k)\| = 0$$

Furthermore, if X^* is nonempty, then the gradient descent algorithm converges to the optimal point.

Proof: By using the Lipschitz constant,

$$f(x_{k+1}) \leq f(x_k) - \frac{\alpha}{2}(2 - \alpha M)\|f(x_k)\|^2.$$

Then with $2 - M\alpha \geq 0$, $\sum_{k=1}^{\infty} \|\nabla^2 f(x_k)\|^2 < \infty$.

Gradient Descent with Strong Convexity

Theorem

Suppose that f is strongly convex, i.e., $mI \leq \nabla^2 f(x) \leq MI$ for all x. Then with the constant stepsize $\alpha < \frac{\min(2,m)}{M}$,

$$\|x_k - x^*\| \leq cq^k \quad 0 < q < 1$$

That is, the gradient descent algorithm has the geometric convergence rate.

Using the strong convexity assumption,

$$\|x_{k+1} - x^*\|^2 \leq (1 - m\alpha + \alpha^2 M)^{k+1} \|x_0 - x^*\|^2$$

- Geometric convergence is not bad, but there are not many functions that satisfy the strong convexity assumption
- Note that with the strongly convexity, x^* is unique