Optimization Techniques

Optimization Techniques

2013 Lecture 7

Optimization problem

Consider the problem

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

Optimization problem

Consider the problem

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

Recall that a maximization problem can be reduced to a minimization problem as follows:

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

Optimization problem

Consider the problem

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

Recall that a maximization problem can be reduced to a minimization problem as follows:

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

In this part of the course we introduce notations

$$\begin{aligned} g(x) &\equiv \nabla f(x) = \left(\frac{\partial f}{\partial x_i} \right)_{i=1}^{n} \\ H(x) &= \left(\frac{\partial^2 f}{\partial x_i \partial x_j} \right)_{i,j=1}^{n} \end{aligned}$$

Optimization problem

Consider the problem

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

Recall that a maximization problem can be reduced to a minimization problem as follows:

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

In this part of the course we introduce notations

$$\begin{aligned} g(x) &\equiv \nabla f(x) = \left(\frac{\partial f}{\partial x_i} \right)_{i=1}^{n} \\ H(x) &= \left(\frac{\partial^2 f}{\partial x_i \partial x_j} \right)_{i,j=1}^{n} \end{aligned}$$

Also, all vectors $x \in \mathbb{R}^n$ are column vectors, $x = (x_1, x_2, \ldots, x_n)^T$

Generic Linesearch Method

Definition

A direction $p_k \in \mathbb{R}^n$ is called a descent direction if

$$p_k^T g_k < 0 \quad \textit{if} \quad g_k \neq 0$$

Generic Linesearch Method

Definition

A direction $p_k \in \mathbb{R}^n$ is called a descent direction if

$$p_k^T g_k < 0 \quad \textit{if} \quad g_k \neq 0$$

1. Given an initial guess x_0, let $k = 0$
2. Until convergence:
    1. Find a descent direction p_k at x_k.
    2. Compute a stepsize α_k using a backtracking-Armijo linesearch along p_k.
    3. Set $x_{k+1} = x_k + \alpha_k p_k$, and increase k by 1.

Generic Linesearch Method

Definition

A direction $p_k \in \mathbb{R}^n$ is called a descent direction if

$$p_k^T g_k < 0 \quad \textit{if} \quad g_k \neq 0$$

1. Given an initial guess x_0, let $k = 0$
2. Until convergence:
   1. Find a descent direction p_k at x_k.
   2. Compute a stepsize α_k using a backtracking-Armijo linesearch along p_k.
   3. Set $x_{k+1} = x_k + \alpha_k p_k$, and increase k by 1.

In the steepest descent direction:$p_k = -g_k$.

Generic Linesearch Method

Definition

A direction $p_k \in \mathbb{R}^n$ is called a descent direction if

$$p_k^T g_k < 0 \quad \textit{if} \quad g_k \neq 0$$

1. Given an initial guess x_0, let $k = 0$
2. Until convergence:
    1. Find a descent direction p_k at x_k.
    2. Compute a stepsize α_k using a backtracking-Armijo linesearch along p_k.
    3. Set $x_{k+1} = x_k + \alpha_k p_k$, and increase k by 1.

In the steepest descent direction:$p_k = -g_k$. But convergence is slow, and numerical convergence sometimes does not occur at all.

Generic Linesearch Method

Definition

A direction $p_k \in \mathbb{R}^n$ is called a descent direction if

$$p_k^T g_k < 0 \quad if \quad g_k \neq 0$$

1. Given an initial guess x_0, let $k = 0$
2. Until convergence:
    1. Find a descent direction p_k at x_k.
    2. Compute a stepsize α_k using a backtracking-Armijo linesearch along p_k.
    3. Set $x_{k+1} = x_k + \alpha_k p_k$, and increase k by 1.

In the steepest descent direction:$p_k = -g_k$. But convergence is slow, and numerical convergence sometimes does not occur at all. Therefore, steepest-descent is rarely used in most cases. Instead other linesearch methods are being used.

Newton and Newton-like methods

Lemma

Let B_k be a symmetric, positive definite matrix. Then p_k such that $B_k p_k = -g_k$ is a search direction.

Newton and Newton-like methods

Lemma

Let B_k be a symmetric, positive definite matrix. Then p_k such that $B_k p_k = -g_k$ is a search direction.

Proof.

Suppose B_k is positive definite matrix and $B_k p_k = -g_k$. Therefore for any direction p_k:

$$\begin{aligned} p_k^T B_k p_k &> 0 \\ p_k^T(-g_k) &> 0 \\ p_k^T g_k &< 0 \end{aligned}$$

Newton and Newton-like methods

Lemma

Let B_k be a symmetric, positive definite matrix. Then p_k such that $B_k p_k = -g_k$ is a search direction.

Proof.

Suppose B_k is positive definite matrix and $B_k p_k = -g_k$. Therefore for any direction p_k:

$$
\begin{aligned}
p_k^T B_k p_k &> 0 \\
p_k^T (-g_k) &> 0 \\
p_k^T g_k &< 0
\end{aligned}
$$

So p_k is a search direction. □

Newton and Newton-like methods

Of particular interest is the possibility that $B_k = H_k$.

Newton and Newton-like methods

Of particular interest is the possibility that $B_k = H_k$. The resulting direction for which

$$H_k p_k = -g_k$$

is known as the **Newton direction**, and any method which uses it is a **Newton method**.

Newton and Newton-like methods

Of particular interest is the possibility that $B_k = H_k$. The resulting direction for which

$$H_k p_k = -g_k$$

is known as the **Newton direction**, and any method which uses it is a **Newton method**. But notice that the Newton direction is only guaranteed to be useful in a linesearch context if the Hessian H_k is positive definite,

Newton and Newton-like methods

Of particular interest is the possibility that $B_k = H_k$. The resulting direction for which

$$H_k p_k = -g_k$$

is known as the **Newton direction**, and any method which uses it is a **Newton method**. But notice that the Newton direction is only guaranteed to be useful in a linesearch context if the Hessian H_k is positive definite, otherwise p_k might turn out to be an ascent direction.

Newton and Newton-like methods

Of particular interest is the possibility that $B_k = H_k$. The resulting direction for which

$$H_k p_k = -g_k$$

is known as the **Newton direction**, and any method which uses it is a **Newton method**. But notice that the Newton direction is only guaranteed to be useful in a linesearch context if the Hessian H_k is positive definite, otherwise p_k might turn out to be an ascent direction.

If H_k is positive definite, then the descent search direction is

$$-H_k^{-1} g_k$$

Newton and Newton-like methods

Theorem

Suppose that $f \in C^1$ and that g is Lipschitz continuous on $\mathbb{R}^n$. Then, for the iterates generated by the Generic Linesearch Method using the Newton or Newton-like direction, either

$$g_l = 0 \quad \textit{for some } l$$

or

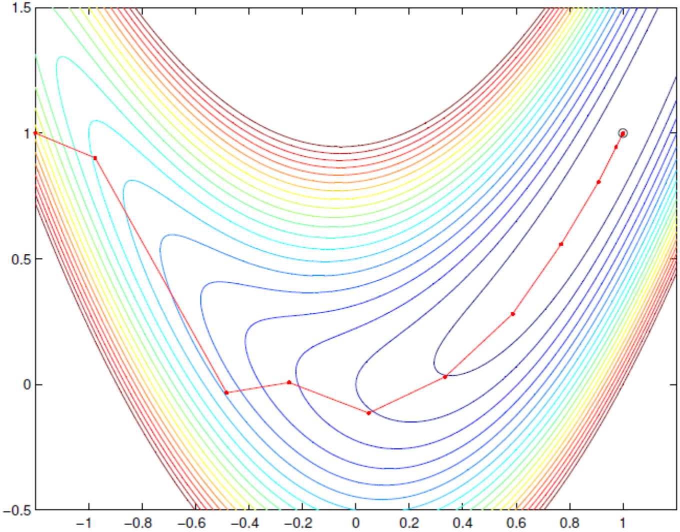
$$\lim_{k \to \infty} f_k = -\infty$$

or

$$\lim_{k \to \infty} g_k = 0$$

provided that the eigenvalues of B_k are uniformly bounded and bounded away from zero.

Newton and Newton-like methods



Newton and Newton-like methods

Indeed, one can regard such methods as "scaled" steepest descent, but they have the advantage that they can be made scale invariant for suitable B_k, and

Newton and Newton-like methods

Indeed, one can regard such methods as "scaled" steepest descent, but they have the advantage that they can be made scale invariant for suitable B_k, and their convergence is often significantly faster than steepest descent.

Newton and Newton-like methods

Indeed, one can regard such methods as "scaled" steepest descent, but they have the advantage that they can be made scale invariant for suitable B_k, and their convergence is often significantly faster than steepest descent. In particular, in the case of the Newton direction, the Generic Linesearch method will usually converge very rapidly.

Newton and Newton-like methods

Indeed, one can regard such methods as "scaled" steepest descent, but they have the advantage that they can be made scale invariant for suitable B_k, and their convergence is often significantly faster than steepest descent. In particular, in the case of the Newton direction, the Generic Linesearch method will usually converge very rapidly.

It can be shown theoretically, that convergence of Newton method under suitable conditions will be quadratic.

Newton and Newton-like methods

Indeed, one can regard such methods as "scaled" steepest descent, but they have the advantage that they can be made scale invariant for suitable B_k, and their convergence is often significantly faster than steepest descent. In particular, in the case of the Newton direction, the Generic Linesearch method will usually converge very rapidly.

It can be shown theoretically, that convergence of Newton method under suitable conditions will be quadratic.
In other words, there exists a positive constant C such that

$$\lim_{k \to \infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|^2} = C$$

Modified Newton methods

Away from a local minimizer there is no reason to believe that H_k will be positive definite, so precautions need to be taken to ensure that Newton and Newton-like linesearch methods, for which B_k is (or is close to) H_k, satisfy the assumptions of the global convergence

Modified Newton methods

Away from a local minimizer there is no reason to believe that H_k will be positive definite, so precautions need to be taken to ensure that Newton and Newton-like linesearch methods, for which B_k is (or is close to) H_k, satisfy the assumptions of the global convergence

If H_k is indefinite, it is usual to solve instead

$$(\underbrace{H_k + M_k}_{\equiv B_k})pk \equiv B_k p_k = -g_k;$$

where Mk is chosen so that $B_k = H_k + M_k$ is "sufficiently" positive definite and $M_k = 0$ when H_k is itself "sufficiently" positive definite.

Away from a local minimizer there is no reason to believe that H_k will be positive definite, so precautions need to be taken to ensure that Newton and Newton-like linesearch methods, for which B_k is (or is close to) H_k, satisfy the assumptions of the global convergence

If H_k is indefinite, it is usual to solve instead

$$(\underbrace{H_k + M_k}_{\equiv B_k})pk \equiv B_k p_k = -g_k;$$

where Mk is chosen so that $B_k = H_k + M_k$ is "sufficiently" positive definite and $M_k = 0$ when H_k is itself "sufficiently" positive definite.

This may be achieved in a number of ways.

Modified Newton methods

Away from a local minimizer there is no reason to believe that H_k will be positive definite, so precautions need to be taken to ensure that Newton and Newton-like linesearch methods, for which B_k is (or is close to) H_k, satisfy the assumptions of the global convergence

If H_k is indefinite, it is usual to solve instead

$$(\underbrace{H_k + M_k}_{\equiv B_k})pk \equiv B_k p_k = -g_k;$$

where Mk is chosen so that $B_k = H_k + M_k$ is "sufficiently" positive definite and $M_k = 0$ when H_k is itself "sufficiently" positive definite.

This may be achieved in a number of ways. There several choices for choosing M_k : using spectral decomposition or Cholesky decomposition.

Quasi-Newton methods

It was fashionable in the 1960s and 1970s to attempt to build suitable approximations B_k to the Hessian, H_k.

Quasi-Newton methods

It was fashionable in the 1960s and 1970s to attempt to build suitable approximations B_k to the Hessian, H_k.

Activity in this area has subsequently died down, possibly because people started to realize that computing exact second derivatives was not as onerous as they had previously contended,

Quasi-Newton methods

It was fashionable in the 1960s and 1970s to attempt to build suitable approximations B_k to the Hessian, H_k.

Activity in this area has subsequently died down, possibly because people started to realize that computing exact second derivatives was not as onerous as they had previously contended,

but these techniques are still of interest particularly when gradients are awkward to obtain (such as when the function values are simply given as the result of some other, perhaps hidden, computation).

Quasi-Newton methods

It was fashionable in the 1960s and 1970s to attempt to build suitable approximations B_k to the Hessian, H_k.

Activity in this area has subsequently died down, possibly because people started to realize that computing exact second derivatives was not as onerous as they had previously contended,

but these techniques are still of interest particularly when gradients are awkward to obtain (such as when the function values are simply given as the result of some other, perhaps hidden, computation).

There are broadly two classes of what may be called quasi-Newton methods: by finite differencies or by secant approximations.