Overview of Algorithms

Oscar Dalmau
dalmau@cimat.mx

Centro de Investigación en Matemáticas
CIMAT A.C. Mexico

February 9, 2021

Outline

1. Algorithm overview
2. Convergence order

Example

Example 1.1

Minimize $f(x, y) = xe^{-x^2-y^2}$

Example

Example

Minimize $f(x,y) = xe^{-x^2-y^2}$

Gradient:

$$\nabla f(x,y) \quad = \quad e^{-x^2-y^2} \left[\begin{array}{c} 1-2x^2 \\ -2xy \end{array} \right]$$

Stationary points:

$$\left[\begin{array}{c} x^* \\ y^* \end{array} \right] \quad = \quad \left[\begin{array}{c} \pm\frac{\sqrt{2}}{2} \\ 0 \end{array} \right]$$

Example

Example

Minimize $f(x,y) = xe^{-x^2-y^2}$

Hessian:

$$\nabla^2 f(x,y) = e^{-x^2-y^2} \begin{bmatrix} 2x(2x^2-3) & 2y(2x^2-1) \\ 2y(2x^2-1) & 2x(2y^2-1) \end{bmatrix}$$

Example

Example

Minimize $f(x,y) = xe^{-x^2-y^2}$

Hessian at $[x^*, y^*]^T = [\frac{\sqrt{2}}{2}, 0]^T$

$$\nabla^2 f(x,y) \quad = \quad \sqrt{\frac{2}{e}} \begin{bmatrix} -2 & 0 \\ 0 & -1 \end{bmatrix}$$

then is al local maximum!

Example

Example

Minimize $f(x, y) = xe^{-x^2-y^2}$

Hessian at $[x^*, y^*]^T = [-\frac{\sqrt{2}}{2}, 0]^T$

$$\nabla^2 f(x, y) \quad = \quad \sqrt{\frac{2}{e}} \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$$

then is al local minimum!

Example

Example 1.2

Minimize $f(x, y) = x^2 + y^2 + e^{x+y}$

Example

Example

Minimize $f(x,y) = x^2 + y^2 + e^{x+y}$

Gradient:

$$\nabla f(x,y) = \begin{bmatrix} 2x + e^{x+y} \\ 2y + e^{x+y} \end{bmatrix}$$

Stationary points: solve the following system of equation!!

$$\begin{bmatrix} 2x + e^{x+y} \\ 2y + e^{x+y} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Then $x = y$ and $-2x = e^{2x}$. It requires a numerical method!, see matlab script

General Framework

- Algorithms for unconstrained minimization are iterative methods that find an approximate solution.
- All algorithms for unconstrained minimization require the user to supply a starting point, which we usually denote by $\boldsymbol{x}_0$.
- The user with knowledge about the application and the data set may be in a good position to choose $\boldsymbol{x}_0$ to be a reasonable estimate of the solution.
- Otherwise, the starting point must be chosen by the algorithm, either by a systematic approach or in some arbitrary manner.

General Framework

- Starting at $\boldsymbol{x}_0$, optimization algorithms generate a sequence of iterates $\{\boldsymbol{x}_k\}_{k=0}^{\infty}$ that terminate when either no more progress can be made or when it seems that a solution point has been approximated with sufficient accuracy.
- In deciding how to move from one iterate $\boldsymbol{x}_k$ to the next, the algorithms use information about the function f at $\boldsymbol{x}_k$, and possibly also information from earlier iterates $\boldsymbol{x}_0, \boldsymbol{x}_1, ..., \boldsymbol{x}_{k-1}$.
- They use this information to find a new iterate $\boldsymbol{x}_{k+1}$ with a lower function value than $\boldsymbol{x}_k$.

General Framework

1. Start at $\boldsymbol{x}_0$, $k = 0$
2. While not converge
   - Find $\boldsymbol{x}_{k+1}$ such that $f(\boldsymbol{x}_{k+1}) < f(\boldsymbol{x}_k)$
   - $k = k + 1$
3. Return $\boldsymbol{x}^* = \boldsymbol{x}_k$

General Framework: Comment

- However, there exist non-monotone algorithms in which f does not decrease at every step, but f should decrease after some number m of iterations that is, $f(\boldsymbol{x}_{k+1}) < f(\boldsymbol{x}_{k-j})$ for some $j \in \mathcal{M} = \{0, 1, \cdots, M\}$ with $M = m - 1$ if $k \geq m - 1$ otherwise $M = k$.
  For example, select $x_{k+1} = x_k + \alpha d_k$, $d_k = -g(x_k)/\|g(x_k)\|$ if

$$f(x_k + \alpha d_k) < \max_{j \in \mathcal{M}} f(\boldsymbol{x}_{k-j}) + \gamma \alpha g(x_k)^T d_k$$

Note: See details, in Grippo86NonMonotoneLineSearch.pdf, in internet download Grippo86.pdf

General Framework

1. How to choose $\boldsymbol{x}_0$?
2. Find a convergence or stop criteria?
3. How to update $\boldsymbol{x}_{k+1}$?

Updating formula

The algorithm chooses a direction $\boldsymbol{d}_k$ and searches along this direction from the current iterate $\boldsymbol{x}_k$ for a new iterate with a lower function value (line search strategy).

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha \boldsymbol{d}_k$$

Descent direction

Definition 1.3

A descent direction is a vector $\boldsymbol{d} \in \mathbb{R}^n$ such that $f(\boldsymbol{x} + t\boldsymbol{d}) < f(\boldsymbol{x})$, $t \in (0, T)$ i.e., allows to move a point $\boldsymbol{x}$ closer towards a local minimum $\boldsymbol{x}^*$ of the objective function $f : \mathbb{R}^n \to \mathbb{R}$.

There are several methods that compute descent directions, for example: use gradient descent, conjugate gradient method.

Descent direction

Descent direction

If $g(\boldsymbol{x})^T \boldsymbol{d} < 0$ then $\boldsymbol{d}$ is a descent direction.

There exists $\hat{t}$ such that $g(\boldsymbol{x} + t\boldsymbol{d})^T \boldsymbol{d} < 0$ for all $t \in [0, \hat{t}]$ (sign preserving theorem).
Using Taylor, there exist $\tau \in (0, 1)$ such that

$$f(\boldsymbol{x} + \hat{t}\boldsymbol{d}) \quad = \quad f(\boldsymbol{x}) + \hat{t}\boldsymbol{g}(\boldsymbol{x} + \tau\hat{t}\boldsymbol{d})^T \boldsymbol{d}$$

as $0 < t = \tau\hat{t} < \hat{t}$ then $\boldsymbol{g}(\boldsymbol{x} + \tau\hat{t}\boldsymbol{d})^T \boldsymbol{d} = \boldsymbol{g}(\boldsymbol{x} + t\boldsymbol{d})^T \boldsymbol{d} < 0$ and therefore $f(\boldsymbol{x} + \hat{t}\boldsymbol{d}) < f(\boldsymbol{x})$ then $\boldsymbol{d}$ is a descent direction.

Line search methods

Line search methods

- First, the algorithm chooses a direction $\boldsymbol{d}_k$
- Then, it searches along this direction from the current iterate $\boldsymbol{x}_k$ for a new iterate with a lower function value. The distance to move along $\boldsymbol{d}_k$ can be found by approximately solving the following one-dimensional minimization problem to find a step length α:

$$\alpha_k = \arg\min_{\alpha>0} f(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k)$$

Search directions for line search method

The steepest descent direction

For example $\boldsymbol{d}_k = -\boldsymbol{g}(\boldsymbol{x}_k)$ is the most obvious choice for search direction

- The *steepest descent method* is a line search method that moves along $\boldsymbol{d}_k = -\boldsymbol{g}(\boldsymbol{x}_k)$ at every step.
- Line search methods may use search directions other than the steepest descent direction.
- In general, any descent direction, one that makes an angle of strictly less than $\pi/2$ radians with $-\boldsymbol{g}(\boldsymbol{x}_k)$, is guaranteed to produce a decrease in f, i.e. if $\boldsymbol{g}(\boldsymbol{x}_k)^T\boldsymbol{d}_k < 0$ then $|\angle(\boldsymbol{g}(\boldsymbol{x}_k), \boldsymbol{d}_k)| > \pi/2$, i.e. $|\angle(-\boldsymbol{g}(\boldsymbol{x}_k), \boldsymbol{d}_k)| < \pi/2$, due to $\boldsymbol{g}(\boldsymbol{x}_k)^T\boldsymbol{d}_k = \|\boldsymbol{g}(\boldsymbol{x}_k)\|\|\boldsymbol{d}_k\| \cos\angle(\boldsymbol{g}(\boldsymbol{x}_k), \boldsymbol{d}_k)$.

Newton direction

- Another important search direction, perhaps the most important one of all, is the *Newton direction*.
- This direction is derived from the second-order Taylor series approximation to $f(\boldsymbol{x}_k + \boldsymbol{d})$

$$f(\boldsymbol{x}_k + \boldsymbol{d}) \quad \approx \quad f(\boldsymbol{x}_k) + \boldsymbol{g}(\boldsymbol{x}_k)^T \boldsymbol{d} + \frac{1}{2}\boldsymbol{d}^T \mathbf{H}(\boldsymbol{x}_k)\boldsymbol{d} \stackrel{def}{=} m_k(\boldsymbol{d})$$

Newton direction

$\nabla_{\boldsymbol{d}} m_k(\boldsymbol{d}) = 0$ then $\boldsymbol{d}_k^N = -\mathbf{H}(\boldsymbol{x}_k)^{-1}\boldsymbol{g}(\boldsymbol{x}_k)$ if there exists $\mathbf{H}(\boldsymbol{x}_k)^{-1}$

Newton direction

- The Newton direction can be used in a line search method when $\mathbf{H}(\boldsymbol{x}_k)$ is positive definite.
- Most line search implementations of Newton's method use the unit step $\alpha = 1$
- When $\mathbf{H}(\boldsymbol{x}_k)$ is not positive definite, the Newton direction may not even be defined, since $\mathbf{H}(\boldsymbol{x}_k)^{-1}$ may not exist.
- Even when it is defined, it may not satisfy the descent property $\boldsymbol{g}_k^T \boldsymbol{d}_k^N < 0$, in which case it is unsuitable as a search direction. In these situations, line search methods modify the definition of $\boldsymbol{d}_k$ to make it satisfy the descent condition.

Quasi-Newton methods

- Quasi-Newton methods are alternatives to Newton's methods which do not require computation of the Hessian.
- Instead of the true Hessian $\mathbf{H}_k$, they use an approximation $\mathbf{B}_k$, which is updated after each step.

$$m_k(\boldsymbol{d}) \quad \stackrel{def}{=} \quad f(\boldsymbol{x}_k) + \boldsymbol{g}(\boldsymbol{x}_k)^T \boldsymbol{d} + \frac{1}{2} \boldsymbol{d}^T \mathbf{B}_k \boldsymbol{d}$$

Quasi-Newton methods

- The approximation $\mathbf{B}_k$ to the Hessian is updated by using successive gradient vectors $\boldsymbol{g}_{k-1}, \boldsymbol{g}_k$ and positions $\boldsymbol{x}_{k-1}, \boldsymbol{x}_k$.
- Quasi-Newton methods are a generalization of the secant method to find the root of the first derivative for multidimensional problems.

Quasi-Newton methods

- **Recall**: The secant method is defined by the recurrence relation

$$\begin{aligned} x_{k+1} &= x_k - f(x_k)\frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} \\ x_{k+1} &= x_k - \frac{f(x_k)}{f'(x_k)}, \text{ (Newton)} \\ f'(x_k) &\approx \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}}, \text{ (Finite difference)} \end{aligned}$$

- Therefore, the *secant method* can be interpreted as a method in which the derivative is replaced by an approximation and then is a *Quasi-Newton method*.

Quasi-Newton methods

- Using Taylor Theorem, for the gradient function

$$\nabla f(\boldsymbol{x}+\boldsymbol{h}) \;=\; \nabla f(\boldsymbol{x}) + \nabla^2 f(\boldsymbol{x})\boldsymbol{h} + o(\|h\|)$$

defining $\boldsymbol{x}_k = \boldsymbol{x}$ and $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{h}$ then $\boldsymbol{h} = \boldsymbol{x}_{k+1} - \boldsymbol{x}_k$ and

$$\nabla f(\boldsymbol{x}_{k+1}) \;\approx\; \nabla f(\boldsymbol{x}_k) + \nabla^2 f(\boldsymbol{x}_k)(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k)$$

- Te previous approximation yields the known *secant equation* that should satisty $\mathbf{B}_k$ (approximation of $\mathbf{H}_k$)

$$\begin{aligned}
\mathbf{B}_{k+1}\boldsymbol{s}_k &= \boldsymbol{y}_k \\
\boldsymbol{s}_k &= \boldsymbol{x}_{k+1} - \boldsymbol{x}_k \\
\boldsymbol{y}_k &= \nabla f(\boldsymbol{x}_{k+1}) - \nabla f(\boldsymbol{x}_k)
\end{aligned}$$

General descent direction

Descent direction

If $\mathbf{A}_k$ is any positive definite matrix and $\boldsymbol{g}(\boldsymbol{x}_k) \neq \mathbf{0}$ then $\boldsymbol{d}_k = -\mathbf{A}_k \boldsymbol{g}(\boldsymbol{x}_k)$ is a descent direction

Convergence order

- Algorithms may differ significantly in their *computational efficiency*.
- A *fast or efficient algorithm* is one that requires only a small number of iterations to converge to a solution and the amount of computation is small.
- In general, in application one uses (or tries to use) the most efficient algorithm.
- How to measure *the rate of convergence* or the efficiency of the algorithms? .
- The most basic criterion is the *order of convergence* of a sequence.

Definition 2.1

Given a sequence $\{\boldsymbol{x}_k\}$ that converges to $\boldsymbol{x}^*$, that is, $\lim_{k\to\infty} \|\boldsymbol{x}_k - \boldsymbol{x}^*\| = 0$, we say that the *order of convergence*, of the sequence, is p, where $p \in \mathbb{R}$, if $0 < \beta < \infty$ where

$$\beta \quad := \quad \lim_{k\to\infty} \frac{\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|}{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^p}$$

If $\beta = 0$ for all p we say that the convergence order is ∞. The parameter β is called the convergence ratio (or rate of convergence).

- If $p = 1$ it is said that the sequence converges linearly to $\boldsymbol{x}^*$. The expression linear convergence is often reserved in the literature to the situation where $0 < \beta < 1$, whereas the situation $\beta = 1$ is referred to as sublinear and $\beta = 0$ superlinear convergence.
- If $p = 2, 3, \cdots$ it is said that the sequence converges quadratically, cubically, ... to x^*.

Convergence order: Examples

Example 2.2

Let $x_k = \frac{1}{k^n}$ for some fixed $n > 0$.

This sequence converges to $x^* = 0$. Now, we can compute the *rate of convergence*:

$$\begin{aligned}
\beta &= \lim_{k\to\infty} \frac{|x_{k+1} - x^*|}{|x_k - x^*|^p} \\
&= \lim_{k\to\infty} \frac{|x_{k+1}|}{|x_k|^p} \\
&= \lim_{k\to\infty} \left(\frac{k^p}{k+1}\right)^n = \begin{cases} 0, & \text{for } p < 1 \\ 1, & \text{for } p = 1 \\ \infty, & \text{for } p > 1 \end{cases}
\end{aligned}$$

We get convergence for $p = 1$, so this sequence converges linearly with rate of convergence $\beta = 1$, i.e., converges sublinearly

Convergence order: Examples

Example 2.3

Let $x_k = \gamma^k$ for $0 < \gamma < 1$.

This sequence converges to $x^* = 0$. Now, we can compute the *rate of convergence*:

$$\begin{aligned}
\beta &= \lim_{k\to\infty} \frac{|x_{k+1} - x^*|}{|x_k - x^*|^p} \\
&= \lim_{k\to\infty} \frac{|x_{k+1}|}{|x_k|^p} \\
&= \lim_{k\to\infty} \gamma^{k(1-p)+1} = \begin{cases} 0, & \text{for } p < 1 \\ \gamma, & \text{for } p = 1 \\ \infty, & \text{for } p > 1 \end{cases}
\end{aligned}$$

We get convergence for $p = 1$, so this sequence converges linearly with rate of convergence $\beta = \gamma$.

Convergence order: Examples

Example 2.4

Let $x_k = \gamma^{(q^k - \frac{1}{q-1})}$ for $0 < \gamma < 1$ and $q > 1$.

This sequence converges to $x^* = 0$. Now, we can compute the *rate of convergence*:

$$\begin{aligned}
\beta &= \lim_{k\to\infty} \frac{|x_{k+1} - x^*|}{|x_k - x^*|^p} \\
&= \lim_{k\to\infty} \frac{|x_{k+1}|}{|x_k|^p} \\
&= \lim_{k\to\infty} \gamma^{\frac{p-1}{q-1}} \gamma^{q^k(q-p)} = \begin{cases} 0, & \text{for } p < q \\ \gamma, & \text{for } p = q \\ \infty, & \text{for } p > q \end{cases}
\end{aligned}$$

We get convergence for $p = q$, so the order of convergence is 'q' with *rate of convergence* $\beta = \gamma$.

Convergence order

- The order of convergence can be interpreted using the notion of the order symbol O (big Oh).
- **Recall**: We say $f(x) = O(g(x))$ as $x \to a$ if there exists a constant C such that $|f(x)| \leq C|g(x)|$ in some neighborhood of a, that is, for $x \in (a - \delta, a + \delta) \setminus \{x\}$ for some $\delta > 0$.
- **Recall**: if $\lim_{x \to a} \frac{h(x)}{g(x)} = L$ then $h(x) = O(g(x))$.
- Then, the order of convergence is at least p if $\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\| = O(\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^p)$. (See next theorem)
- The order of convergence is at least 2 if $\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\| = O(\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2)$

Convergence order

Theorem 2.5

Let $\{x_k\}$ be a sequence that converges to x^*. If

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\| = O(\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^p)$$

then the order of convergence (if it exists) is at least p.

Let s be the order of convergence of $\{x_k\}$. On the other hand, there exists a content C such that

$$\begin{aligned} \|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\| &\leq C\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^p \\ \beta = \lim_{k\to\infty} \frac{\|\boldsymbol{x}_{k+1} - x^*\|}{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^s} &\leq C \lim_{k\to\infty} \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^{p-s} \end{aligned}$$

by definition of order of convergence $\beta > 0$

Convergence order

Theorem 2.6

Let $\{x_k\}$ be a sequence that converges to x^*. If

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\| = O(\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^p)$$

then the order of convergence (if it exists) is at least p.

$$\begin{aligned} 0 < \beta = \lim_{k\to\infty} \frac{\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|}{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^s} \quad &\leq \quad C \lim_{k\to\infty} \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^{p-s} \\ \lim_{k\to\infty} \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^{p-s} \quad &> \quad 0 \end{aligned}$$

As $\lim_{k\to\infty} \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^{p-s} = 0$ for $s < p$ then $s \geq p$. That is,the order of convergence is at least p.

Convergence order: Example

Example 2.7

Consider the problem of finding a minimizer of the function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = x^2 - \frac{1}{3}x^3$. Suppose we use the algorithm $x_{k+1} = x_k - \alpha f'(x_k)$ with step size $\alpha = 0.5$ and initial condition $x_0 = 1$.

We first show that the algorithm converges to a local minimizer of f. We have $f'(x) = 2x - x^2$. Therefore

$$x_{k+1} \quad = \quad x_k - \alpha f'(x_k) = 0.5(x_k)^2$$

with $x_0 = 1$ we obtain that $x_k = (1/2)^{2^k - 1}$ that converges to $x^* = 0$, note that $x^* = 0$ is a local minimizer of f.

$$\beta \quad = \quad \lim_{k \to \infty} \frac{|x_{k+1} - x^*|}{|x_k - x^*|^2} = \lim_{k \to \infty} \frac{x_{k+1}}{(x_k)^2} = 0.5$$

The order of convergence is 2 with rate of convergence 0.5

Let $x_0 = y_0 = 1$ and the following convergence order with convergence rate 'r'

- Order 1 (linear): $x_k = rx_{k-1}$

$$x_k = rx_{k-1} = r\ r\ r \cdots r\ x_0 = r^k$$

- Order 2 (quadratic) : $y_k = ry_{k-1}^2$

$$\begin{aligned} y_k &= ry_{k-1}^2 = r(ry_{k-2}^2)^2 = r^{2^0} r^{2^1} y_{k-2}^{2^2} = r^{2^0} r^{2^1} r^{2^2} \cdots r^{2^{k-1}} y_0^{2^k} \\ &= r^{\sum_{i=0}^{k-1} 2^i} = r^{2^k - 1} \end{aligned}$$

Then plot (k, x_k) and (k, y_k) in the same figure, for different values of r.