Numerical Optimization

A quick primer and recap

Melanie Ganz-Benjaminsen
Kenny Erleben
Department of Computer Science

University of Copenhagen

Taylor Series

If $f(x) : \mathbb{R} \mapsto \mathbb{R}$ then we know

$$f(x+p) = f(x) + f'(x)p + \frac{1}{2}f''(x)p^2 + \cdots$$

or written more compactly

$$f(x+p) = \sum_{n=0}^{\infty} \frac{f^{(n)}(x)}{n!} p^n$$

Taylor's (Formula) Theorem

By definition one has

$$f(x+p) = f(x) + f'(x)p + \frac{1}{2}f''(x)p^2 + \cdots + \frac{f^{(n)}(x)}{n!}p^n + R_n$$

where

$$R_n = \frac{f^{(n+1)}(x+tp)}{(n+1)!}p^{n+1}$$

for some $0 < t < 1$.

For $n = 0$ and $n = 1$

We get

$$f(x+p) = f(x) + \underbrace{f'(x+tp)p}_{R_0}$$

and

$$f(x+p) = f(x) + f'(x)p + \underbrace{\frac{f^{(2)}(x+tp)p^2}{2}}_{R_1}$$

Compare this with Theorem 2.1 in Nocedal and Wright (page 14).

Connection to Taylor Series for $n = 0$ case

The Mean Value Theorem (MVT) says

$$f'(x + tp) = \frac{f(x + p) - f(x)}{p}$$

So from the Taylor series, we have

$$f(x + p) - f(x) = f'(x)p + \frac{1}{2}f''(x)p^2 + \cdots$$

Use MVT

$$f(x + p) = f(x) + f'(x + tp)p$$

So Taylor's Theorem is really just MVT on Taylor Series.

The Fundamental Theorem of Calculus

It says

$$\int_a^b f'(x)dx = f(b) - f(a)$$

Now from Taylor's Theorem

$$f(x+p) = f(x) + f'(x+tp)p, \quad \text{for some } t \in (0..1)$$

$$\int_0^1 \left(f(x+p) - f(x)\right) dt = \int_0^1 f'(x+tp)pdt.$$

So Taylor's Theorem can be written as

$$f(x+p) = f(x) + \int_0^1 f'(x+tp)pdt$$

Definitions of Order Notation – big O

We write

$$f(x) \in \mathcal{O}(g(x))$$

if there exist $C > 0$ and x_0 such that

$$|f(x)| \leq C\,|g(x)|$$

holds for all $x > x_0$. This definition describes the growth rate as x grows.

Definitions of Order Notation – big O

Big O can be generalized to describe the behavior near some value x_0

$$|f(x)| \leq C\,|g(x)| \quad \text{as} \quad x \to x_0$$

Meaning that there exist there exist $C > 0$ and $\delta > 0$ such that

$$|f(x)| \leq C\,|g(x)| \quad \text{as} \quad |x - x_0| < \delta$$

Typically in finite difference approximations when we study how error terms scale we use $x_0 = 0$.

Generalization of big O

From the last generalized definition, we may write that

$$f(x) \in \mathcal{O}(g(x))$$

means

$$\lim_{x \to x_0} \sup \left| \frac{f(x)}{g(x)} \right| < \infty$$

Definitions of Order Notation – little o

We write

$$f(x) \in \mathbf{o}(g(x))$$

if

$$\lim_{x \to x_0} \frac{f(x)}{g(x)} = 0$$

Lesson Learned

big-O meaning "grows no faster than" (i.e. grows at the same rate or slower) and little-o meaning "grows strictly slower than"

Hence little-o is considered a stronger statement than big-O.

Using little-$\mathbf{o}$ in Taylor Series

We know

$$f(x+p) = f(x) + f'(x)p + \frac{1}{2}f''(x)p^2 + \cdots$$

Clearly

$$\lim_{p \to 0} \frac{\frac{1}{2}f''(x)p^2 + \cdots}{p} = 0$$

and by definition, we can write

$$f(x+p) = f(x) + f'(x)p + \mathbf{o}(p)$$

Intuition: this tells how $f(x+p)$ behaves when p gets smaller.

Using Big-$\mathcal{O}$ in Taylor Series

We know

$$f(x+p) = f(x) + f'(x)p + \frac{1}{2}f''(x)p^2 + \cdots$$

Use Taylor formula

$$f(x+p) = f(x) + f'(x)p + R_1$$

where $R_1 = \frac{1}{2}f''(x+tp)p^2$ for some $0 \leq t \leq 1$. Clearly

$$|R_1| \leq C\left|p^2\right|$$

where $C = \sup_t \left(\frac{1}{2}f''(x+tp)\right)$ and by definition we can write

$$f(x+p) = f(x) + f'(x)p + \mathcal{O}(p^2)$$

Intuition: this tells a loose upper bound on $f(x+p)$.

Notice the Difference

So we have

$$f(x+p) = f(x) + f'(x)p + \mathcal{O}(p^2)$$
$$f(x+p) = f(x) + f'(x)p + \mathbf{o}(p)$$

Which one is the better one?

Convergence Rate Definitions

Let $\varepsilon_k \in \mathbb{R}$ be the error in the k^{th} iteration then

Linear Convergence Rate

$$\lim_{k\to\infty} \frac{|\varepsilon_{k+1}|}{|\varepsilon_k|} = c$$

where $0 < c < 1$ is the convergence constant.

Super Linear Convergence Rate

$$\lim_{k\to\infty} \frac{|\varepsilon_{k+1}|}{|\varepsilon_k|} = 0$$

Quadratic Convergence Rate

$$\frac{|\varepsilon_{k+1}|}{|\varepsilon_k|^2} \leq M$$

for some $0 < M \in \mathbb{R}$ and for all k sufficiently large.

Linear Convergence Rate

Basically means that

$$\varepsilon_{k+1} \leq c\, \varepsilon_k$$

for some constant $0 < c < 1$. In the worst case, equality holds

$$\begin{aligned} \varepsilon_1 &= c\, \varepsilon_0 \\ \varepsilon_2 &= c\, \varepsilon_1 = c^2 \varepsilon_0 \\ &\vdots \\ \varepsilon_k &= c\, \varepsilon_{k-1} = c^k \varepsilon_0 \end{aligned}$$

Linear Convergence Rate

So for linear convergence rate, we have in the worst case,

$$\varepsilon_k = c^k \varepsilon_0 \,.$$

Now we take the logarithm,

$$\log \varepsilon_k = \log(c^k \varepsilon_0) = k \log(c) + \log(\varepsilon_0)$$

Hence in a log plot, we have a straight line with an intersection with y-axis given by $\log(\varepsilon_0)$ and slope $\log(c)$.

Quadratic Convergence Rate

Basically means that

$$\varepsilon_{k+1} \le M\varepsilon_k^2$$

for some constant $0 < M$. In worst case equality holds

$$\begin{aligned}
\varepsilon_1 &= M\varepsilon_0^2 \\
\varepsilon_2 &= M\varepsilon_1^2 = M^3\varepsilon_0^4 \\
\varepsilon_3 &= M\varepsilon_2^2 = M^7\varepsilon_0^8 \\
\vdots &\quad \vdots \\
\varepsilon_k &= M\varepsilon_{k-1}^2 = M^{2^k-1}\,\varepsilon_0^{2^k}
\end{aligned}$$

Quadratic Convergence Rate

So we have

$$\varepsilon_k = M^{2^k-1}\varepsilon_0^{2^k} .$$

Now we take the logarithm

$$\begin{aligned} \log \varepsilon_k &= \log(M^{2^k-1}\,\varepsilon_0^{2^k}) , \\ &= (2^k-1)\log(M) + 2^k \log(\varepsilon_0) , \\ &= 2^k \log(M\varepsilon_0) - \log M . \end{aligned}$$

Hence in a log plot, we have a decreasing power function if $\log(M\varepsilon_0) < 0$ this will occur if the initial error is sufficiently small.

Assignment

Given

$$f(x) = \begin{cases} e^{-\frac{1}{x^2}} & ; x \neq 0, \\ 0 & ; \text{otherwise}, \end{cases}$$

- Show that $f(x)$ is infinitely differentiable at $x = 0$
- Write up its Taylor series around $x = 0$
- Discuss if $f(x)$ is equal to its Taylor series around $x = 0$
- What requirements should one have to a function $f(x)$ in order for it to be equal to its Taylor series?

Assignment

- Find out for what functions $f(x) : [a..b] \mapsto \mathbb{R}$ that MVT holds for. Here we assume that $a, b \in \mathbb{R}$ and $a < b$.
- Discuss if these requirements are fulfilled for functions that are equal to their Taylor series

Assignment

- Try and differentiate $f(x + tp)$ wrt. t and use the result to apply the fundamental theorem of calculus to $\int_0^1 f'(x + tp)pdt$. What have you found?
- Discuss what equation 2.5 in Nocedal and Wright really is. Hint consider MVT and the Fundamental Theorem of Calculus.

Assignment

- Use formal definitions to show whether

$$x^2 \in \mathcal{O}(x^2)$$
$$x^2 \in \mathcal{O}(x^2 + x)$$
$$x^2 \in \mathcal{O}(200 * x^2)$$
$$x^2 \in \mathbf{o}(x^3)$$
$$x^2 \in \mathbf{o}(x!)$$
$$ln(x) \in \mathbf{o}(x)$$

- Assume $f(x) \in \mathbf{o}(g(x))$ does this imply $f(x) \in \mathcal{O}(g(x))$
- Assume $f(x) \in \mathcal{O}(g(x))$ discuss if this imply $f(x) \in \mathbf{o}(g(x))$

Assignment

- Prove or disprove $f(x) \in \mathcal{O}(x^2) \Rightarrow f(x) \in \mathbf{o}(x)$
- Prove or disprove $f(x) \in \mathbf{o}(x) \Rightarrow f(x) \in \mathcal{O}(x^2)$

Hint:

$$\lim_{x \to 0} \frac{f(x)}{x} = 0$$

Means that for any value $0 < \varepsilon \in \mathbb{R}$ there exist $0 < \delta \in \mathbb{R}$ such that for all x

$$|x| < \delta \quad \Rightarrow \quad \left| \frac{f(x)}{x} \right| < \varepsilon$$

Assignment

- Let $\varepsilon \in \mathbb{R}$ and $k \in \mathbb{N}_+$ then define the sequence $L \equiv \{\varepsilon_k\}_{k=1}^N$ for some sufficiently large $N > 1$ given by the recurrence relation $\varepsilon_{k+1} = \frac{1}{2}\varepsilon_k$ with $\varepsilon_1 = \frac{1}{10}$.
- Define $Q \equiv \{\varepsilon_k\}_{k=1}^N$ given by $\varepsilon_{k+1} = 10\varepsilon_k^2$ with $\varepsilon_1 = \frac{1}{10}$.
- Define $S \equiv \{\varepsilon_k\}_{k=1}^N$ given by $\varepsilon_{k+1} = c(k)\varepsilon_k^2$ with $\varepsilon_1 = \frac{1}{10}$ and $c(k) = e^{-\frac{1}{k^2}}$

Make plots of each sequence and discuss the shape of the plots. Can you identify which is which just by looking at their shapes?

Assignment

- Reconsider the sequences of error measures L, S, and Q. Prove/disprove for each whether it has a linear, super linear, or quadratic convergence rate. (Hint: use the formal definitions of convergence rates).
- Try to plot the sequences $\varepsilon_k \equiv 1 + \frac{1}{2}^k$, $\varepsilon_k \equiv 1 + k^{-k}$, and $\varepsilon_k \equiv 1 + \frac{1}{2}^{2^k}$. Determine which ones have linear, super linear, and quadratic convergence rates.

Assignment

- Try and take all previously listed sequences from previous slides and plot these in log plots.
- What can you observe from the plots?
- Consider what happens with the log plot of linear convergence rate if you let c be a decreasing function towards zero as a function of k. What kind of curve shape do you get?
- Try for quadratic convergence rate to plot the $\log \log \varepsilon_k$ as a function of k, what do you discover? (Extra prove that slope of $\log \log \varepsilon_k$ plot for quadratic converge is $\log 2$