Midway presentation

Dahyun Kang (20192702) from Group 7

CSED490Y: Optimization for machine learning
Department of Computer Science and Engineering

May 11, 2022

Overview

1. **Paper presentation**

2. **Term project progress**

A rapidly convergent descent method for minimization

R. Fletcher and M. J. D. Powell

The computer journal, 1963.

Preliminary: Newton's method

We are interested in finding the minimum of an unrestricted, twice-differentiable at all points, and convex function f:

$$\min_{x} f(x)$$

Gradient descent:

$$x_{t+1} = x_t - \eta \nabla f(x_t)$$

Newton's method:

$$x_{t+1} = x_t - G(x_t)^{-1} \nabla f(x_t),$$

where G is the second-order derivative.

Preliminary: Newton's method (continued)

Taylor series second-order approximation of f at a local point:

$$f(x) \approx f(x_t) + \nabla f(x_t)(x - x_t) + \frac{1}{2}\nabla^2 f(x_t)(x - x_t)^2$$

The minimum of $f(x)$ is found by setting its gradient to 0:

$$\begin{aligned} \nabla f(x) &= \nabla f(x_t) + \nabla^2 f(x_t)(x - x_t) \\ &= 0 \\ \Leftrightarrow x &= x_t - G(x_t)^{-1}\nabla f(x_t) \end{aligned}$$

This works because the second-order terms in the Taylor series expansion dominate near the minimum.

Quasi-Newton method

Computing $G(x_t)^{-1}$ is extremely expensive.

- Computing Hessian takes $\mathcal{O}(n^2)$.
- Matrix inverse takes $\mathcal{O}(n^3)$.

Solution: let us approximate $G^{-1}(x_t)$ iteratively.

- Let us denote the approximation as $H_t :\approx G^{-1}(x_t)$.
- $x_{t+1} = x_t - H_t \nabla f(x_t)$ for each t^{th} iteration
- Relevant method [Householder 1953] frequently fails to converge from a poor approximation to the minimum.

Secant equation for approximating Hessian

Recall the **first**-order derivative:

$$\frac{d}{dx}f(x) = \lim_{\triangle x \to 0} \frac{f(x + \triangle x) - f(x)}{x + \triangle x - x}$$

Approximating the **second**-order derivative of $G_t \approx \nabla^2 f(x_t)$:

$$G_{t+1} \approx \frac{\nabla f(x_{t+1}) - \nabla f(x_t)}{x_{t+1} - x_t}$$

$$G_{t+1}(x_{t+1} - x_t) \approx \nabla f(x_{t+1}) - \nabla f(x_t)$$

By setting $H_{t+1} := G_{t+1}^{-1}, s := x_{t+1} - x_t$ and $y := \nabla f(x_{t+1}) - \nabla f(x_t)$:

$$H_{t+1} y = s$$

Symmetric rank-1 update (Davidon) [1]

Assumption: H_{t+1} from H_t follows rank-1 update such as:

$$H_{t+1} = H_t + auu^\top,$$

where a is a scalar value and u is an arbitrary vector.
Combining the secant equation $H_{t+1}y = s$ and setting $u = \alpha(H_t y - s)$ leads to:

$$H_t y + a(\alpha(H_t y - s))(\alpha(H_t y - s))^\top y = s.$$

$$\Rightarrow H_{t+1} = H_t + \frac{(s - H_t y)(s - H_t y)^\top}{(s - H_t y)^\top y}$$

This update has following limitations:

- $(s - H_t y)^\top y \approx 0$ may fail to update.
- H_t is not guaranteed to be possitive semi-definite.

[1] Derivation taken from a lecture note, CMU [Javier Pẽna 2016]

Symmetric rank-2 update (Davidon-Fletcher-Powell)

Assumption: H_{t+1} from H_t follows rank-2 update such as:

$$
\begin{aligned}
H_{t+1} &= H_t + auu^\top + bvv^\top \\
H_{t+1}y &= H_t y + auu^\top y + bvv^\top y = s \\
\Leftrightarrow s - H_t y &= au^\top yu + bv^\top yv
\end{aligned}
$$

where a and b are scalar values and u and v are arbitrary vectors.
By setting $u := s$ and $v := H_t y$:

$$
H_{t+1} = H_t - \frac{H_t yy^\top H_t}{y^\top H_t y} + \frac{ss^\top}{y^\top s}
$$

Stability

H_t is positive definite $\rightarrow$ the convergence is stable.
Let z be an arbitrary vector.

$$\begin{aligned}
H_{t+1} &= H_t - \frac{H_t y y^\top H_t}{y^\top H_t y} + \frac{s s^\top}{y^\top s} \\
\Rightarrow z^\top H_{t+1} z &= z^\top H_t z - \frac{z^\top H_t y y^\top H_t z}{y^\top H_t y} + \frac{z^\top s s^\top z}{y^\top s} \\
&= \frac{p^\top p q^\top q - (p^\top q)^2}{q^\top q} + \frac{(s^\top z)^2}{y^\top s} \quad \text{where} \quad p = H_t^{1/2} z \ \text{ and } \ q = H_t^{1/2} y \\
&\geq \frac{(s^\top z)^2}{y^\top s} \\
&> 0
\end{aligned}$$

on account of Schwartz's inequality.

Experiment

Function (1):

- $f(x_1, x_2) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2$.
  - This function is difficult to minimize on account of its having a steep sided valley following $x_1^2 = x_2$.

Function (2):

- $f(x_1, x_2, x_3) = 100[x_3 - 10\theta(x_1, x_2)]^2 + [r(x_1, x_2) - 1]^2 + x_3^2$.
  - $2\pi\theta(x_1, x_2) = \begin{cases} \arctan(x_2/x_1) & \text{if } x_1 > 0, \\ \pi + \arctan(x_2/x_1) & \text{otherwise.} \end{cases}$
  - $r(x_1, x_2) = (x_1^2 + x_2^2)^{1/2}$
  - This function has a helical valley in the x_3 direction with pitch 10 and radius 1.

Experiment: function (1)

- $f(x_1, x_2) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2$.
- $\min f(x_1, x_2) = (1, 1)$

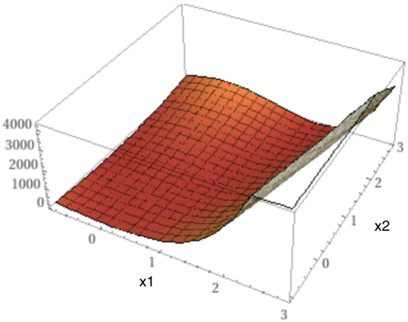

Table 1

A comparison in two dimensions

EQUIVALENT n	STEEPEST DESCENTS $f(x_1, x_2)$	POWELL'S METHOD $f(x_1, x_2)$	OUR METHOD $f(x_1, x_2)$
0	24·200	24·200	24·200
3	3·704	3·643	3·687
6	3·339	2·898	1·605
9	3·077	2·195	0·745
12	2·869	1·412	0·196
15	2·689	0·831	0·012
18	2·529	0·432	1×10^{-8}
21	2·383	0·182	—
24	2·247	0·052	—
27	2·118	0·004	—
30	1·994	5×10^{-5}	—
33	1·873	8×10^{-9}	—

Experiment: function (2)

- $f(x_1, x_2, x_3) = 100[x_3 - 10\theta(x_1, x_2)]^2 + [r(x_1, x_2) - 1]^2 + x_3^2$.
- $\min f(x_1, x_2, x_3) = (1, 0, 0)$

Table 3

A function with a steep-sided helical valley

n	x_1	x_2	x_3	f
0	−1·000	0·000	0·000	$2{\cdot}5 \times 10^4$
1	−1·000	2·278	1·431	$5{\cdot}2 \times 10^3$
2	−0·023	2·004	2·649	$1{\cdot}1 \times 10^3$
3	−0·856	1·559	3·429	74·080
4	−0·372	1·127	3·319	24·190
5	−0·499	0·908	3·285	10·942
6	−0·314	0·900	3·075	9·841
7	0·059	1·069	2·408	6·304
8	0·146	1·086	2·261	6·093
9	0·774	0·725	1·218	1·889
10	0·746	0·706	1·242	1·752
11	0·894	0·496	0·772	0·762
12	0·994	0·298	0·441	0·382
13	0·994	0·191	0·317	0·141
14	1·017	0·085	0·133	0·058
15	0·997	0·070	0·110	0·013
16	1·002	0·009	0·014	8×10^{-4}
17	1·000	0·002	0·040	3×10^{-6}
18	1·000	10^{-5}	10^{-5}	7×10^{-8}

Conclusion

Takeaway:

- This paper presents a Quasi-Newton method that iteratively approximates the inverse of Hessian using rank-2 update.

What I learned from reading this paper:

- Valuable experience of reading a classic paper
- Not easy to fully understand due to classical notations and unkind writing.
- Studied background of the second-order gradient methods.

Midterm progress

Analysis on second-order optimization method:
Newton's and Quasi-Newton method

The goal of the project

Understanding & in-depth analysis on second-order gradient methods.

Three representative second-order gradient methods that we chose are:

- Vanilla Newton's method
- A Quasi-Newton method (DFP [Fletcher and Powell 1963])
- A recent method (AdaHessian [Yao et al. 2021])

We will implement these methods and analyze them in two-variate convex functions.

Progress

- Studied the background of second-order methods.
- Implemented code skeleton

Plan

- ~~Apr. 14th - Apr. 30th : Survey & study~~
- May. 1st - May. 19th : Implement Newton's, Quasi-Newton, AdaHessian method
- May. 20th - May. 26th : Implement evaluation pipeline.
- May. 26th - Jun. 1st : Analysis
- Jun. 2nd - Jun. 4th : Final report & prepare for presentation

References

Javier Pẽna (2016)
Lecture note: Quasi-Newton Methods.
Statistics & Data Science, Carnegie Mellon university.

A. S. Householder (1953)
Principles of numerical analysis.
New York: McGraw-Hill.

R. Fletcher and M. J. D. Powell (1963)
A rapidly convergent descent method for minimization.
The computer journal.

Z Yao et al. (2021)
AdaHessian: An adaptive second order optimizer for machine learning.
Proceedings of the AAAI Conference on Artificial Intelligence.

Thank you

- Any questions?

Stability (continued)

H_t is positive definite $\rightarrow$ the convergence is stable.
Let z be an arbitrary vector.

$$
\begin{aligned}
H_{t+1} &= H_t - \frac{H_t y y^\top H_t}{y^\top H_t y} + \frac{s s^\top}{y^\top s} \\
&\geq \frac{(s^\top z)^2}{y^\top s}
\end{aligned}
$$

$$
\begin{aligned}
y^\top s &= (\nabla f(x_{t+1}) - \nabla f(x_t))^\top (x_{t+1} - x_t) \\
&= -\nabla f(x_t)^\top (x_{t+1} - x_t) \\
&= \nabla f(x_t)^\top H_t \nabla f(x_t) \\
&> 0
\end{aligned}
$$