

Optimization for Machine Learning

CS-439

Lecture 7: Newton's and Quasi-Newton Methods

Martin Jaggi

EPFL – github.com/epfml/OptML_course

April 19, 2024

Chapter 8

Newton's Method

1-dimensional case: Newton-Raphson method

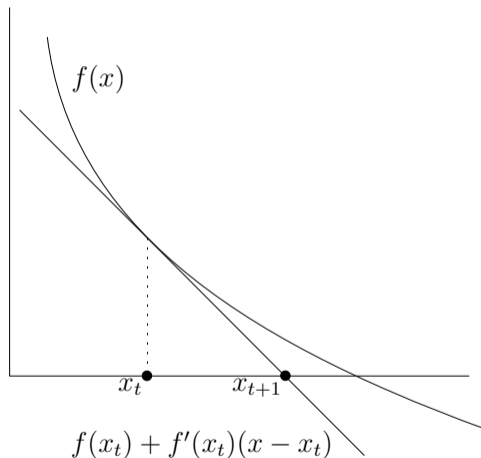
Goal: find a zero of differentiable
 $f : \mathbb{R} \rightarrow \mathbb{R}$.

Method:

$$x_{t+1} := x_t - \frac{f(x_t)}{f'(x_t)}, \quad t \geq 0.$$

x_{t+1} solves

$$f(x_t) + f'(x_t)(x - x_t) = 0,$$



The Babylonian method

Computing square roots: find a zero of $f(x) = x^2 - R, R \in \mathbb{R}_+$.

Newton-Raphson step:

$$x_{t+1} = x_t - \frac{f(x_t)}{f'(x_t)} = x_t - \frac{x_t^2 - R}{2x_t} = \frac{1}{2} \left(x_t + \frac{R}{x_t} \right).$$

Starting far (large $x_0 > 0$), we move slowly:

$$x_{t+1} = \frac{1}{2} \left(x_t + \frac{R}{x_t} \right) \geq \frac{x_t}{2}.$$

E.g., from $x_0 = R \geq 1$, it takes $\mathcal{O}(\log R)$ steps to get $x_t - \sqrt{R} < 1/2$ (Exercise 45).

The Babylonian method - Takeoff

Starting close, $x_0 - \sqrt{R} < 1/2$ (achievable after $\mathcal{O}(\log R)$ steps), things will speed up:

$$x_{t+1} - \sqrt{R} = \frac{1}{2} \left(x_t + \frac{R}{x_t} \right) - \sqrt{R} = \frac{x_t}{2} + \frac{R}{2x_t} - \sqrt{R} = \frac{1}{2x_t} \left(x_t - \sqrt{R} \right)^2.$$

Assume $R \geq 1/4$. Then all iterates have value at least $\sqrt{R} \geq 1/2$. Hence we get

$$x_{t+1} - \sqrt{R} \leq \left(x_t - \sqrt{R} \right)^2.$$

$$x_T - \sqrt{R} \leq \left(x_0 - \sqrt{R} \right)^{2^T} < \left(\frac{1}{2} \right)^{2^T}, \quad T \geq 0.$$

To get $x_T - \sqrt{R} < \varepsilon$, we only need $T = \log \log(\frac{1}{\varepsilon})$ steps!

The Babylonian method - Example

$R = 1000$, IEEE 754 double arithmetic

- ▶ 7 steps to get $x_7 - \sqrt{1000} < 1/2$
- ▶ 3 more steps to get x_{10} equal to $\sqrt{1000}$ up to machine precision (53 binary digits).
- ▶ First phase: \approx one more correct digit per iteration
- ▶ Last phase, \approx double the number of correct digits in each iteration!

Once you're close, you're there...

Newton's method for optimization

1-dimensional case: Find a global minimum x^* of a differentiable convex function $f : \mathbb{R} \rightarrow \mathbb{R}$.

Can equivalently search for a zero of the derivative f' : Apply the Newton-Raphson method to f' .

Update step:

$$x_{t+1} := x_t - \frac{f'(x_t)}{f''(x_t)} = x_t - f''(x_t)^{-1} f'(x_t)$$

(needs f twice differentiable).

d -dimensional case: Newton's method for minimizing a convex function $f : \mathbb{R}^d \rightarrow \mathbb{R}$:

$$\mathbf{x}_{t+1} := \mathbf{x}_t - \nabla^2 f(\mathbf{x}_t)^{-1} \nabla f(\mathbf{x}_t)$$

Newton's method = adaptive gradient descent

General update scheme:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - H(\mathbf{x}_t)\nabla f(\mathbf{x}_t),$$

where $H(\mathbf{x}) \in \mathbb{R}^{d \times d}$ is some matrix.

Newton's method: $H = \nabla^2 f(\mathbf{x}_t)^{-1}$.

Gradient descent: $H = \gamma I$.

Newton's method: “adaptive gradient descent”, adaptation is w.r.t. the local geometry of the function at \mathbf{x}_t .

Convergence in one step on quadratic functions

A **nondegenerate** quadratic function is a function of the form

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top M\mathbf{x} - \mathbf{q}^\top \mathbf{x} + c,$$

where $M \in \mathbb{R}^{d \times d}$ is an invertible symmetric matrix, $\mathbf{q} \in \mathbb{R}^d$, $c \in \mathbb{R}$. Let $\mathbf{x}^* = M^{-1}\mathbf{q}$ be the unique solution of $\nabla f(\mathbf{x}) = \mathbf{0}$.

► \mathbf{x}^* is the unique global minimum if f is convex.

Lemma

On nondegenerate quadratic functions, with any starting point $\mathbf{x}_0 \in \mathbb{R}^d$, Newton's method yields $\mathbf{x}_1 = \mathbf{x}^$.*

Proof.

We have $\nabla f(\mathbf{x}) = M\mathbf{x} - \mathbf{q}$ (this implies $\mathbf{x}^* = M^{-1}\mathbf{q}$) and $\nabla^2 f(\mathbf{x}) = M$. Hence,

$$\mathbf{x}_1 = \mathbf{x}_0 - \nabla^2 f(\mathbf{x}_0)^{-1} \nabla f(\mathbf{x}_0) = \mathbf{x}_0 - M^{-1}(M\mathbf{x}_0 - \mathbf{q}) = M^{-1}\mathbf{q} = \mathbf{x}^*.$$



Affine Invariance

Newton's method is **affine invariant**

(invariant under any invertible affine transformation):

Lemma (Exercise 46)

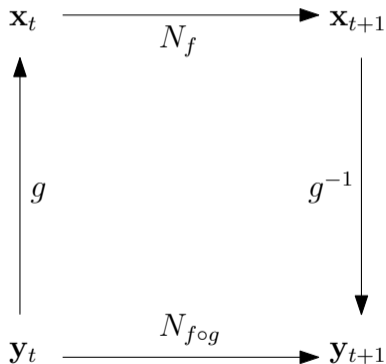
Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be twice differentiable, $A \in \mathbb{R}^{d \times d}$ an invertible matrix, $\mathbf{b} \in \mathbb{R}^d$. Let $g : \mathbb{R}^d \rightarrow \mathbb{R}$ be the (bijective) affine function $g(\mathbf{y}) = A\mathbf{y} + \mathbf{b}, \mathbf{y} \in \mathbb{R}^d$. Finally, for a twice differentiable function $h : \mathbb{R}^d \rightarrow \mathbb{R}$, let $N_h : \mathbb{R}^d \rightarrow \mathbb{R}^d$ denote the Newton step for h , i.e.

$$N_h(\mathbf{x}) := \mathbf{x} - \nabla^2 h(\mathbf{x})^{-1} \nabla h(\mathbf{x}),$$

whenever this is defined. Then we have $N_{f \circ g} = g^{-1} \circ N_f \circ g$.

Affine Invariance

Newton step for $f \circ g$ on \mathbf{y}_t : transform \mathbf{y}_t to $\mathbf{x}_t = g(\mathbf{y}_t)$, perform the Newton step for f on \mathbf{x} and transform the result \mathbf{x}_{t+1} back to $\mathbf{y}_{t+1} = g^{-1}(\mathbf{x}_{t+1})$. This means, the following diagram commutes:



Gradient descent suffers if coordinates are at different scales; Newton's method doesn't.

Minimizing the second-order Taylor approximation

Alternative interpretation of Newton's method:

Each step minimizes the local [second-order Taylor approximation](#).

Lemma (Exercise 49)

Let f be convex and twice differentiable at $\mathbf{x}_t \in \text{dom}(f)$, with $\nabla^2 f(\mathbf{x}_t) \succ 0$ being invertible. The vector \mathbf{x}_{t+1} resulting from the Newton step satisfies

$$\mathbf{x}_{t+1} = \underset{\mathbf{x} \in \mathbb{R}^d}{\operatorname{argmin}} f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top (\mathbf{x} - \mathbf{x}_t) + \frac{1}{2} (\mathbf{x} - \mathbf{x}_t)^\top \nabla^2 f(\mathbf{x}_t) (\mathbf{x} - \mathbf{x}_t).$$

Local Convergence

We will prove: under suitable conditions, and starting close to the global minimum, Newton's method will reach distance at most ε to the minimum within $\log \log(1/\varepsilon)$ steps.

- ▶ much faster than anything we have seen so far. . .
- ▶ . . . but we need to start close to the minimum already.

This is a **local convergence** result.

Global convergence results that hold for every starting point were unknown for Newton's method until very recently [KSJ18].

Once you're close, you're there...

Theorem

Let $f : \text{dom}(f) \rightarrow \mathbb{R}$ be twice differentiable with a critical point \mathbf{x}^* . Suppose there is a ball $X \subseteq \text{dom}(f)$ with center \mathbf{x}^* , s.t.

(i) *Bounded inverse Hessians:* There exists a real number $\mu > 0$ such that

$$\|\nabla^2 f(\mathbf{x})^{-1}\| \leq \frac{1}{\mu}, \quad \forall \mathbf{x} \in X.$$

(ii) *Lipschitz continuous Hessians:* There exists a real number $B \geq 0$ such that

$$\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\| \leq B\|\mathbf{x} - \mathbf{y}\| \quad \forall \mathbf{x}, \mathbf{y} \in X.$$

Then, for $\mathbf{x}_t \in X$ and \mathbf{x}_{t+1} resulting from the Newton step, we have

$$\|\mathbf{x}_{t+1} - \mathbf{x}^*\| \leq \frac{B}{2\mu} \|\mathbf{x}_t - \mathbf{x}^*\|^2.$$

Super-exponentially fast

Corollary (Exercise 47)

With the assumptions and terminology of the convergence theorem, and if

$$\|\mathbf{x}_0 - \mathbf{x}^*\| \leq \frac{\mu}{B},$$

then Newton's method yields

$$\|\mathbf{x}_T - \mathbf{x}^*\| \leq \frac{\mu}{B} \left(\frac{1}{2}\right)^{2^T - 1}, \quad T \geq 0.$$

Starting close to a critical point \mathbf{x}^* , we will reach distance at most ε to \mathbf{x}^* within $\mathcal{O}(\log \log(1/\varepsilon))$ steps.

Bound as for the last phase of the Babylonian method.

Super-exponentially fast — intuitive reason

Almost constant Hessians close to optimality...

...so f behaves almost like a quadratic function which has truly constant Hessians and allows Newton's method to convergence in one step.

Lemma (Exercise 48)

With the assumptions and terminology of the convergence theorem, and if $\mathbf{x}_0 \in X$ satisfies

$$\|\mathbf{x}_0 - \mathbf{x}^*\| \leq \frac{\mu}{B},$$

then the Hessians in Newton's method satisfy the relative error bound

$$\frac{\|\nabla^2 f(\mathbf{x}_t) - \nabla^2 f(\mathbf{x}^*)\|}{\|\nabla^2 f(\mathbf{x}^*)\|} \leq \left(\frac{1}{2}\right)^{2^t - 1}, \quad t \geq 0.$$

Proof of convergence theorem

We abbreviate $H := \nabla^2 f$, $\mathbf{x} = \mathbf{x}_t$, $\mathbf{x}' = \mathbf{x}_{t+1}$. Subtracting \mathbf{x}^* from both sides of the Newton step definition:

$$\begin{aligned}\mathbf{x}' - \mathbf{x}^* &= \mathbf{x} - \mathbf{x}^* - H(\mathbf{x})^{-1} \nabla f(\mathbf{x}) \\ &= \mathbf{x} - \mathbf{x}^* + H(\mathbf{x})^{-1} (\nabla f(\mathbf{x}^*) - \nabla f(\mathbf{x})) \\ &= \mathbf{x} - \mathbf{x}^* + H(\mathbf{x})^{-1} \int_0^1 H(\mathbf{x} + t(\mathbf{x}^* - \mathbf{x})) (\mathbf{x}^* - \mathbf{x}) dt,\end{aligned}$$

using the fundamental theorem of calculus

$$\int_a^b h'(t) dt = h(b) - h(a)$$

with

$$\begin{aligned}h(t) &= \nabla f(\mathbf{x} + t(\mathbf{x}^* - \mathbf{x})), \\ h'(t) &= \nabla^2 f(\mathbf{x} + t(\mathbf{x}^* - \mathbf{x})) (\mathbf{x}^* - \mathbf{x}).\end{aligned}$$

Proof of convergence theorem, II

We so far have

$$\mathbf{x}' - \mathbf{x}^* = \mathbf{x} - \mathbf{x}^* + H(\mathbf{x})^{-1} \int_0^1 H(\mathbf{x} + t(\mathbf{x}^* - \mathbf{x}))(\mathbf{x}^* - \mathbf{x})dt.$$

With

$$\mathbf{x} - \mathbf{x}^* = H(\mathbf{x})^{-1}H(\mathbf{x})(\mathbf{x} - \mathbf{x}^*) = H(\mathbf{x})^{-1} \int_0^1 -H(\mathbf{x})(\mathbf{x}^* - \mathbf{x})dt,$$

we further get

$$\mathbf{x}' - \mathbf{x}^* = H(\mathbf{x})^{-1} \int_0^1 (H(\mathbf{x} + t(\mathbf{x}^* - \mathbf{x})) - H(\mathbf{x}))(\mathbf{x}^* - \mathbf{x})dt.$$

Taking norms, we have

$$\|\mathbf{x}' - \mathbf{x}^*\| \leq \|H(\mathbf{x})^{-1}\| \cdot \left\| \int_0^1 (H(\mathbf{x} + t(\mathbf{x}^* - \mathbf{x})) - H(\mathbf{x}))(\mathbf{x}^* - \mathbf{x})dt \right\|,$$

because $\|A\mathbf{y}\| \leq \|A\| \cdot \|\mathbf{y}\|$ for any A, \mathbf{y} (by def. of spectral norm).

Proof of convergence theorem, III

We so far have

$$\begin{aligned}\|\mathbf{x}' - \mathbf{x}^*\| &\leq \|H(\mathbf{x})^{-1}\| \cdot \left\| \int_0^1 (H(\mathbf{x} + t(\mathbf{x}^* - \mathbf{x})) - H(\mathbf{x}))(\mathbf{x}^* - \mathbf{x}) dt \right\| \\ &\leq \|H(\mathbf{x})^{-1}\| \int_0^1 \|(H(\mathbf{x} + t(\mathbf{x}^* - \mathbf{x})) - H(\mathbf{x}))(\mathbf{x}^* - \mathbf{x})\| dt \quad (\text{Ex. 51}) \\ &\leq \|H(\mathbf{x})^{-1}\| \int_0^1 \|H(\mathbf{x} + t(\mathbf{x}^* - \mathbf{x})) - H(\mathbf{x})\| \cdot \|\mathbf{x}^* - \mathbf{x}\| dt \\ &= \|H(\mathbf{x})^{-1}\| \cdot \|\mathbf{x}^* - \mathbf{x}\| \int_0^1 \|H(\mathbf{x} + t(\mathbf{x}^* - \mathbf{x})) - H(\mathbf{x})\| dt.\end{aligned}$$

We can now use the properties (i) and (ii) (bounded inverse Hessians, Lipschitz continuous Hessians) to conclude that

$$\|\mathbf{x}' - \mathbf{x}^*\| \leq \frac{1}{\mu} \|\mathbf{x}^* - \mathbf{x}\| \int_0^1 B \|t(\mathbf{x}^* - \mathbf{x})\| dt = \frac{B}{\mu} \|\mathbf{x}^* - \mathbf{x}\|^2 \underbrace{\int_0^1 t dt}_{1/2} = \frac{B}{2\mu} \|\mathbf{x} - \mathbf{x}^*\|^2.$$

□

Strong convexity \Rightarrow Bounded inverse Hessians

One way to ensure bounded inverse Hessians is to require strong convexity over X .

Lemma (Exercise 52)

Let $f : \mathbf{dom}(f) \rightarrow \mathbb{R}$ be twice differentiable and strongly convex with parameter μ over an open convex subset $X \subseteq \mathbf{dom}(f)$ meaning that

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{\mu}{2} \|\mathbf{x} - \mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in X.$$

Then $\nabla^2 f(\mathbf{x})$ is invertible and $\|\nabla^2 f(\mathbf{x})^{-1}\| \leq 1/\mu$ for all $\mathbf{x} \in X$, where $\|\cdot\|$ is the spectral norm.

Downside of Newton's method

Computational bottleneck in each step:

- ▶ compute and invert the **Hessian matrix**
- ▶ or solve the linear system $\nabla^2 f(\mathbf{x}_t)\Delta\mathbf{x} = -\nabla f(\mathbf{x}_t)$ for the next step $\Delta\mathbf{x}$.

Matrix / system has size $d \times d$, taking up to $\mathcal{O}(d^3)$ time to invert / solve.

In many applications, d is large. . .

The secant method

Another iterative method for finding zeros in dimension 1

Start from Newton-Raphson step

$$x_{t+1} := x_t - \frac{f(x_t)}{f'(x_t)},$$

Use **finite difference approximation** of $f'(x_t)$:

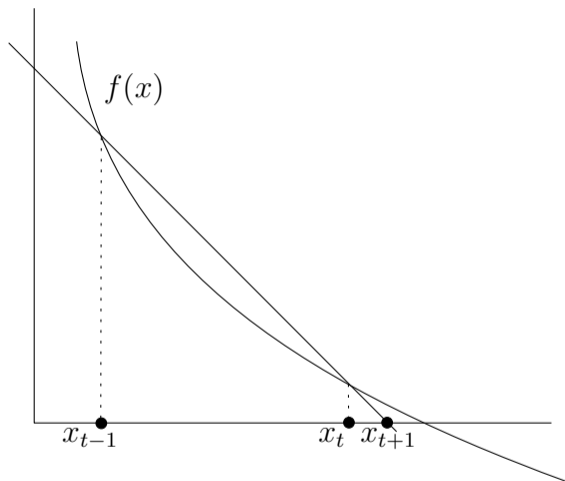
$$f'(x_t) \approx \frac{f(x_t) - f(x_{t-1})}{x_t - x_{t-1}}.$$

(for $|x_t - x_{t-1}|$ small)

Obtain the **secant method**:

$$x_{t+1} := x_t - f(x_t) \frac{x_t - x_{t-1}}{f(x_t) - f(x_{t-1})}$$

The secant method II



- ▶ construct the line through the two points $(x_{t-1}, f(x_{t-1}))$ and $(x_t, f(x_t))$;
- ▶ next iterate x_{t+1} is where this line intersects the x -axis (Exercise 53)

The secant method III

We now have a **derivative-free** version of the Newton-Raphson method.

Secant method for optimization: Can we also **optimize** a differentiable univariate function f ? — Yes, apply the secant method to f' :

$$x_{t+1} := x_t - f'(x_t) \frac{x_t - x_{t-1}}{f'(x_t) - f'(x_{t-1})}$$

- ▶ a **second-derivative-free** version of Newton's method for optimization.

Can we generalize this to higher dimensions to obtain a **Hessian-free** version of Newton's method on \mathbb{R}^d ?

The secant condition

Apply finite difference approximation to f'' (still 1-dim),

$$H_t := \frac{f'(x_t) - f'(x_{t-1})}{x_t - x_{t-1}} \approx f''(x_t)$$

\Leftrightarrow

$$f'(x_t) - f'(x_{t-1}) = H_t(x_t - x_{t-1}),$$

the [secant condition](#).

- ▶ Newton's method: $x_{t+1} := x_t - f''(x_t)^{-1} f'(x_t)$
- ▶ Secant method: $x_{t+1} := x_t - H_t^{-1} f'(x_t)$

In higher dimensions: Let $H_t \in \mathbb{R}^{d \times d}$ be a symmetric matrix satisfying the [d-dimensional secant condition](#)

$$\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1}) = H_t(\mathbf{x}_t - \mathbf{x}_{t-1}).$$

The secant method step then becomes

$$\mathbf{x}_{t+1} := \mathbf{x}_t - H_t^{-1} \nabla f(\mathbf{x}_t). \quad (1)$$

Quasi-Newton methods

$$\text{Newton: } \mathbf{x}_{t+1} := \mathbf{x}_t - \nabla^2 f(\mathbf{x}_t)^{-1} \nabla f(\mathbf{x}_t)$$

$$\text{Secant } \mathbf{x}_{t+1} := \mathbf{x}_t - H_t^{-1} \nabla f(\mathbf{x}_t), \text{ where } \nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1}) = H_t(\mathbf{x}_t - \mathbf{x}_{t-1})$$

If f is twice differentiable, secant condition and first-order approximation of $\nabla f(\mathbf{x})$ at \mathbf{x}_t yield:

$$\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1}) = H_t(\mathbf{x}_t - \mathbf{x}_{t-1}) \approx \nabla^2 f(\mathbf{x}_t)(\mathbf{x}_t - \mathbf{x}_{t-1}).$$

Might therefore hope that $H_t \approx \nabla^2 f(\mathbf{x}_t) \dots$

\dots meaning that the secant method approximates Newton's method.

- ▶ $d = 1$: unique number H_t satisfying the secant condition
- ▶ $d > 1$: Secant condition $\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1}) = H_t(\mathbf{x}_t - \mathbf{x}_{t-1})$ has infinitely many symmetric solutions H_t (underdetermined linear system).

Any scheme of choosing in each step of the secant method a **symmetric** H_t that satisfies the secant condition defines a **Quasi-Newton method**.

Quasi-Newton methods II

- ▶ Exercise 54: Newton's method is a Quasi-Newton method if and only if f is a nondegenerate quadratic function.
- ▶ Hence, Quasi-Newton methods do not generalize Newton's method but form a family of related algorithms.
- ▶ The first Quasi-Newton method was developed by William C. Davidon in 1956; he desperately needed iterations that were faster than those of Newton's method in order obtain results in the short time spans between expected failures of the room-sized computer that he used to run his computations on.
- ▶ But the paper he wrote about his new method got rejected for lacking a convergence analysis, and for allegedly dubious notation. It became a very influential Technical Report in 1959 [Dav59] and was finally officially published in 1991, with a foreword giving the historical context [Dav91]. Ironically, Quasi-Newton methods are today the methods of choice in a large number of optimization applications.
- ▶ Here: no convergence analysis (for a change), we focus on development of algorithms from first principles.

Developing a Quasi-Newton method

For efficiency reasons (want to avoid matrix inversions!), directly deal with the inverse matrices H_t^{-1} .

Given: iterates $\mathbf{x}_{t-1}, \mathbf{x}_t$ as well as the matrix H_{t-1}^{-1} .

Wanted: next matrix H_t^{-1} needed in next Quasi-Newton step

$$\mathbf{x}_{t+1} := \mathbf{x}_t - H_t^{-1} \nabla f(\mathbf{x}_t).$$

How should we choose H_t^{-1} ?

Newton's method: $\nabla^2 f(\mathbf{x}_t)$ fluctuates only very little in the region of extremely fast convergence.

Hence, in a Quasi-Newton method, it also makes sense to have that $H_t \approx H_{t-1}$, or $H_t^{-1} \approx H_{t-1}^{-1}$.

Greenstadt's family of Quasi-Newton methods

Given: iterates $\mathbf{x}_{t-1}, \mathbf{x}_t$ as well as the matrix H_{t-1}^{-1} .

Wanted: next matrix H_t^{-1} needed in next Quasi-Newton step

$$\mathbf{x}_{t+1} := \mathbf{x}_t - H_t^{-1} \nabla f(\mathbf{x}_t).$$

Greenstadt [Gre70]: Update

$$H_t^{-1} := H_{t-1}^{-1} + E_t,$$

E_t an error matrix.

Try to minimize the error subject to H_t satisfying the secant condition!

Simple error measure: Frobenius norm

$$\|E\|_F^2 := \sum_{i=1}^d \sum_{j=1}^d E_{ij}^2.$$

Greenstadt's family of Quasi-Newton methods II

Greenstadt: minimizing $\|E\|_F$ gives just one method, this is “too specialized”.

Greenstadt searched for a compromise between variability in the method and simplicity of the resulting formulas.

More general error measure

$$\|AEA^\top\|_F^2,$$

where $A \in \mathbb{R}^{d \times d}$ is some fixed invertible transformation matrix.

$A = I$: squared Frobenius norm of E , the “specialized” method.

The Greenstadt Update $H_{t-1}^{-1} \rightarrow H_t^{-1}$

Secant condition in terms of H_t^{-1} :

$$H_t^{-1}(\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})) = (\mathbf{x}_t - \mathbf{x}_{t-1}).$$

Fix t and simplify notation:

H	$:= H_{t-1}^{-1}$	(old inverse)
H'	$:= H_t^{-1}$	(new inverse)
E	$:= E_t,$	(error matrix)
$\boldsymbol{\sigma}$	$:= \mathbf{x}_t - \mathbf{x}_{t-1}$	(step in solutions)
\mathbf{y}	$= \nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})$	(step in gradients)
\mathbf{r}	$= \boldsymbol{\sigma} - Hy$	(error of old inverse in secant condition)

The update formula is

$$H' = H + E,$$

Secant condition becomes

$$H'\mathbf{y} = \boldsymbol{\sigma} \quad (\Leftrightarrow E\mathbf{y} = \mathbf{r}).$$

The Greenstadt Update $H_{t-1}^{-1} \rightarrow H_t^{-1}$ II

Minimizing the error becomes a convex constrained minimization problem in the d^2 variables E_{ij} :





$$\begin{aligned} \text{minimize} \quad & \frac{1}{2} \|AEA^\top\|_F^2 && \text{(error function)} \\ \text{subject to} \quad & E\mathbf{y} = \mathbf{r} && \text{(secant condition)} \\ & E^\top - E = 0 && \text{(symmetry)} \end{aligned}$$

Don't need to solve it computationally (for numbers E_{ij}) ...

... but mathematically (formula for E)

Minimize **convex quadratic** function subject to **linear equations** \rightarrow analytic formula for the minimizer from the **method of Lagrange multipliers**.

Bibliography

-  William C. Davidon.
Variable metric method for minimization.
Technical Report ANL-5990, AEC Research and Development, 1959.
-  William C. Davidon.
Variable metric method for minimization.
SIAM J. Optimization, 1(1):1–17, 1991.
-  J. Greenstadt.
Variations on variable-metric methods.
Mathematics of Computation, 24(109):1–22, 1970.
-  Sai Praneeth Karimireddy, Sebastian U Stich, and Martin Jaggi.
Global linear convergence of Newton's method without strong-convexity or Lipschitz gradients.
arXiv, 2018.