Optimization for Machine Learning
CS-439

Lecture 8: Newton & Quasi-Newton

Martin Jaggi

EPFL – github.com/epfml/OptML_course

April 12, 2019

Affine Invariance

Newton's method is **affine invariant**
(invariant under any invertible affine transformation):

Lemma (Exercise 41)

Let $f : \mathbb{R}^d \to \mathbb{R}$ *be twice differentiable,* $A \in \mathbb{R}^{d\times d}$ *an invertible matrix,* $\mathbf{b} \in \mathbb{R}^d$*. Let* $g : \mathbb{R}^d \to \mathbb{R}$ *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* $h : \mathbb{R}^d \to \mathbb{R}$*, let* $N_h : \mathbb{R}^d \to \mathbb{R}^d$ *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:

$$\begin{tikzcd}
\mathbf{x}_t \arrow[r, "N_f"'] & \mathbf{x}_{t+1} \arrow[d, "g^{-1}"'] \\
\mathbf{y}_t \arrow[u, "g"'] \arrow[r, "N_{f \circ g}"] & \mathbf{y}_{t+1}
\end{tikzcd}$$

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 44)

Let f be convex and twice differentiable at $\mathbf{x}_t \in \mathbf{dom}(f)$, with $\nabla^2 f(\mathbf{x}_t) \succ 0$ being invertible. The vector $\mathbf{x}_{t+1}$ resulting from the Netwon 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 are unknown for Newton's method.

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

Theorem

Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be convex with a unique global minimum $\mathbf{x}^\star$. Suppose there is a ball $X \subseteq \mathbf{dom}(f)$ with center $\mathbf{x}^\star$, 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 $B > 0$ 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

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

Super-exponentially fast

Corollary (Exercise 42)

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

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

then Newton's method yields

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

Starting close to the global minimum, we will reach distance at most ε to the minimum within $\mathcal{O}\big(\log\log(1/\varepsilon)\big)$ 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 43)

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

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

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

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

Proof of convergence theorem

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

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

using the fundamental theorem of calculus

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

with

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

Proof of convergence theorem, II

We so far have

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

With

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

we further get

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

Taking norms, we have

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

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

$$
\begin{aligned}
\|\mathbf{x}' - \mathbf{x}^\star\| &\leq \|H(\mathbf{x})^{-1}\| \cdot \left\| \int_0^1 \big(H(\mathbf{x} + t(\mathbf{x}^\star - \mathbf{x})) - H(\mathbf{x})\big)(\mathbf{x}^\star - \mathbf{x}) dt \right\| \\
&\leq \|H(\mathbf{x})^{-1}\| \int_0^1 \left\| \big(H(\mathbf{x} + t(\mathbf{x}^\star - \mathbf{x})) - H(\mathbf{x})\big)(\mathbf{x}^\star - \mathbf{x}) \right\| dt \quad \text{(Ex. 46)} \\
&\leq \|H(\mathbf{x})^{-1}\| \int_0^1 \left\| H(\mathbf{x} + t(\mathbf{x}^\star - \mathbf{x})) - H(\mathbf{x}) \right\| \cdot \|\mathbf{x}^\star - \mathbf{x}\| dt \\
&= \|H(\mathbf{x})^{-1}\| \cdot \|\mathbf{x}^\star - \mathbf{x}\| \int_0^1 \left\| H(\mathbf{x} + t(\mathbf{x}^\star - \mathbf{x})) - H(\mathbf{x}) \right\| dt.
\end{aligned}
$$

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

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

$\square$

Strong convexity $\Rightarrow$ Bounded inverse Hessians

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

Lemma (Exercise 47)

Let $f : \mathbf{dom}(f) \to \mathbb{R}$ 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 $f'(x_t)$:

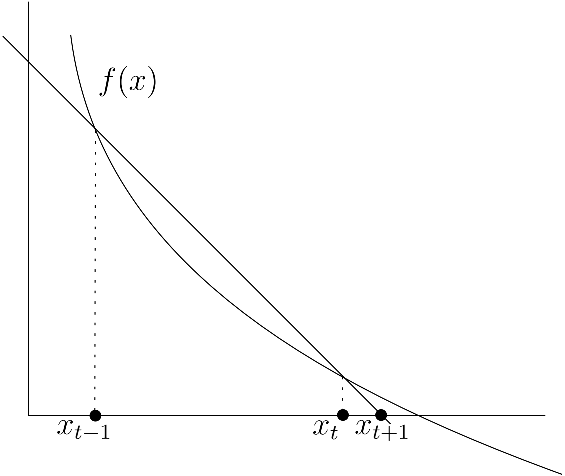
$$f'(x_t) \approx \frac{f(x_t) - f(x_{t-1})}{x_t - x_{t-1}}.$$

(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 48)

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

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

Quasi-Newton methods

Newton: $\mathbf{x}_{t+1} := \mathbf{x}_t - \nabla^2 f(\mathbf{x}_t)^{-1} \nabla f(\mathbf{x}_t)$
Secant $\mathbf{x}_{t+1} := \mathbf{x}_t - H_t^{-1} \nabla f(\mathbf{x}_t)$, 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 $H_t \approx \nabla^2 f(\mathbf{x}_t)\ldots$
...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 49: 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 number of relevant machine learning applications.
- Here: no convergence analysis (for a change), we focus on development of algorithms from first principles.

Developing a Quasi-Newton method

For efficieny 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: $\nabla f^2(\mathbf{x}_t)$ 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 errror 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.

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.