Optimization for Machine Learning
CS-439

Lecture 7: Newton and Quasi-Newton

Martin Jaggi

EPFL – github.com/epfml/OptML_course

April 20, 2018

Affine Invariance

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

Lemma (Exercise 27)

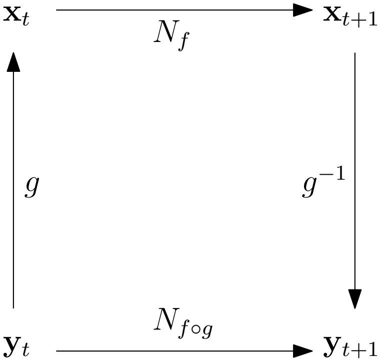
Let $f : \mathbb{R}^d \to \mathbb{R}$ *be twice differentiable,* $A \in \mathbb{R}^{d\times d}$ *an invertible matrix,* $\mathbf{b} \in \mathbb{R}^d$*. Let* $g : \mathbb{R}^d \to \mathbb{R}$ *be the (bijective) affine function* $g(\mathbf{y}) = A\mathbf{y} + \mathbf{b}, \mathbf{y} \in \mathbb{R}^d$*. Finally, let* $N_h : \mathbb{R}^d \to \mathbb{R}^d$ *denote the Newton step for function* 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 $\mathbf{y}_t$: can 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 $\mathbf{y}_{t+1} = g^{-1}(\mathbf{x}_{t+1})$. I.e., the following diagram commutes:



Hence, while gradient descent suffers if the coordinates are at very different scales, Newton's method doesn't.

Affine Invariance

Invariance to scaling of the input problem

Minimizing the second-order Taylor approximation

Alternative interpretation of Newton's method:
Each step minimizes the local second-order Taylor approximation.

Lemma (Exercise 30)

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

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

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 an open 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 $L > 0$ such that*
$$\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\| \quad \forall \mathbf{x}, \mathbf{y} \in X.$$

Matrix norm is spectral norm. Note: (i) $\Rightarrow$ Hessian invertible at all $\mathbf{x} \in X$.

Then, for $\mathbf{x}_t \in X$ and $\mathbf{x}_{t+1}$ resulting from the Newton step, we have
$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\| \leq \frac{L}{2\mu}\|\mathbf{x}_t - \mathbf{x}^\star\|^2.$$

Super-exponentially fast?

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.

Corollary (Exercise 28)

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

$$\|\mathbf{x}_0 - \mathbf{x}^\star\| < \frac{\mu}{L},$$

then Newton's method yields

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

Proof of convergence theorem

Lemma (Exercise 29)

Let f be twice differentiable over a convex domain $\mathbf{dom}(f)$, $\mathbf{x}, \mathbf{y} \in \mathbf{dom}(f)$. Then

$$\int_0^1 \nabla^2 f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))(\mathbf{y} - \mathbf{x})dt = \nabla f(\mathbf{y}) - \nabla f(\mathbf{x}).$$

Proof of Thm. 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 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 previous Lemma.

Proof of convergence theorem, II

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

Also,
$$\left\|\int_0^1 \mathbf{g}(t)dt\right\| \leq \int_0^1 \|\mathbf{g}(t)\|dt$$
for any vector-valued function $\mathbf{g}$ (Exercise 32) , so we can bound

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

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

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

□

Strong convexity?

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

Lemma (Exercise 33)

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.*

Chapter 7

Quasi-Newton Methods

Downside of Newton's method

Computational bottleneck in each step:

- compute and invert the Hessian matrix.

Matrix has size $d \times d$, taking up to $\mathcal{O}(d^3)$ time to invert
— or to solve the linear system $\nabla^2 f(\mathbf{x}_t)\Delta\mathbf{x} = -\nabla f(\mathbf{x}_t)$ for $\Delta\mathbf{x}$.

The secant method

Back to 1-dim.

Another iterative methods for finding zeros?

Newton-Raphson step

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

Lazy: use finite difference approximation

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

(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

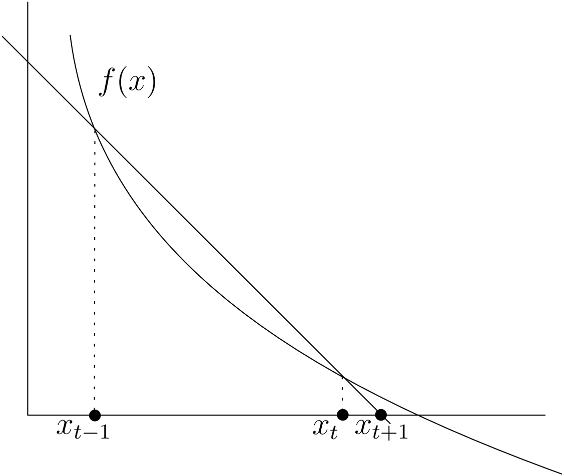


Figure: One step of the secant method

The secant method III

Why? now have a derivative-free version of Newton's 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 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

Applying 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 Newton step then becomes

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

Quasi-Newton methods

If f is twice differentiable, join the secant condition along with the first-order Taylor approximation of $\nabla f(\mathbf{x})$:

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

$\Rightarrow$ (QN) approximates Newton's method.

Quasi-Newton method: Whenever (QN) is used with a symmetric matrix satisfying the secant condition.

- How to find good H_t^{-1} matrices?
  BFGS, L-BFGS, etc.
- Newton's method is a Quasi-Newton method if and only if f is a nondegenerate quadratic function (Exercise 35). Hence, Quasi-Newton methods do not generalize Newton's method but form a family of related algorithms.