Lecture 16

Last time
- Analysis continued
- Convex guarantees
- Extensions

Today
- What's to come
- One-dimensional Newton's method.
- Newton's in \( \mathbb{R}^d \).

What's to come? Winter
Second-order Methods
- Quasi-Newton Methods.
- Conjugate gradient.
- Trust Region Methods.

Newton's Method
Imagine we had a system of nonlinear equations

\[ F(x) = 0 \]
with \( F: \mathbb{R}^d \to \mathbb{R}^d \) and we want to solve for \( x \). This recovers finding stationary points if \( F = \nabla f \).

**One-dimensional setup**

Assume \( F: \mathbb{R} \to \mathbb{R} \) is smooth.

The idea of Newton's method is to linearize and then look for a root (zero).

Thus, we update

Pick \( x_{k+1} \) s.t.

\[
F(x_k) + F'(x_k)(x_{k+1} - x_k) = 0
\]

Note that reordering this amounts to

\[
x_{k+1} = x_k - \frac{F(x_k)}{F'(x_k)}
\]

If \( F = f' \), then this is

\[
x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}
\]

we need second order information
If \( f''(x_k) > 0 \) this also corresponds
\[
\chi_{k+1} = \arg\min_x \left\{ f(x_k) + f'(x_k)(x-x_k) + \frac{f''(x_k)}{2}(x-x_k)^2 \right\}
\]

When \( f''(x_k) < 0 \) we don't have a convex model.

This method is really fast. As an example: Consider \( F(x) = x^2 - a \), then
\[
F(x) = 0 \iff x = \pm \sqrt{a}.
\]

In this case, Newton's method reduces to
\[
\chi_{k+1} = \chi_k - \frac{F(x_k)}{F'(x_k)} = \chi_k - \frac{x_k^2 - a}{2x_k} = \frac{1}{2} (x_k - \frac{a}{x_k}).
\]

For \( a = 2 \) and \( x = 1 \), we obtain
\[
\chi_0 = 1 \ldots \\
\chi_1 = 1.5 \ldots \\
\chi_2 = 1.41 \ldots \\
\chi_3 = 1.41421 \ldots \\
\chi_4 = 1.41421356237 \ldots
\]
Aside: This algorithm was used in the video game *Quake 3* (1998) to find $1/\sqrt{x}$.

Quick review of convergence naming

Suppose $s_k \to 0$ (This could be the objective gap, the distance to a solution or $\|\nabla f(x_k)\|_2$).

We say that

- $s_k$ converges linearly if $\exists c \in (0, 1)$, $N \geq 0$ s.t.
  \[ \forall k \geq N \quad s_{k+1} \leq c s_k \]

- $s_k$ converges sublinearly if no such $c$ exists.

- $s_k$ converges superlinearly if $\exists c_n \in (0, 1)$, $N \geq 0$ s.t. $c_k \to 0$ and $\forall k \geq N$,
  \[ s_{k+1} \leq c_k s_k \]

- $s_k$ converges quadratically if $\exists c \in (0, 1)$, $N \geq 0$ s.t.
  \[ \forall k \geq N \quad s_{k+1} \leq c s_k \]
  This is super linear since $c s_k \to 0$.

Secant Method

If we don't know $F'(x_k)$ it is reasonable to approximate it with
\[ E'(x_k) \approx \frac{F(x_k) - F(x_{k-1})}{x_k - x_{k-1}} \]

and thus
\[ x_{k+1} = x_k - \left( \frac{x_k - x_{k-1}}{F(x_k) - F(x_{k-1})} \right) F(x_k). \]

Under modest regularity conditions, we have \( x_k \to x^* \) with \( F(x^*) = 0 \). Moreover
\[ |x_{k+1} - x^*| \leq c \cdot |x_k - x^*|^\eta \]

where \( \eta = \frac{\sqrt{5} - 1}{2} \approx 1.618 \ldots \) is the Golden ratio. Thus convergence is superlinear but not quadratic.

**Newton in \( \mathbb{R}^d \)**

In a bunch of applications we want to solve systems of equations
\[ F(x) = 0 \quad \text{with} \quad F: \mathbb{R}^d \to \mathbb{R}^d \text{ smooth.} \]

For example:
- Optimization \( \nabla f(x) = 0. \)
Key idea: Linearize $F(X)$, then solve linear system.

Recall that the Jacobian of $F(x)$ is

$$
\nabla F(x) = \begin{bmatrix}
\frac{\partial F_1(x)}{\partial x_1} & \cdots & \frac{\partial F_i(x)}{\partial x_d} \\
\vdots & \ddots & \vdots \\
\frac{\partial F_d(x)}{\partial x_1} & \cdots & \frac{\partial F_d(x)}{\partial x_d}
\end{bmatrix}.
$$

When $f: \mathbb{R}^d \to \mathbb{R}$ is $C^2$, the Hessian $\nabla^2 f$ is the Jacobian of $\nabla f$.

Then, Newton's method updates by finding $x_{k+1}$ s.t.

$$
F(x_k) + \nabla F(x_k)(x_{k+1} - x_k) = 0.
$$

If $\nabla F(x_{k+1})$ is full rank, then the system has a unique solution and Newton's direction

$$
x_{k+1} = x_k - \nabla F(x_k)^{-1} F(x_k).
$$
In optimization land this is equivalent to
\[ x_{k+1} = x_k - \nabla^2 f(x_k) \nabla f(x_k). \]

Notice that this is equivalent to constructing a second-order approximation of \( f \) at \( x_k \):

\[ f_k(x) = f(x_k) + \nabla f(x_k)^T (x - x_k) + \frac{1}{2} (x - x_k)^T \nabla^2 f(x_k) (x - x_k) \]

and finding a critical point of \( f_k \).

Unlike before we don't have that \( f_k \) is convex:

- If \( \nabla^2 f(x_k) < 0 \) \( \Rightarrow \) \( f_k \) is concave
  - Ascent direction.
- If \( \nabla^2 f(x_k) > 0 \) \( \Rightarrow \) \( f_k \) is convex
  - Descent direction
- If \( \nabla^2 f(x_k) \) is indefinite \( \Rightarrow \) \( f_k \) has a saddle.

**Convergence of Newton's method**

We state the following without a proof, but we'll get back to a
nonasymptotic version next class

**Theorem (Local convergence)**

Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be continuously differentiable and assume $F(x^*) = 0$ for some $x^*$. If $\nabla F(x^*)$ is nonsingular, then some neighborhood $S$ of $x^*$ we have that if $x_0 \in S$, the iterates of Newton's method satisfy $x_k \in S$, $x_k \to x^*$, $\nabla F(x_k)$ nonsingular.

**Warnings**

- Global convergence is not granted.
- If $\nabla F(x_k)$ is singular the method is not well-defined.
- Even if $\nabla F(x_k)$ is nonsingular, we can have numerical issues

$$F(x) = \exp\left(-\frac{1}{x^2}\right)$$

Iterates do not move much.