Lecture 18

Last time	Today
▷ Convergence guarantee	▷ Exam results.
▷ Computational complexity	▷ Modified Newton
▷ Quasi-Newton intro.	▷ 3 variants

New idea from last class

Instead of using Taylor's approximation, consider

$$m_k(x) = f_k + g_k^T(x - x_k) + \frac{1}{2}(x - x_k)^T B_k (x - x_k)$$

Thus, a natural strategy is to consider

x_{k+1} is such that $\nabla m_k(x_{k+1}) = 0.$

which in turn reduces to

$$x_{k+1} = x_k - \overbrace{B_k^{-1} g_k}^{p_k}.$$

when B_k is invertible.

Natural questions:

▷ How do we pick B_k so that we have descent?

▷ Can we make it cheaper per-iteration?

We will focus on the first question in this lecture.

Let's look at the geometry of a Newton step.

$\nabla^2 f(x_k)$ is a symmetric, real matrix (and let's assume nonsingular).
We might take an spectral decomposition:

$$\nabla^2 f(x_k) = V \underset{\text{Diagonal.}}{\Lambda} \underset{\text{orthogonal}}{V^T} \quad \leftarrow \text{cost } O(d^3).$$

$$\Lambda = \underbrace{\begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_d \end{pmatrix}}_{\text{Eigenvalues}} = \begin{pmatrix} \Lambda_+ & \\ & \Lambda_- \end{pmatrix}$$

$$V = \underbrace{\begin{pmatrix} | & & | \\ v_1 & \cdots & v_d \\ | & & | \end{pmatrix}}_{\text{Eigenvectors}} = \begin{pmatrix} V_+ & V_- \end{pmatrix}$$

Now we can decompose the Newton step:

$$p_k = -(V \Lambda V^T)^{-1} \nabla f(x_k)$$

$$= -V \Lambda^{-1} V^T \nabla f(x_k)$$

$$= -\begin{pmatrix} V_+ & V_- \end{pmatrix} \begin{pmatrix} \Lambda_+^{-1} & \\ & \Lambda_-^{-1} \end{pmatrix} \begin{bmatrix} V_+^T \nabla f(x_k) \\ V_-^T \nabla f(x_k) \end{bmatrix}$$

invert diagonals

$$= \underbrace{-V_+ \Lambda_+^{-1} V_+^T \nabla f(x_k)}_{p_k^+} \underbrace{- V_- \Lambda_-^{-1} V_- \nabla f(x_k)}_{p_k^-}$$

Claim: p_k^+ is a "descent" direction $(\nabla f(x_k)^T p_k^+ \le 0)$.

We can easily check

$$\nabla f(x)^T p_k^+ = -\nabla f(x_k)^T V_+ \Lambda_+^{-1} V_+^T \nabla f(x_k) \nabla f(x_k) \le 0.$$

Symmetrically p_k^- satisfies $\nabla f(x_k)^T p_k^- \ge 0$.

Thus if all eigenvalues are positive $\Rightarrow$ Descent
all eigenvalues are negative $\Rightarrow$ Ascent
mixture $\Rightarrow$ Could do anything.

Lemma: If $B_k \succ 0$, and $g_k \neq 0$, then $p_k = \arg\min_p \{g_k^T p + p^T B_k p\}$

$\Rightarrow$ $g_k^T p_k < 0$.

In particular, if $g_k = \nabla f(x_k)$, then p_k is a descent direction.

Proof: Since B_k is positive definite, then $p \mapsto g_k^T p + p^T B_k p$ is strongly convex, then p_k is well-defined. Then $p_k = -B_k g_k$, thus

$$g_k^T p_k = -g_k^T B_k g_k < 0 \qquad \square$$

Warning: This doesn't guarantee that we have $f(x_{k+1}) \le f(x_k)$ via

$$x_{k+1} \leftarrow x_k - B_k^{-1} \nabla f(x_k).$$

We only have

$$f(x_k + \alpha p_k) = f(x_k) + \underbrace{\alpha \nabla f(x_k)^T p_k}_{<0} + o(\alpha^2).$$

Thus we need an stepsize!

Linesearch could we appied. The Armijo condition reduces to: for some $\eta \in (0,1)$

$$f(x_k - \alpha_k p_k) \le f(x_k) + \eta \alpha_k g_k^T p_k$$

with α_k exponentially shrinking until this holds.

Modified Newton's Method

Consider the following template

Loop $k = 0, 1, \dots$

Compute $\nabla f(x_k)$ and $\nabla^2 f(x_k)$

3 methods today. $\to$ Build $B_k \succ 0$ (Based on $\nabla^2 f(x_k)$)

Compute $p_k \leftarrow B_k^{-1} \nabla f(x_k)$

Pick α_k ensuring descent (Armijo)

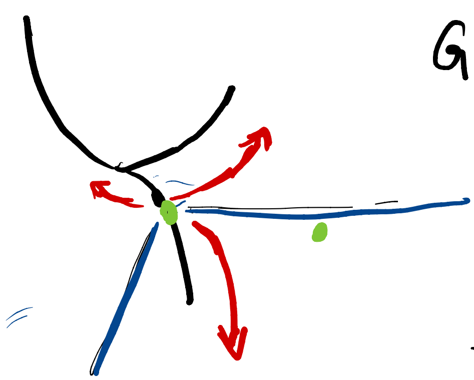
HW 5 you'll prove constant stepsizes also work.

$x_{k+1} \leftarrow x_k + p_k$

End loop.

▷ Option 1

Discard nonpositive eigenvalues



Get the factorization

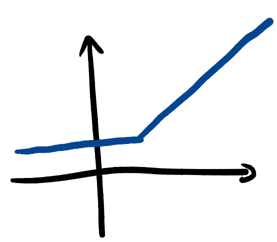
$$\nabla^2 f(x_k) = V \Lambda V^T$$

Define

$$\bar{\Lambda} = \operatorname{diag}(\bar{\lambda}_i) \quad \text{with}$$

$$\bar{\lambda}_i = \max\{\lambda_i, \varepsilon\}$$

$\varepsilon \geq 0$



Then take

$$B_k = V \bar{\Lambda} V^T.$$

The downside is that we loose the "mag

nitude" of the negative λ_i.
We move little when $\nabla f(x_k)$ is aligned with negative components.
Pretty bad unless $\nabla^2 f(x_k) \succeq \varepsilon I$, in which case was good too.

Option 2

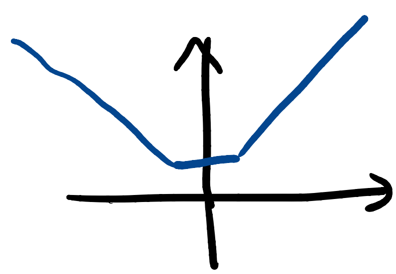
Keep eigenvalues with large magnitude, but make them positive

$$\nabla f(x_k) = V \Lambda V^T$$

Pick $\varepsilon > 0$ and set

$$\bar{\Lambda} = \operatorname{diag}(\bar{\lambda}_i) \quad \text{where} \quad \bar{\lambda}_i = \max\{|\lambda_i|, \varepsilon\}$$

$$B_k = V \bar{\Lambda} V^T.$$



$$\Rightarrow p_k = -B_k^{-1} \nabla f(x_k)$$

$$= -\left((V_+\ V_\varepsilon\ V_-) \begin{pmatrix} \Lambda_+ & & \\ & \varepsilon I & \\ & & -\Lambda_- \end{pmatrix} \begin{pmatrix} V_+^T \\ V_\varepsilon^T \\ V_-^T \end{pmatrix} \right)^{-1} \nabla f(x_k).$$

$$= -V_+ \Lambda_+^{-1} V_+^T \nabla f(x_k) \leftarrow \text{descent}$$

$$- \frac{1}{\varepsilon} V_\varepsilon V_\varepsilon^T \nabla f(x_k)$$

$$+ V_- \Lambda_-^{-1} V_-^T \nabla f(x_k).$$

"null space"

previous ascent

Option 3

Shift the entire spectrum

Compute $\lambda_{min} = \lambda_{min}(\nabla^2 f(x_k))$

Pick $\varepsilon > 0$

If $\lambda_{min} \geq \varepsilon \Rightarrow B_k = 0$

Otherwise, set $\gamma = \varepsilon - \lambda_{min}$ and

$$B_k = \nabla f(x_k) + \gamma I.$$

Clearly

$$\lambda_i(B_k) = \lambda_i - \lambda_{min} + \varepsilon \geq \varepsilon.$$

Moreover if $p = -(\nabla^2 f(x_k) + \gamma I)^{-1} \nabla f(x_k)$

$\Rightarrow$ as $\gamma \downarrow 0$, $p \to -\nabla^2 f(x_k)$ (Newton)

$\Rightarrow$ as $\gamma \uparrow \infty$, $\dfrac{p}{\|p\|} \to \dfrac{\nabla f(x_k)}{\|\nabla f(x_k)\|}$ (Gradient descent)

Next time we will cover convergence guarantees.