OPTIMIZATION (SI 416) – LECTURE 6

Harsha Hutridurga

IIT Bombay

- Start with a point $x^{(0)} \in \mathbb{R}^n$ and a direction $p^{(0)} \in \mathbb{R}^n$ such that

$$\left\langle \nabla f(x^{(0)}), p^{(0)} \right\rangle < 0$$

  i.e. $p^{(0)}$ is a descent direction at the point $x^{(0)}$
- Find the next iterate $x^{(1)}$ along the line $x^{(0)} + \alpha p^{(0)}$ with $\alpha > 0$ such that

$$f(x^{(1)}) \leq f(x^{(0)})$$

- General principle of line search algorithms:
  - At the current iterate $x^{(n)}$, choose a descent direction $p^{(n)}$, i.e.

$$\left\langle \nabla f(x^{(n)}), p^{(n)} \right\rangle < 0$$

  - Pick the next iterate $x^{(n+1)}$ along the line $x^{(n)} + \alpha p^{(n)}$ with $\alpha > 0$ such that

$$f(x^{(n+1)}) \leq f(x^{(n)})$$

RECAP: LINE SEARCH ALGORITHMS

- We saw an example (Lecture 5) which demonstrated that this approach may not succeed always
- The root cause for this behaviour stems from the choice of step lengths α_n in each iteration step
- Here we encounter certain sufficient decrease conditions to avoid such scenarios

♣ Recall: While at the point $x^{(n)}$, the search direction $p^{(n)}$ is said to be a descent direction if

$$\left\langle \nabla f(x^{(n)}), p^{(n)} \right\rangle < 0$$

Definition (Wolfe conditions)

For a descent direction $p^{(n)}$ at the point $x^{(n)}$, the step length α_n is said to satisfy the WOLFE CONDITIONS *if*

$$f(x^{(n)} + \alpha_n p^{(n)}) \leq f(x^{(n)}) + c_1 \alpha_n \left\langle \nabla f(x^{(n)}), p^{(n)} \right\rangle$$

$$\left\langle \nabla f(x^{(n)} + \alpha_n p^{(n)}), p^{(n)} \right\rangle \geq c_2 \left\langle \nabla f(x^{(n)}), p^{(n)} \right\rangle$$

for some constants

$$0 < c_1 < c_2 < 1$$

- Let $p^{(n)}$ be a descent direction at the point $x^{(n)}$
- Consider the function $\varphi : [0,\infty) \to \mathbb{R}$ defined as follows:

$$\varphi(\alpha) := f(x^{(n)} + \alpha p^{(n)}) \qquad \text{for } \alpha \in [0,\infty).$$

- Note that

$$\varphi'(0) = \left\langle \nabla f(x^{(n)}), p^{(n)} \right\rangle < 0$$

- Let $c_1 \in (0,1)$ and take a linear function $\psi : [0,\infty) \to \mathbb{R}$ defined as

$$\psi(\alpha) := f(x^{(n)}) + c_1\alpha \left\langle \nabla f(x^{(n)}), p^{(n)} \right\rangle \qquad \text{for } \alpha \in [0,\infty)$$

$$\text{Note that} \qquad \psi'(\alpha) = c_1 \left\langle \nabla f(x^{(n)}), p^{(n)} \right\rangle = c_1\varphi'(0)$$

- As $c_1 \in (0,1)$, we have $\varphi(\alpha) \leq \psi(\alpha)$ for small enough α
- The first condition in the Wolfe conditions is referred to as the SUFFICIENT DECREASE condition

wolfe conditions (contd.)

- We have seen that sufficient decrease condition is satisfied by any α_n as long as it is small
- To ensure that sufficient progress is made in each iteration step, a second condition is imposed:

$$\left\langle \nabla f(x^{(n)} + \alpha_n p^{(n)}), p^{(n)} \right\rangle \geq c_2 \left\langle \nabla f(x^{(n)}), p^{(n)} \right\rangle$$

- Written in terms of the function φ defined earlier, it reads

$$\varphi'(\alpha_n) \geq c_2 \varphi'(0)$$

♣ Is it possible to find step lengths satisfying Wolfe conditions?

Lemma

Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable.
Let $p^{(n)}$ be a descent direction at $x^{(n)}$.
Let $M \in \mathbb{R}$ be such that

$$f(x^{(n)} + \alpha p^{(n)}) \geq M \qquad \text{for all } \alpha \geq 0.$$

If $0 < c_1 < c_2 < 1$, then there exist interval of step lengths satisfying the Wolfe conditions:

$$f(x^{(n)} + \alpha_n p^{(n)}) \leq f(x^{(n)}) + c_1 \alpha_n \left\langle \nabla f(x^{(n)}), p^{(n)} \right\rangle$$

$$\left\langle \nabla f(x^{(n)} + \alpha_n p^{(n)}), p^{(n)} \right\rangle \geq c_2 \left\langle \nabla f(x^{(n)}), p^{(n)} \right\rangle$$

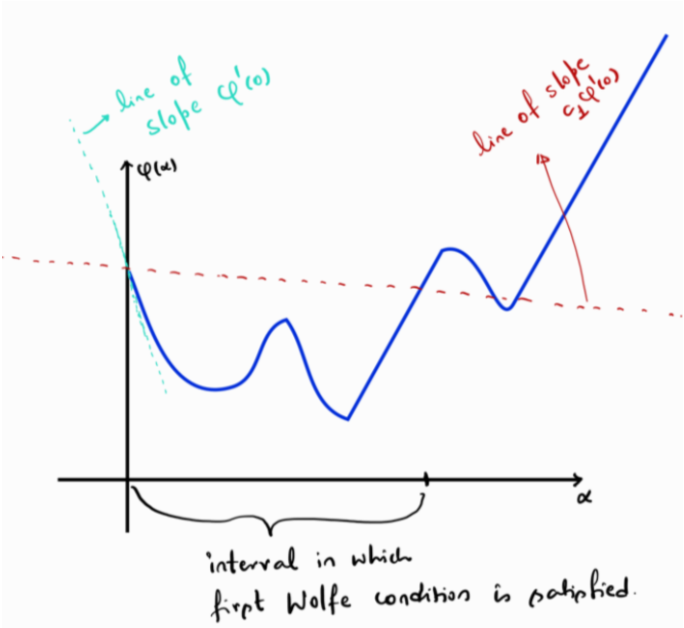
- Consider the functions $\varphi, \psi : [0, \infty) \to \mathbb{R}$ defined as

$$\varphi(\alpha) := f(x^{(n)} + \alpha p^{(n)}) \qquad \text{for } \alpha \in [0, \infty),$$
$$\psi(\alpha) := f(x^{(n)}) + c_1 \alpha \left\langle \nabla f(x^{(n)}), p^{(n)} \right\rangle \qquad \text{for } \alpha \in [0, \infty),$$

  where $0 < c_1 < 1$.
- Note that $\varphi'(0) < 0$ and $\psi'(\alpha) = c_1 \varphi'(0) < 0$
- Hence it follows that there exists a $\alpha_* > 0$ such that

$$\varphi(\alpha) \leq \psi(\alpha) \qquad \text{for } \alpha \in [0, \alpha_*]$$

- i.e. the graph of φ falls below the line ψ in $[0, \alpha_*]$


line of slope φ'(0)
φ(α)
line of slope c1φ'(0)
α
interval in which first Wolfe condition is satisfied.

SECOND WOLFE CONDITION

- Consider the function $\varphi : [0, \infty) \to \mathbb{R}$ from before

$$\varphi(\alpha) := f(x^{(n)} + \alpha p^{(n)}) \qquad \text{for } \alpha \in [0, \infty),$$

- Second Wolfe condition reads as follows:

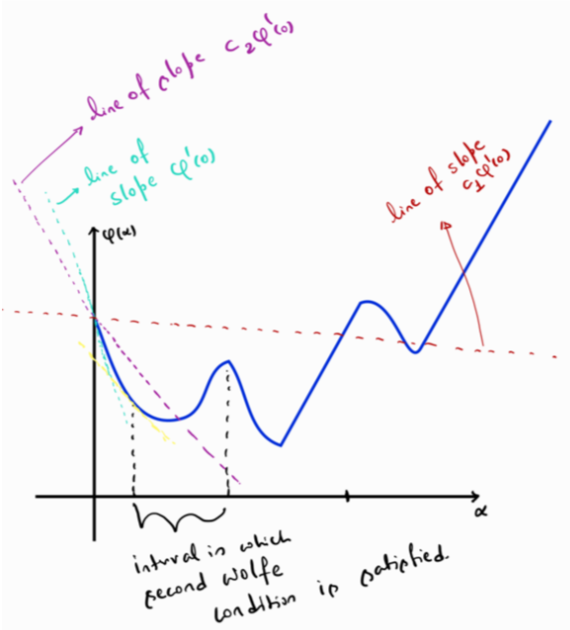
$$\left\langle \nabla f(x^{(n)} + \alpha_n p^{(n)}), p^{(n)} \right\rangle \geq c_2 \left\langle \nabla f(x^{(n)}), p^{(n)} \right\rangle$$

- i.e. slope of φ at α_n is greater than a multiple of its initial slope:

$$\varphi'(\alpha_n) \geq c_2 \varphi'(0) \qquad \text{for some } c_2 \in (c_1, 1)$$

- Note that $c_2 > c_1$ guarantees that $x^{(n+1)}$ is not too close to $x^{(n)}$

SECOND WOLFE CONDITION – GRAPHICAL ILLUSTRATION



♣ Is it possible to find step lengths satisfying Wolfe conditions?

Lemma

Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable.
Let $p^{(n)}$ be a descent direction at $x^{(n)}$.
Let $M \in \mathbb{R}$ be such that

$$f(x^{(n)} + \alpha p^{(n)}) \geq M \qquad \text{for all } \alpha \geq 0.$$

If $0 < c_1 < c_2 < 1$, then there exist interval of step lengths satisfying the Wolfe conditions:

$$f(x^{(n)} + \alpha_n p^{(n)}) \leq f(x^{(n)}) + c_1 \alpha_n \left\langle \nabla f(x^{(n)}), p^{(n)} \right\rangle$$

$$\left\langle \nabla f(x^{(n)} + \alpha_n p^{(n)}), p^{(n)} \right\rangle \geq c_2 \left\langle \nabla f(x^{(n)}), p^{(n)} \right\rangle$$

- Consider the functions $\varphi, \psi : [0, \infty) \to \mathbb{R}$ defined earlier:

$$\varphi(\alpha) := f(x^{(n)} + \alpha p^{(n)}) \qquad \text{for } \alpha \in [0, \infty),$$
$$\psi(\alpha) := f(x^{(n)}) + c_1 \alpha \left\langle \nabla f(x^{(n)}), p^{(n)} \right\rangle \qquad \text{for } \alpha \in [0, \infty),$$

- It is given that $\varphi(\alpha) \geq M$ for all $\alpha \geq 0$
- As $c_1 > 0$ and as $\langle \nabla f(x^{(n)}), p^{(n)} \rangle < 0$, it follows that ψ is unbounded below
- Note that $\varphi(0) = \psi(0) = f(x^{(n)})$
- Hence the graph of ψ must intersect the graph of φ at least once
- Let $\alpha' > 0$ be the smallest such that

$$\varphi(\alpha') = \psi(\alpha') \quad \text{and} \quad \varphi(\alpha) < \psi(\alpha) \quad \forall \alpha \in (0, \alpha')$$

- This is the first of the Wolfe conditions

- We saw that there exists a $\alpha' > 0$ such that

$$f(x^{(n)} + \alpha' p^{(n)}) = f(x^{(n)}) + c_1 \alpha' \left\langle \nabla f(x^{(n)}), p^{(n)} \right\rangle$$

- Employing Taylor's theorem we have for some $\alpha'' \in (0, \alpha')$

$$f(x^{(n)} + \alpha' p^{(n)}) = f(x^{(n)}) + \alpha' \left\langle \nabla f(x^{(n)} + \alpha'' p^{(n)}), p^{(n)} \right\rangle$$

- Comparing the above two equalities, we deduce

$$c_1 \left\langle \nabla f(x^{(n)}), p^{(n)} \right\rangle = \left\langle \nabla f(x^{(n)} + \alpha'' p^{(n)}), p^{(n)} \right\rangle$$

- As $c_1 < c_2$ and as $\left\langle \nabla f(x^{(n)}), p^{(n)} \right\rangle < 0$, it follows that

$$\left\langle \nabla f(x^{(n)} + \alpha'' p^{(n)}), p^{(n)} \right\rangle > c_2 \left\langle \nabla f(x^{(n)}), p^{(n)} \right\rangle$$

- This is the second of the Wolfe conditions

- We showed the first inequality (strict) holds true for all $\alpha \in (0, \alpha')$
- Further, the second inequality (strict) holds true for an $\alpha'' \in (0, \alpha')$
- As f is continuously differentiable, both the inequalities hold true for all α in an interval around α''
- There is a stronger version of the Wolfe conditions:

$$f(x^{(n)} + \alpha_n p^{(n)}) \leq f(x^{(n)}) + c_1 \alpha_n \left\langle \nabla f(x^{(n)}), p^{(n)} \right\rangle$$

$$\left| \left\langle \nabla f(x^{(n)} + \alpha_n p^{(n)}), p^{(n)} \right\rangle \right| \leq c_2 \left| \left\langle \nabla f(x^{(n)}), p^{(n)} \right\rangle \right|$$

  for some $0 < c_1 < c_2 < 1$
- Observe that the first condition is the same as before
- The second condition makes sure that slope of φ isn't too positive
- If the step lengths in a line search algorithm satisfy the Wolfe conditions, then to what limit the sequence generated converges?

Theorem

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a once continuously differentiable function such that $f(x) \geq M$ for all $x \in \mathbb{R}^n$. Let $x^{(0)}$ be starting point of the algorithm:

$$x^{(k+1)} = x^{(k)} + \alpha_k p^{(k)} \qquad \text{for } k = 0, 1, 2, \dots$$

Here $p^{(k)}$ is the descent direction at the point $x^{(k)}$.
Suppose the step lengths α_k satisfy the Wolfe conditions.
Consider the set $\Lambda := \left\{ x \in \mathbb{R}^n \text{ such that } f(x) \leq f(x^{(0)}) \right\}$.
Further suppose that f is β-smooth in an open set $\Omega \supset \Lambda$, i.e.

$$\|\nabla f(x) - \nabla f(y)\| \leq \beta \|x - y\| \qquad \text{for all } x, y \in \Omega.$$

Then

$$\sum_{k=0}^{\infty} \left(\cos^2 \theta_k\right) \left\| \nabla f(x^{(k)}) \right\|^2 < \infty \quad \text{with} \quad \cos\theta_k := \frac{-\left\langle \nabla f(x^{(k)}), p^{(k)} \right\rangle}{\left\| \nabla f(x^{(k)}) \right\| \left\| p^{(k)} \right\|}$$

- ♣ Function being bounded below is not too restrictive as minimization pb for unbounded function (from below) is ill-defined
- ♣ Note that

$$\sum_{k=0}^{\infty} \left(\cos^2 \theta_k\right) \left\| \nabla f(x^{(k)}) \right\|^2 < \infty \implies \lim_{k\to\infty} \left(\cos^2 \theta_k\right) \left\| \nabla f(x^{(k)}) \right\|^2 = 0$$

- ♣ Suppose we choose the descent directions such that

$$\cos \theta_k \geq \delta > 0 \qquad \text{for all } k$$

- ♣ Then it follows that

$$\lim_{k\to\infty} \left\| \nabla f(x^{(k)}) \right\|^2 = 0$$

- ♣ Hence the iterates converge to a stationary point
- ♣ It doesn't guarantee that the iterates converge to a minimizer

- Recall the second Wolfe condition:

$$\left\langle \nabla f(x^{(k)} + \alpha_k p^{(k)}), p^{(k)} \right\rangle \geq c_2 \left\langle \nabla f(x^{(k)}), p^{(k)} \right\rangle$$

- Hence it follows that

$$\left\langle \nabla f(x^{(k+1)}) - \nabla f(x^{(k)}), p^{(k)} \right\rangle \geq \left(c_2 - 1\right) \left\langle \nabla f(x^{(k)}), p^{(k)} \right\rangle$$

- Note that Cauchy-Schwarz inequality says

$$\begin{aligned} \left\langle \nabla f(x^{(k+1)}) - \nabla f(x^{(k)}), p^{(k)} \right\rangle &\leq \left\| \nabla f(x^{(k+1)}) - \nabla f(x^{(k)}) \right\| \left\| p^{(k)} \right\| \\ &\leq \beta\, \alpha_k \left\| p^{(k)} \right\|^2 \end{aligned}$$

thanks to the β-smoothness assumption on f

- Putting it all together, we obtain

$$\beta\,\alpha_k \left\|p^{(k)}\right\|^2 \ge \left(c_2 - 1\right) \left\langle \nabla f(x^{(k)}), p^{(k)} \right\rangle$$

- Hence we deduce

$$-\,\alpha_k \le \frac{\left(1 - c_2\right)}{\beta} \frac{\left\langle \nabla f(x^{(k)}), p^{(k)} \right\rangle}{\left\|p^{(k)}\right\|^2}$$

- Recall the first Wolfe condition:

$$f(x^{(k)} + \alpha_k p^{(k)}) \le f(x^{(k)}) + c_1 \alpha_k \left\langle \nabla f(x^{(k)}), p^{(k)} \right\rangle$$

- Using the above bound on $-\alpha_k$, we deduce

$$f(x^{(k+1)}) \le f(x^{(k)}) - \frac{c_1\left(1 - c_2\right)}{\beta} \frac{\left|\left\langle \nabla f(x^{(k)}), p^{(k)} \right\rangle\right|^2}{\left\|p^{(k)}\right\|^2}$$

- Recall that the angle between $p^{(k)}$ and $-\nabla f(x^{(k)})$ satisfies

$$\cos\theta_k := \frac{-\left\langle \nabla f(x^{(k)}), p^{(k)} \right\rangle}{\left\|\nabla f(x^{(k)})\right\| \, \left\|p^{(k)}\right\|}$$

- Define a positive constant $C := \frac{c_1(1-c_2)}{\beta}$
- We have thus shown that

$$f(x^{(k+1)}) \le f(x^{(k)}) - C\left(\cos^2\theta_k\right)\left\|\nabla f(x^{(k)})\right\|^2 \qquad \text{for } k = 0, 1, 2, \ldots$$

- Above recurrence inequality leads to

$$f(x^{(k)}) \le f(x^{(0)}) - C\sum_{\ell=0}^{k-1}\left(\cos^2\theta_\ell\right)\left\|\nabla f(x^{(\ell)})\right\|^2 \qquad \text{for } k = 0, 1, 2, \ldots$$

- Recall that we have assumed that f is bounded below. Hence

$$f(x^{(k)}) \geq M \implies -f(x^{(k)}) \leq -M$$
$$\implies f(x^{(0)}) - f(x^{(k)}) \leq f(x^{(0)}) - M$$

- Recall that we had derived the inequality

$$C \sum_{\ell=0}^{k-1} \left(\cos^2 \theta_\ell\right) \left\| \nabla f(x^{(\ell)}) \right\|^2 \leq f(x^{(0)}) - f(x^{(k)})$$

- Now letting $k \to \infty$ and using the boundedness property from above, we deduce

$$\lim_{k\to\infty} \sum_{\ell=0}^{k} \left(\cos^2 \theta_\ell\right) \left\| \nabla f(x^{(\ell)}) \right\|^2 < \infty$$

- Take A to be a $n \times n$ symmetric positive definite matrix
- Take $b \in \mathbb{R}^n$
- Consider the function $f : \mathbb{R}^n \to \mathbb{R}$ defined as follows

$$f(x) := \frac{1}{2} \langle Ax, x \rangle - \langle b, x \rangle \qquad \text{for } x \in \mathbb{R}^n.$$

- Recall that f is convex and its gradient

$$\nabla f(x) = Ax - b$$

- Hence the unique minimizer x_* of f solves the system $Ax = b$
- Employing steepest descent for this objective functions amounts to taking $-\nabla f(x^{(k)})$ as the descent direction
- Here, we can indeed find the best step length α_k that minimizes

$$f(x^{(k)} - \alpha \nabla f(x^{(k)}))$$

♣ Consider the function $\varphi : [0,\infty) \to \mathbb{R}$ defined as

$$\varphi(\alpha) := f(x^{(k)} - \alpha \nabla f(x^{(k)})) \quad \text{for } \alpha \in [0,\infty)$$

♣ Note that

$$\begin{aligned}
\varphi(\alpha) &= \frac{1}{2}\left\langle A(x^{(k)} - \alpha\nabla f(x^{(k)})), x^{(k)} - \alpha\nabla f(x^{(k)})\right\rangle \\
&\qquad - \left\langle b, x^{(k)} - \alpha\nabla f(x^{(k)})\right\rangle \\
&= \frac{1}{2}\left\langle Ax^{(k)}, x^{(k)}\right\rangle - \alpha\left\langle Ax^{(k)}, \nabla f(x^{(k)})\right\rangle \\
&+ \frac{\alpha^2}{2}\left\langle A\nabla f(x^{(k)}), \nabla f(x^{(k)})\right\rangle - \left\langle b, x^{(k)}\right\rangle + \alpha\left\langle b, \nabla f(x^{(k)})\right\rangle \\
&= \frac{1}{2}\left\langle Ax^{(k)}, x^{(k)}\right\rangle - \alpha\left\langle \nabla f(x^{(k)}), \nabla f(x^{(k)})\right\rangle \\
&+ \frac{\alpha^2}{2}\left\langle A\nabla f(x^{(k)}), \nabla f(x^{(k)})\right\rangle - \left\langle b, x^{(k)}\right\rangle
\end{aligned}$$

- Hence we have

$$\varphi'(\alpha) = -\left\langle \nabla f(x^{(k)}), \nabla f(x^{(k)}) \right\rangle + \alpha \left\langle A \nabla f(x^{(k)}), \nabla f(x^{(k)}) \right\rangle$$

- Thus the best step length α_k is given by

$$\alpha_k = \frac{\left\| \nabla f(x^{(k)}) \right\|^2}{\left\langle A \nabla f(x^{(k)}), \nabla f(x^{(k)}) \right\rangle}$$

- This approach is called EXACT LINE SEARCH approach
- Hence the steepest descent algorithm takes the form

$$x^{(k+1)} = x^{(k)} - \frac{\left\| \nabla f(x^{(k)}) \right\|^2}{\left\langle A \nabla f(x^{(k)}), \nabla f(x^{(k)}) \right\rangle} \nabla f(x^{(k)})$$

- Fact: For any symmetric positive definite matrix A, the following is a norm on $\mathbb{R}^n$

$$\|x\|_A := \sqrt{\langle Ax, x\rangle} \qquad \text{for } x \in \mathbb{R}^n$$

- Consider the quadratic function f defined earlier using A and b
- We claim that

$$\frac{1}{2}\|x - x_*\|_A^2 = f(x) - f(x_*) \qquad \text{for } x \in \mathbb{R}^n,$$

where $Ax_* = b$.
- Note that

$$\begin{aligned} f(x) - f(x_*) &= \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle - \frac{1}{2}\langle Ax_*, x_*\rangle + \langle b, x_*\rangle \\ &= \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle + \frac{1}{2}\langle b, x_*\rangle \end{aligned}$$

- Note further that

$$\begin{aligned}\frac{1}{2}\|x-x_*\|_A^2 &= \frac{1}{2}\langle Ax-Ax_*, x-x_*\rangle \\ &= \frac{1}{2}\langle Ax, x\rangle - \langle Ax_*, x\rangle + \frac{1}{2}\langle Ax_*, x_*\rangle \\ &= \frac{1}{2}\langle Ax, x\rangle - -\langle b, x\rangle + \frac{1}{2}\langle b, x_*\rangle\end{aligned}$$

- We have thus proved the claim
- Further note that

$$\nabla f(x) = Ax - b = Ax - Ax_* \implies A^{-1}\nabla f(x) = x - x_*$$

- Hence it follows that

$$\begin{aligned}\|x-x_*\|_A^2 &= \left\langle AA^{-1}\nabla f(x), A^{-1}\nabla f(x)\right\rangle \\ &= \left\langle \nabla f(x), A^{-1}\nabla f(x)\right\rangle\end{aligned}$$

Proposition

Let $f : \mathbb{R}^n \to \mathbb{R}$ *be the following quadratic function*

$$f(x) := \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle \qquad \textit{for } x \in \mathbb{R}^n,$$

where A *is symmetric positive definite and* $b \in \mathbb{R}^n$*.*
Consider the steepest descent algorithm with exact line searches:

$$x^{(k+1)} = x^{(k)} - \alpha_k \nabla f(x^{(k)}) \qquad \textit{for } k = 0, 1, 2, \ldots$$

where the step length α_k *defined earlier. Then, for* $k = 0, 1, 2, \ldots$

$$\frac{\left\|x^{(k+1)} - x_*\right\|_A^2}{\left\|x^{(k)} - x_*\right\|_A^2} = \left\{1 - \frac{\left\|\nabla f(x^{(k)})\right\|^4}{\left\langle A\nabla f(x^{(k)}), \nabla f(x^{(k)})\right\rangle \langle \nabla f(x), A^{-1}\nabla f(x)\rangle}\right\}$$

- Note that

$$\left\| x^{(k+1)} - x_* \right\|_A^2 = \left\langle A(x^{(k)} - x_*), x^{(k)} - x_* \right\rangle - 2\alpha_k \left\langle A(x^{(k)} - x_*), \nabla f(x^{(k)}) \right\rangle + \alpha_k^2 \left\langle A\nabla f(x^{(k)}), \nabla f(x^{(k)}) \right\rangle$$

- Using the earlier observation that $A(x - x_*) = \nabla f(x)$, we get

$$\left\| x^{(k+1)} - x_* \right\|_A^2 = \left\| x^{(k)} - x_* \right\|_A^2 - 2\alpha_k \left\| \nabla f(x^{(k)}) \right\|^2 + \alpha_k^2 \left\langle A\nabla f(x^{(k)}), \nabla f(x^{(k)}) \right\rangle$$

- The result follows from using the definition of α_k and the following observation from earlier:

$$\| x - x_* \|_A^2 = \left\langle \nabla f(x), A^{-1} \nabla f(x) \right\rangle$$

A MORE REFINED STATEMENT

♣ If $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ are the strictly positive eigenvalues of A, then

$$\left\| x^{(k+1)} - x_* \right\|_A^2 \leq \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1} \right)^2 \left\| x^{(k)} - x_* \right\|_A^2$$

END OF LECTURE 6

THANK YOU FOR YOUR ATTENTION