6.252 NONLINEAR PROGRAMMING

LECTURE 7: ADDITIONAL METHODS

LECTURE OUTLINE

- Least-Squares Problems and Incremental Gradient Methods
- Conjugate Direction Methods
- The Conjugate Gradient Method
- Quasi-Newton Methods
- Coordinate Descent Methods

- Recall the least-squares problem:

$$\text{minimize} \quad f(x) = \tfrac{1}{2}\|g(x)\|^2 = \tfrac{1}{2}\sum_{i=1}^{m} \|g_i(x)\|^2$$

$$\text{subject to} \quad x \in \Re^n,$$

where $g = (g_1, \ldots, g_m)$, $g_i : \Re^n \to \Re^{r_i}$.

INCREMENTAL GRADIENT METHODS

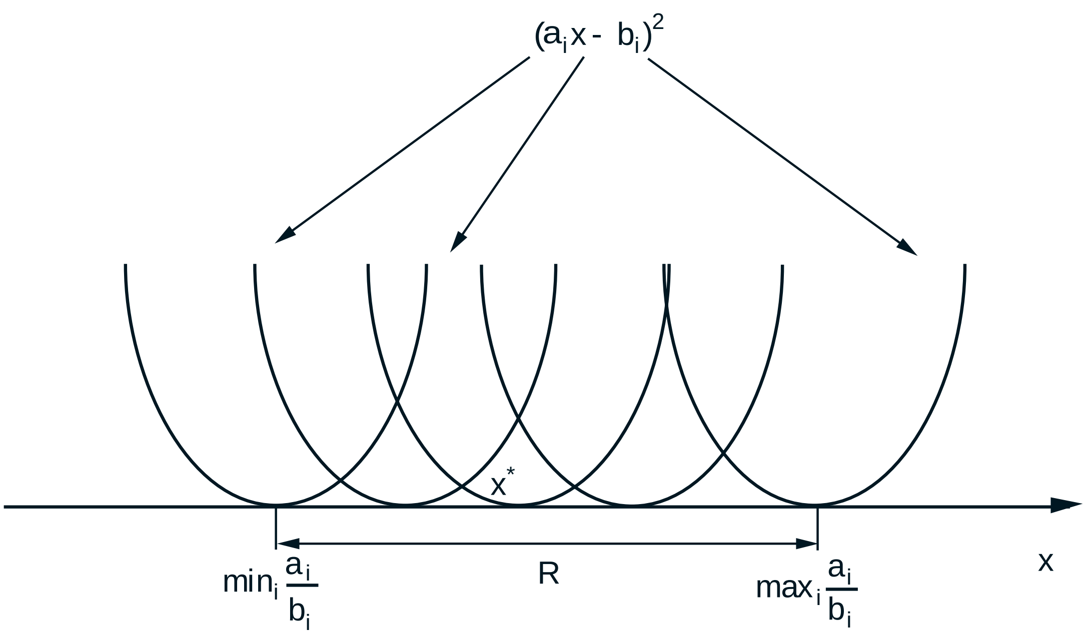
- Steepest descent method

$$x^{k+1} = x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \sum_{i=1}^{m} \nabla g_i(x^k) g_i(x^k)$$

- Incremental gradient method:

$$\psi_i = \psi_{i-1} - \alpha^k \nabla g_i(\psi_{i-1}) g_i(\psi_{i-1}), \quad i = 1, \ldots, m$$

$$\psi_0 = x^k, \qquad x^{k+1} = \psi_m$$



Advantage of incrementalism

VIEW AS GRADIENT METHOD W/ ERRORS

- Can write incremental gradient method as

$$x^{k+1} = x^k - \alpha^k \sum_{i=1}^{m} \nabla g_i(x^k) g_i(x^k)$$
$$+ \alpha^k \sum_{i=1}^{m} \big(\nabla g_i(x^k) g_i(x^k) - \nabla g_i(\psi_{i-1}) g_i(\psi_{i-1}) \big)$$

- Error term is proportional to stepsize α^k
- Convergence (generically) for a diminishing stepsize (under a Lipschitz condition on $\nabla g_i g_i$)
- Convergence to a “neighborhood” for a constant stepsize

CONJUGATE DIRECTION METHODS

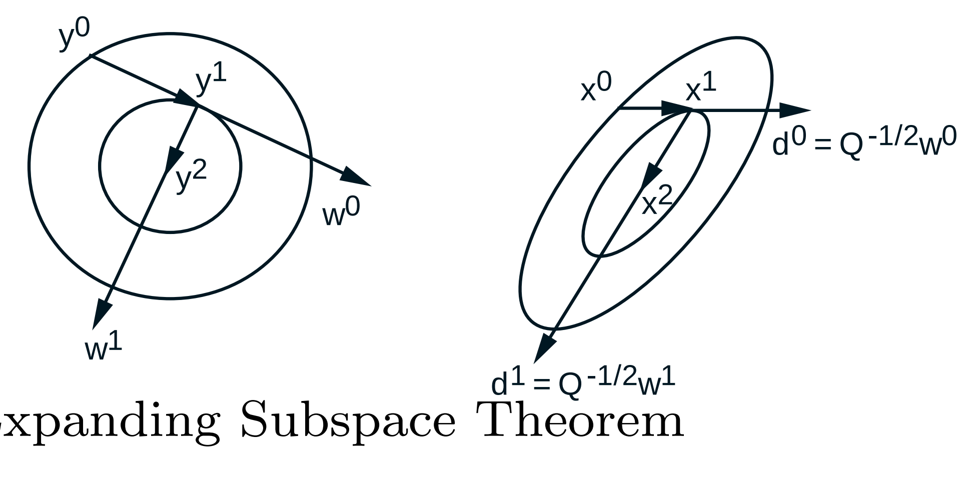
- Aim to improve convergence rate of steepest descent, without incurring the overhead of Newton's method

- Analyzed for a quadratic model. They require n iterations to minimize $f(x) = (1/2)x'Qx - b'x$ with Q an $n \times n$ positive definite matrix $Q > 0$.

- Analysis also applies to nonquadratic problems in the neighborhood of a nonsingular local min

- Directions $d^1, \ldots, d^k$ are Q-*conjugate*, if $d^{i'}Qd^j = 0$ for all $i \neq j$

- Generic conjugate direction method: $x^{k+1} = x^k + \alpha^k d^k$ where the d^ks are Q-conjugate and α^k is obtained by line minimization



Expanding Subspace Theorem

GENERATING Q-CONJUGATE DIRECTIONS

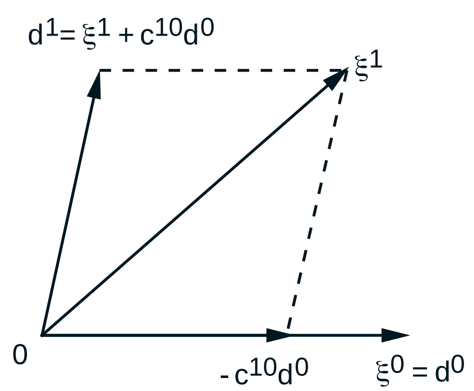
- Given set of linearly independent vectors $\xi^0, \ldots, \xi^k$, we can construct a set of Q-conjugate directions $d^0, \ldots, d^k$ s.t. $Span(d^0, \ldots, d^i) = Span(\xi^0, \ldots, \xi^i)$

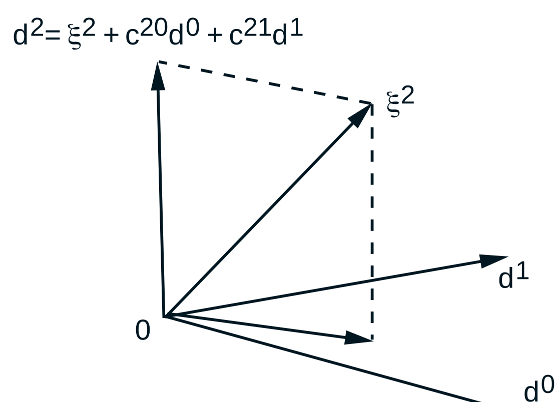
- *Gram-Schmidt procedure*. Start with $d^0 = \xi^0$. If for some $i < k$, $d^0, \ldots, d^i$ are Q-conjugate and the above property holds, take

$$d^{i+1} = \xi^{i+1} + \sum_{m=0}^{i} c^{(i+1)m} d^m;$$

choose $c^{(i+1)m}$ so d^{i+1} is Q-conjugate to $d^0, \ldots, d^i$,

$$d^{i+1'} Q d^j = \xi^{i+1'} Q d^j + \left(\sum_{m=0}^{i} c^{(i+1)m} d^m \right)' Q d^j = 0.$$





CONJUGATE GRADIENT METHOD

- Apply Gram-Schmidt to the vectors $\xi^k = g^k = \nabla f(x^k)$, $k = 0, 1, \ldots, n-1$

$$d^k = -g^k + \sum_{j=0}^{k-1} \frac{g^{k\prime} Q d^j}{d^{j\prime} Q d^j} d^j$$

- Key fact: Direction formula can be simplified.

Proposition : The directions of the CGM are generated by $d^0 = -g^0$, and

$$d^k = -g^k + \beta^k d^{k-1}, \qquad k = 1, \ldots, n-1,$$

where β^k is given by

$$\beta^k = \frac{g^{k\prime} g^k}{g^{k-1\prime} g^{k-1}} \quad \text{or} \quad \beta^k = \frac{(g^k - g^{k-1})' g^k}{g^{k-1\prime} g^{k-1}}$$

Furthermore, the method terminates with an optimal solution after at most n steps.

- Extension to nonquadratic problems.

QUASI-NEWTON METHODS

- $x^{k+1} = x^k - \alpha^k D^k \nabla f(x^k)$, where D^k is an inverse Hessian approximation
- Key idea: Successive iterates x^k, x^{k+1} and gradients $\nabla f(x^k)$, $\nabla f(x^{k+1})$, yield curvature info

$$q^k \approx \nabla^2 f(x^{k+1}) p^k,$$

$$p^k = x^{k+1} - x^k, \quad q^k = \nabla f(x^{k+1}) - \nabla f(x^k).$$

$$\nabla^2 f(x^n) \approx \left[\, q^0 \ \cdots \ q^{n-1} \,\right] \left[\, p^0 \ \cdots \ p^{n-1} \,\right]^{-1}$$

- Most popular Quasi-Newton method is a clever way to implement this idea

$$D^{k+1} = D^k + \frac{p^k p^{k\prime}}{p^{k\prime} q^k} - \frac{D^k q^k q^{k\prime} D^k}{q^{k\prime} D^k q^k} + \xi^k \tau^k v^k v^{k\prime},$$

$$v^k = \frac{p^k}{p^{k\prime} q^k} - \frac{D^k q^k}{\tau^k}, \quad \tau^k = q^{k\prime} D^k q^k, \quad 0 \le \xi^k \le 1$$

and $D^0 > 0$ is arbitrary, α^k by line minimization, and $D^n = Q^{-1}$ for a quadratic.

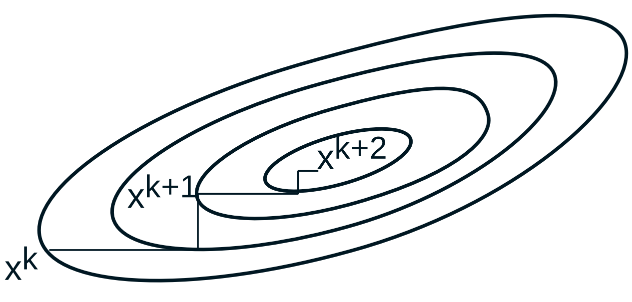
NONDERIVATIVE METHODS

- Finite difference implementations
- Forward and central difference formulas

$$\frac{\partial f(x^k)}{\partial x^i} \approx \frac{1}{h}\Big(f(x^k + he_i) - f(x^k)\Big)$$

$$\frac{\partial f(x^k)}{\partial x^i} \approx \frac{1}{2h}\Big(f(x^k + he_i) - f(x^k - he_i)\Big)$$

- Use central difference for more accuracy near convergence



- Coordinate descent. Applies also to the case where there are bound constraints on the variables.

- Direct search methods. Nelder-Mead method.

PROOF OF CONJUGATE GRADIENT RESULT

- Use induction to show that all gradients g^k generated up to termination are linearly independent. True for $k = 1$. Suppose no termination after k steps, and $g^0, \ldots, g^{k-1}$ are linearly independent. Then, $Span(d^0, \ldots, d^{k-1}) = Span(g^0, \ldots, g^{k-1})$ and there are two possibilities:
  - $g^k = 0$, and the method terminates.
  - $g^k \neq 0$, in which case from the expanding manifold property

    $$g^k \text{ is orthogonal to } d^0, \ldots, d^{k-1}$$

    $$g^k \text{ is orthogonal to } g^0, \ldots, g^{k-1}$$

    so g^k is linearly independent of $g^0, \ldots, g^{k-1}$, completing the induction.

- Since at most n lin. independent gradients can be generated, $g^k = 0$ for some $k \leq n$.

- Algebra to verify the direction formula.