6.252 NONLINEAR PROGRAMMING

LECTURE 4

CONVERGENCE ANALYSIS OF GRADIENT METHODS

LECTURE OUTLINE

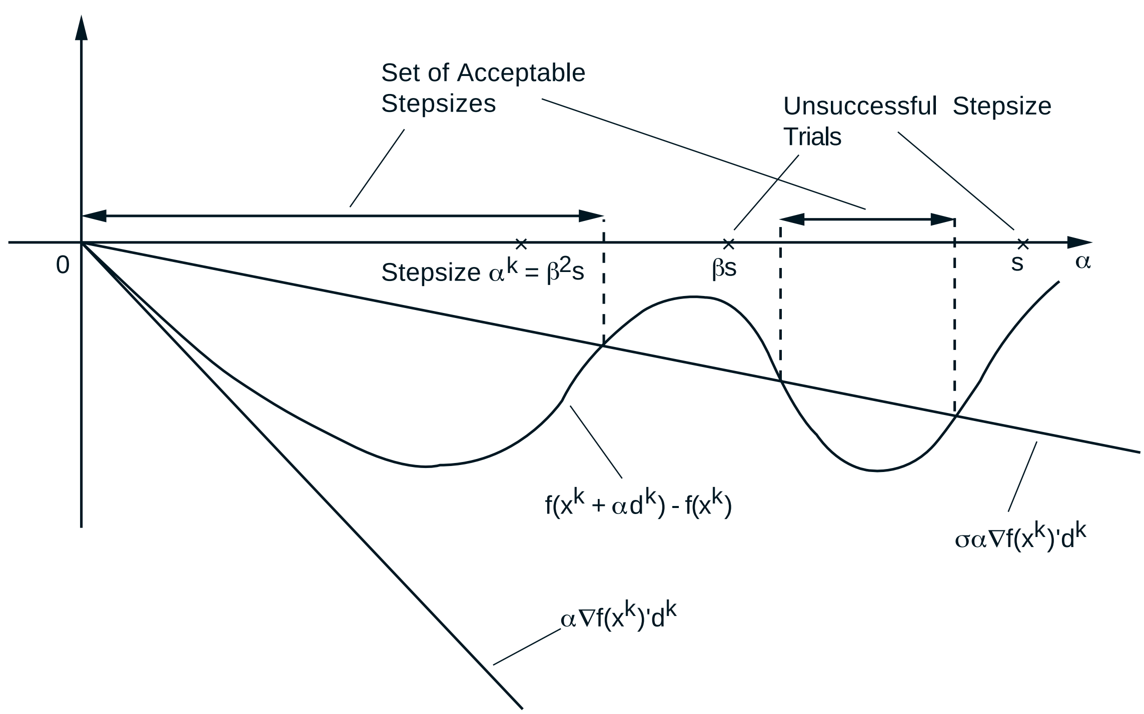
- Gradient Methods - Choice of Stepsize
- Gradient Methods - Convergence Issues

CHOICES OF STEPSIZE I

- Minimization Rule: α^k is such that

$$f(x^k + \alpha^k d^k) = \min_{\alpha \geq 0} f(x^k + \alpha d^k).$$

- Limited Minimization Rule: Min over $\alpha \in [0, s]$
- Armijo rule:



Start with s and continue with $\beta s, \beta^2 s, ...,$ until $\beta^m s$ falls within the set of α with

$$f(x^k) - f(x^k + \alpha d^k) \geq -\sigma \alpha \nabla f(x^k)' d^k.$$

CHOICES OF STEPSIZE II

- Constant stepsize: α^k is such that

$$\alpha^k = s : \text{ a constant}$$

- Diminishing stepsize:

$$\alpha^k \to 0$$

but satisfies the infinite travel condition

$$\sum_{k=0}^{\infty} \alpha^k = \infty$$

GRADIENT METHODS WITH ERRORS

$$x^{k+1} = x^k - \alpha^k(\nabla f(x^k) + e^k)$$

where e^k is an uncontrollable error vector

- Several special cases:
  - e^k small relative to the gradient; i.e., for all k, $\|e^k\| < \|\nabla f(x^k)\|$

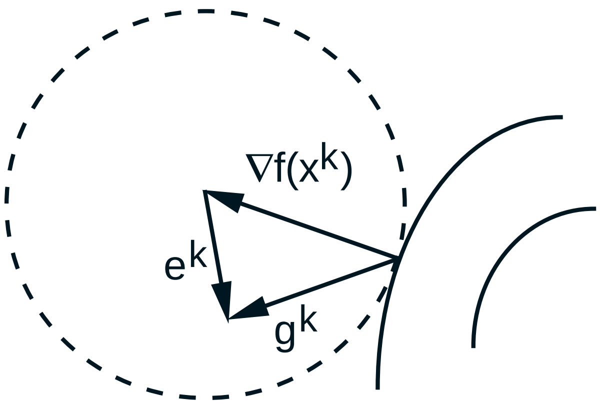


Illustration of the descent property of the direction $g^k = \nabla f(x^k) + e^k$.

  - $\{e^k\}$ is bounded, i.e., for all k, $\|e^k\| \leq \delta$, where δ is some scalar.
  - $\{e^k\}$ is proportional to the stepsize, i.e., for all k, $\|e^k\| \leq q\alpha^k$, where q is some scalar.
  - $\{e^k\}$ are independent zero mean random vectors

CONVERGENCE ISSUES

- Only convergence to stationary points can be guaranteed
- Even convergence to a single limit may be hard to guarantee (capture theorem)
- Danger of nonconvergence if directions d^k tend to be orthogonal to $\nabla f(x^k)$
- Gradient related condition:

For any subsequence $\{x^k\}_{k\in\mathcal{K}}$ that converges to a nonstationary point, the corresponding subsequence $\{d^k\}_{k\in\mathcal{K}}$ is bounded and satisfies

$$\limsup_{k\to\infty,\, k\in\mathcal{K}} \nabla f(x^k)'d^k < 0.$$

- Satisfied if $d^k = -D^k\nabla f(x^k)$ and the eigenvalues of D^k are bounded above and bounded away from zero

CONVERGENCE RESULTS

CONSTANT AND DIMINISHING STEPSIZES

Let $\{x^k\}$ be a sequence generated by a gradient method $x^{k+1} = x^k + \alpha^k d^k$, where $\{d^k\}$ is gradient related. Assume that for some constant $L > 0$, we have

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|, \qquad \forall\, x, y \in \Re^n,$$

Assume that either

(1) there exists a scalar ϵ such that for all k

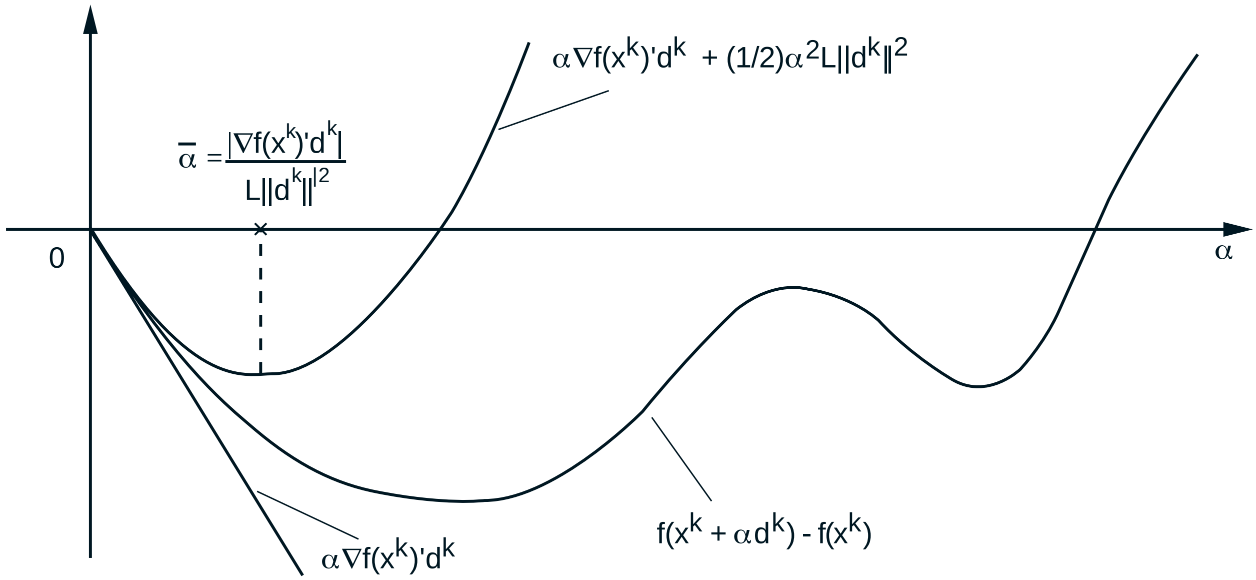
$$0 < \epsilon \le \alpha^k \le \frac{(2-\epsilon)|\nabla f(x^k)' d^k|}{L\|d^k\|^2}$$

or

(2) $\alpha^k \to 0$ and $\sum_{k=0}^{\infty} \alpha^k = \infty$.

Then either $f(x^k) \to -\infty$ or else $\{f(x^k)\}$ converges to a finite value and $\nabla f(x^k) \to 0$.

MAIN PROOF IDEA



The idea of the convergence proof for a constant stepsize. Given x^k and the descent direction d^k, the cost difference $f(x^k + \alpha d^k) - f(x^k)$ is majorized by $\alpha \nabla f(x^k)' d^k + \frac{1}{2}\alpha^2 L \|d^k\|^2$ (based on the Lipschitz assumption; see next slide). Minimization of this function over α yields the stepsize

$$\overline{\alpha} = \frac{|\nabla f(x^k)' d^k|}{L\|d^k\|^2}$$

This stepsize reduces the cost function f as well.

DESCENT LEMMA

Let α be a scalar and let $g(\alpha) = f(x + \alpha y)$. Have

$$
\begin{aligned}
f(x+y) - f(x) = g(1) - g(0) &= \int_0^1 \frac{dg}{d\alpha}(\alpha)\, d\alpha \\
&= \int_0^1 y' \nabla f(x + \alpha y)\, d\alpha \\
&\leq \int_0^1 y' \nabla f(x)\, d\alpha \\
&\quad + \left| \int_0^1 y' \big(\nabla f(x + \alpha y) - \nabla f(x) \big)\, d\alpha \right| \\
&\leq \int_0^1 y' \nabla f(x)\, d\alpha \\
&\quad + \int_0^1 \|y\| \cdot \|\nabla f(x + \alpha y) - \nabla f(x)\| d\alpha \\
&\leq y' \nabla f(x) + \|y\| \int_0^1 L\alpha \|y\|\, d\alpha \\
&= y' \nabla f(x) + \frac{L}{2} \|y\|^2.
\end{aligned}
$$

CONVERGENCE RESULT – ARMIJO RULE

Let $\{x^k\}$ be generated by $x^{k+1} = x^k + \alpha^k d^k$, where $\{d^k\}$ is gradient related and α^k is chosen by the Armijo rule. Then every limit point of $\{x^k\}$ is stationary.

Proof Outline: Assume $\overline{x}$ is a nonstationary limit point. Then $f(x^k) \to f(\overline{x})$, so $\alpha^k \nabla f(x^k)' d^k \to 0$.

- If $\{x^k\}_{\mathcal{K}} \to \overline{x}$, $\limsup_{k\to\infty,\, k\in\mathcal{K}} \nabla f(x^k)' d^k < 0$, by gradient relatedness, so that $\{\alpha^k\}_{\mathcal{K}} \to 0$.
- By the Armijo rule, for large $k \in \mathcal{K}$

$$f(x^k) - f\big(x^k + (\alpha^k/\beta)d^k\big) < -\sigma(\alpha^k/\beta)\nabla f(x^k)' d^k.$$

Defining $p^k = \frac{d^k}{\|d^k\|}$ and $\overline{\alpha}^k = \frac{\alpha^k \|d^k\|}{\beta}$, we have

$$\frac{f(x^k) - f(x^k + \overline{\alpha}^k p^k)}{\overline{\alpha}^k} < -\sigma \nabla f(x^k)' p^k.$$

Use the Mean Value Theorem and let $k \to \infty$. We get $-\nabla f(\overline{x})'\overline{p} \le -\sigma \nabla f(\overline{x})'\overline{p}$, where $\overline{p}$ is a limit point of p^k – a contradiction since $\nabla f(\overline{x})'\overline{p} < 0$.