Gradient Methods for Unconstrained Optimization

- Unconstrained minimization of continuously differentiable function $f: \mathbb{R}^n \to \mathbb{R}$.

Iterative Descent Start at some point x_0, and successively generate $x_1, x_2, \ldots$ s.t.

$$f(x_{k+1}) < f(x_k) \quad k = 0, 1, \ldots$$

Hope to decrease f all the way to minimum.

Steepest Descent Move x_k in direction that decreases function most. This direction is $-\nabla f(x_k)$.

Why? For $v \in \mathbb{R}^n$, $v \neq 0$

$$f(x + \varepsilon v) = f(x) + \varepsilon \nabla f(x)^T v + o(\varepsilon)$$

Rate of change of f along direction v:

$$\lim_{\varepsilon \to 0} \frac{f(x + \varepsilon v) - f(x)}{\varepsilon} = \nabla f(x)^T v$$

By Cauchy-Schwarz inequality

$$-\|\nabla f(x)\| \|v\| \leq \nabla f(x)^T v \leq \|\nabla f(x)\| \|v\|$$

iff $v = -\beta \nabla f(x)$ for some $\beta > 0$

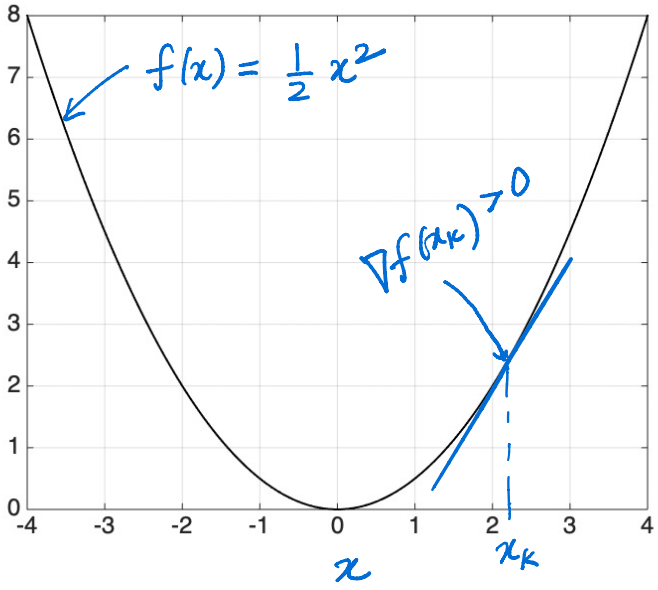
Maximum rate of decrease of f at point x is in direction $-\nabla f(x)$.

Steepest Descent Algorithm

Assuming that $\nabla f(x_k) \neq 0$ to go from x_k to x_{k+1} move along $-\nabla f(x_k)$,

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k)$$

α_k ← Step size



- α_k controls movement along gradient.
- If α_k is too large function value may increase
- Need to choose carefully

General Gradient Descent Algorithm

Assume that $\nabla f(x_k) \neq 0$. Then

$$x_{k+1} = x_k + \alpha_k d_k$$

where d_k is s.t.

$$\nabla f(x_k)^T d_k < 0 \quad \equiv \quad -\nabla f(x_k)^T d_k > 0$$

d_k has a positive projection along $-\nabla f(x_k)$

- If $d_k = -\nabla f(x_k)$ we get steepest descent
- Often d_k is constructed using matrix $D_k > 0$.

$$d_k = -D_k \nabla f(x_k)$$

Then $d_k^T \nabla f(x_k) = -\nabla f(x_k)^T D_k \nabla f(x_k) < 0$ if $\nabla f(x_k) \neq 0$

Methods for choosing α_k

(1) Fixed stepsize : $\alpha_k = \alpha$

- Can have issues with convergence

(2) Optimal line search : choose α_k to solve

$$\min_{\alpha \geq 0} f(x_k + \alpha d_k)$$

- may be difficult in practice

(3) Armijo's Rule (successive stepsize reduction)

$$f(x_k + \alpha_k d_k) = f(x_k) + \alpha_k \underbrace{\nabla f(x_k)^T d_k}_{<0} + o(\alpha_k)$$

- If α_k is suff. small, f decreases
- But don't want α_k to be too small (slow)

Armijo's rule :

(1) Initialize $\alpha_k = \tilde{\alpha}$. Let σ, β be prespecified parameters with both in $(0, 1)$.

(2) If $f(x_k + \alpha_k d_k) \leq f(x_k) + \sigma \alpha_k \nabla f(x_k)^T d_k$ Stop

(3) Else set $\alpha_k = \beta \alpha_k$, and go back to step 2

Termination at smallest integer m s.t.

$$f(x_k + \beta^m \tilde{\alpha} d_k) \leq f(x_k) + \sigma \beta^m \tilde{\alpha} \nabla f(x_k)^T d_k$$

Bersekas book : $\sigma \in [10^{-5}, 10^{-1}]$, $\beta \in [\frac{1}{10}, \frac{1}{2}]$

Convergence of Steepest Descent with Armijo's Rule

Background: Limit Points of sequences

Consider sequence $\{x_k\}$ in $\mathbb{R}^n$. A point $\bar{x} \in \mathbb{R}^n$ is called a limit point of $\{x_k\}$ if there is a subsequence of $\{x_k\}$ that converges to $\bar{x}$

i.e. $\lim_{\substack{k \to \infty \\ k \in \mathcal{K}}} x_k = \bar{x}$ ← indices of subsequence

Results 1) A bounded sequence in $\mathbb{R}^n$ converges iff it has a unique limit point

(2) Every bounded sequence has at least one limit point

Examples $x_k = (-1)^k$ has limit points $-1, 1$

$x_k = \left(\frac{1}{2}\right)^k$ converges to 0

Armijo's Rule for Steepest Descent:

$\alpha_k = \tilde{\alpha}\, \beta^{m_k}$ where m_k is smallest m s.t.

$$f(x_k - \tilde{\alpha}\beta^m \nabla f(x_k)) \leq f(x_k) - \sigma\tilde{\alpha}\beta^m \|\nabla f(x_k)\|^2$$

Result Assume $\inf_x f(x) > -\infty$. Then every limit point of $\{x_k\}$ for steepest descent with Armijo's rule is a stationary point of f

Proof Assume that $\bar{x}$ is a limit point of $\{x_k\}$ s.t. $\nabla f(\bar{x}) \neq 0$.

- Since $\{f(x_k)\}$ is monotonically non-increasing and bounded below, $\{f(x_k)\}$ converges
- f is continuous $\Rightarrow$ $f(\bar{x})$ is a limit point of $\{f(x_k)\}$

$$\Rightarrow \lim_{k\to\infty} f(x_k) = f(\bar{x})$$

$$\Rightarrow f(x_k) - f(x_{k+1}) \to 0 \quad -(*)$$

- By definition of Armijo's rule:

$$f(x_k) - f(x_{k+1}) \geq \sigma \alpha_k \|\nabla f(x_k)\|^2$$

$$\Rightarrow \alpha_k \|\nabla f(x_k)\|^2 \to 0 \quad (\text{by } (*))$$

Let $\{x_k\}_{\mathcal{K}}$ be a subsequence converging to $\bar{x}$. Then by the continuity of $\|\nabla f(x)\|^2$, and by the assumption that $\nabla f(\bar{x}) \neq 0$,

$$\lim_{\substack{k\to\infty \\ k\in\mathcal{K}}} \|\nabla f(x_k)\|^2 = \|\nabla f(\bar{x})\|^2 \neq 0$$

$$\Rightarrow \lim_{\substack{k\to\infty \\ k\in\mathcal{K}}} \alpha_k = 0$$

Recall that $\alpha_k = \bar{\alpha}\,\beta^{m_k} \Rightarrow \ln \alpha_k = \ln \bar{\alpha} + m_k \ln \beta$

$$\Rightarrow m_k = \frac{\ln \alpha_k - \ln \bar{\alpha}}{\ln \beta} \xrightarrow[k \in \mathcal{K}]{k \to \infty} \infty$$

($\ln \beta < 0$)

$\Rightarrow \exists\ \bar{k}$ s.t. $m_k > 1$ for all $k > \bar{k}$, $k \in \mathcal{K}$

- $m_k > 1$ means Armijo's rule does stepsize reduction at least once

$$\Rightarrow f(x_k) - f\left(x_k - \frac{\alpha_k}{\beta} \nabla f(x_k)\right) < \sigma \frac{\alpha_k}{\beta} \|\nabla f(x_k)\|^2$$

$\forall\ k \in \mathcal{K},\ k > \bar{k}$

By Taylor's Theorem,

$$f\left(x_k - \frac{\alpha_k}{\beta} \nabla f(x_k)\right) = f(x_k) - \nabla f\left(x_k - \frac{\bar{\alpha}_k}{\beta} \nabla f(x_k)\right)^T \frac{\alpha_k}{\beta} \nabla f(x_k)$$

for some $\bar{\alpha}_k \in (0, \alpha_k)$

Therefore we get

$$\nabla f\left(x_k - \frac{\bar{\alpha}_k}{\beta} \nabla f(x_k)\right)^T \nabla f(x_k) < \sigma \|\nabla f(x_k)\|^2$$

$\forall\ k \in \mathcal{K},\ k > \bar{k}$

Taking limit as $k \to \infty$ on $\mathcal{K}$, $\bar{\alpha}_k \to 0$

$$\|\nabla f(\bar{x})\|^2 < \sigma \|\nabla f(\bar{x})\|^2 < \|\nabla f(\bar{x})\|^2$$

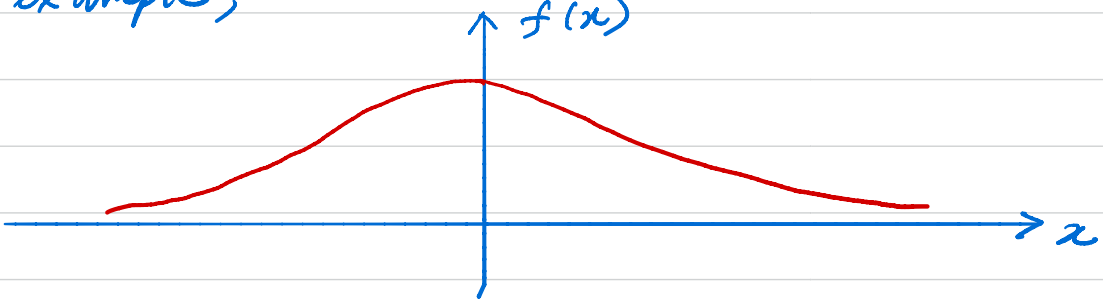
($\sigma < 1$)

$\Rightarrow\Leftarrow$

Remarks:

1. Any stepsize selection rule that decreases f by more than Armijo's rule inherits convergence properties. Also can replace $-\nabla f(x_k)$ by d_k (see book)

2. Lemma does not imply that $\{x_k\}$ has (finite) limit points, even though $\{f(x_k)\}$ converges. For example,



3. Even if $\{x_k\}$ is bounded, it could have multiple limit points, and convergence to single limit point not guaranteed. However, it can be shown that local minima that are isolated stationary points tend to attract most gradient methods. See Prop 1.2.4 of text (Capture Theorem).

4. If f is convex, every stationary point is a global min. Therefore even if $\{x_k\}$ does not converge, we can get arbitrarily close to global min.