End of Important Aside: Second Order conditions for Convexity, Strong Convexity, Lipschitz Continuity of Gradient, Convex Conjugate, Fenchel Duality.

Using Strong Convexity: Revisiting Convergence Analysis

- $f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla^\top f(\mathbf{x})(\mathbf{y}-\mathbf{x}) + \frac{m}{2}\|\mathbf{y}-\mathbf{x}\|^2$
  $\geq$

Using Strong Convexity: Revisiting Convergence Analysis

- $f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla^\top f(\mathbf{x})(\mathbf{y} - \mathbf{x}) + \frac{m}{2}\|\mathbf{y} - \mathbf{x}\|^2$
  $\geq$ minimum value the RHS can take as a function of $\mathbf{y}$
- Minimum value of RHS
  $\nabla f(\mathbf{x}) + m\mathbf{y} - m\mathbf{x} = 0$
  $\implies \mathbf{y} = \mathbf{x} - \frac{1}{m}\nabla f(\mathbf{x})$
- Thus,

Using Strong Convexity: Revisiting Convergence Analysis

- $f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla^\top f(\mathbf{x})(\mathbf{y}-\mathbf{x}) + \frac{m}{2}\|\mathbf{y}-\mathbf{x}\|^2$
  $\geq$ minimum value the RHS can take as a function of $\mathbf{y}$
- Minimum value of RHS
  $\nabla f(\mathbf{x}) + m\mathbf{y} - m\mathbf{x} = 0$
  $\implies \mathbf{y} = \mathbf{x} - \frac{1}{m}\nabla f(\mathbf{x})$
- Thus,
  $f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla^\top f(\mathbf{x})\left(-\frac{1}{m}\nabla f(\mathbf{x})\right) + \frac{m}{2}\left\|-\frac{1}{m}\nabla f(\mathbf{x})\right\|^2$
  $\implies f(\mathbf{y}) \geq f(\mathbf{x}) - \frac{1}{2m}\|\nabla f(\mathbf{x})\|^2$
  - Here, LHS is independent of $\mathbf{x}$, and RHS is independent of $\mathbf{y}$
  - Thus the inequality also holds for $\mathbf{y} = \mathbf{x}^*$ (point of minimum of $f(\mathbf{x})$)

Using Strong Convexity: Revisiting Convergence Analysis (contd.)

$$f(\mathbf{x}) \geq f(\mathbf{x}^*) \geq f(\mathbf{x}) - \frac{1}{2m}\|\nabla f(\mathbf{x})\|^2$$

Using Strong Convexity: Revisiting Convergence Analysis (contd.)

$$f(\mathbf{x}) \geq f(\mathbf{x}^*) \geq f(\mathbf{x}) - \frac{1}{2m}\left\|\nabla f(\mathbf{x})\right\|^2$$

That is, if $f(\mathbf{x}^*) = p^*$ then $f(\mathbf{x}) \in \left[p^* - \frac{1}{2m}\left\|\nabla f(\mathbf{x})\right\|^2, p^* + \frac{1}{2m}\left\|\nabla f(\mathbf{x})\right\|^2\right]$

Using Strong Convexity: Revisiting Convergence Analysis (contd.)

$$f(\mathbf{x}) \geq f(\mathbf{x}^*) \geq f(\mathbf{x}) - \frac{1}{2m}\left\|\nabla f(\mathbf{x})\right\|^2$$

That is, if $f(\mathbf{x}^*) = p^*$ then $f(\mathbf{x}) \in \left[p^* - \frac{1}{2m}\left\|\nabla f(\mathbf{x})\right\|^2, p^* + \frac{1}{2m}\left\|\nabla f(\mathbf{x})\right\|^2\right]$

- If $\left\|\nabla f(\mathbf{x})\right\|$ is small, the point is nearly optimal
  - If $\left\|\nabla f(\mathbf{x})\right\| \leq \sqrt{2m\epsilon}$, then: $f(\mathbf{x}) - p^* \leq \epsilon$
  - As the gradient $\left\|\nabla f(\mathbf{x})\right\|$ approaches 0, we get closer to an optimal solution $\mathbf{x}^*$

Rate of Convergence using Strong Convexity and Lipschitz Continuity for fixed step size ($t = \frac{1}{l}$)

Recap from Convergence using Lipschitz Continuity

- We recap the (necessary) inequality resulting from Lipschitz continuity of $\nabla f(\mathbf{x})$: $f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla^\top f(\mathbf{x})(\mathbf{y} - \mathbf{x}) + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2$
- Considering $\mathbf{x}^k \equiv \mathbf{x}$, and $\mathbf{x}^{k+1} = \mathbf{x}^k - t\nabla f(\mathbf{x}^k) \equiv \mathbf{y}$, we get

Recap from Convergence using Lipschitz Continuity

- We recap the (necessary) inequality resulting from Lipschitz continuity of $\nabla f(\mathbf{x})$: $f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla^{\top} f(\mathbf{x})(\mathbf{y}-\mathbf{x}) + \frac{L}{2}\|\mathbf{y}-\mathbf{x}\|^2$
- Considering $\mathbf{x}^k \equiv \mathbf{x}$, and $\mathbf{x}^{k+1} = \mathbf{x}^k - t\nabla f(\mathbf{x}^k) \equiv \mathbf{y}$, we get

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) - t\nabla^{\top} f(\mathbf{x}^k)\nabla f(\mathbf{x}^k) + \frac{L\,(t)^2}{2}\left\|\nabla f(\mathbf{x}^k)\right\|^2$$

$$\implies f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) + (\frac{Lt^2}{2} - t)\left\|\nabla f(\mathbf{x}^k)\right\|^2$$

See https://www.youtube.com/watch?v=SGZdsQviFYs&list=PLsd82ngobrvcYfCdnSnqM7lKLqE9qUUpX&index=17

Using Strong Convexity: Revisiting Convergence Analysis (contd.)

- Since f is strongly convex, and also Lipschitz continuous, we have for some L :

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) + (\frac{Lt^2}{2} - t)\left\|\nabla f(\mathbf{x}^k)\right\|^2$$

- Also, $0 < t \leq \frac{2}{L}(1 - c_1) \implies \frac{Lt^2}{2} - t \leq -c_1 t$
- Thus, we get the exit condition of backtracking line search

$$f(x^{k+1}) \leq f(x^k) - c_1 t\left\|\nabla f(x^k)\right\|^2$$

$$\implies f\left(x^k - t\nabla f(x^k)\right) \leq f(x^k) - c_1 t\left\|\nabla f(x^k)\right\|^2$$

- Often $c_1 = \frac{1}{2}$.

Putting things together for Strong Convexity

- Let $p^* = f(x^*)$
- $f\left(\mathbf{x} - t\nabla f(\mathbf{x})\right) \leq f(\mathbf{x}) - t\left\|\nabla f(\mathbf{x})\right\|^2 + \frac{Lt^2}{2}\left\|\nabla f(\mathbf{x})\right\|^2$
  - Consider RHS for t=1/L

[7] See a more involved proof for backtracking line search later in the slides and later generalized for proximal/generalized gradient descent

Putting things together for Strong Convexity

- Let $p^* = f(x^*)$
- $f\left(\mathbf{x} - t\nabla f(\mathbf{x})\right) \leq f(\mathbf{x}) - t\left\|\nabla f(\mathbf{x})\right\|^2 + \frac{Lt^2}{2}\left\|\nabla f(\mathbf{x})\right\|^2$
  - Consider RHS for $t^* = \frac{1}{L}$[7]

$$\implies f\left(\mathbf{x} - t\nabla f(\mathbf{x})\right) \leq f(\mathbf{x}) - \frac{1}{2L}\left\|\nabla f(\mathbf{x})\right\|^2$$
$$\implies f\left(\mathbf{x} - t\nabla f(\mathbf{x})\right) - p^* \leq f(\mathbf{x}) - \frac{1}{2L}\left\|\nabla f(\mathbf{x})\right\|^2 - p^*$$

- From strong convexity, we had
  $f(y) \geq f(\mathbf{x}) - \frac{1}{2m}\left\|\nabla f(\mathbf{x})\right\|^2$
  $\implies p^* \geq f(\mathbf{x}) - \frac{1}{2m}\left\|\nabla f(\mathbf{x})\right\|^2$
  $\implies \left\|\nabla f(\mathbf{x})\right\|^2 \geq 2m\left(f(\mathbf{x}) - p^*\right)$

[7] See a more involved proof for backtracking line search later in the slides and later generalized for proximal/generalized gradient descent

Putting things together for Strong Convexity

- Thus,
$$f\left(\mathbf{x} - t^*\nabla f(\mathbf{x})\right) - p^* \leq f(\mathbf{x}) - \frac{1}{2L}\left\|\nabla f(\mathbf{x})\right\|^2 - p^*$$
$$\implies f\left(\mathbf{x} - t^*\nabla f(\mathbf{x})\right) - p^* \leq f(\mathbf{x}) - \frac{2m}{2L}\left(f(\mathbf{x}) - p^*\right) - p^*$$
$$\implies f\left(\mathbf{x} - t^*\nabla f(x)\right) - p^* \leq \left(1 - \frac{m}{L}\right)\left(f(\mathbf{x}) - p^*\right)$$
- That is,
$$f(\mathbf{x}^k) - p^* \leq \left(1 - \frac{m}{L}\right)\left(f(\mathbf{x}^{k-1}) - p^*\right)$$

between 0 and 1

Keep applying this inequality for k-1, k-2,.... 0

Putting things together for Strong Convexity

- Thus,
  $f\left(\mathbf{x} - t^*\nabla f(\mathbf{x})\right) - p^* \leq f(\mathbf{x}) - \frac{1}{2L}\left\|\nabla f(\mathbf{x})\right\|^2 - p^*$
  $\implies f\left(\mathbf{x} - t^*\nabla f(\mathbf{x})\right) - p^* \leq f(\mathbf{x}) - \frac{2m}{2L}\left(f(\mathbf{x}) - p^*\right) - p^*$
  $\implies f\left(\mathbf{x} - t^*\nabla f(x)\right) - p^* \leq \left(1 - \frac{m}{L}\right)\left(f(\mathbf{x}) - p^*\right)$
- That is,
  $f(\mathbf{x}^k) - p^* \leq \left(1 - \frac{m}{L}\right)\left(f(\mathbf{x}^{k-1}) - p^*\right)$
  $\leq \left(1 - \frac{m}{L}\right)^2\left(f(\mathbf{x}^{k-2}) - p^*\right)$
  $\vdots$
  $\leq \left(1 - \frac{m}{L}\right)^k\left(f(\mathbf{x}^{(0)}) - p^*\right)$

Linear Convergence under Strong Convexity Assumption

- We get linear convergence

$$f(\mathbf{x}^k) - p^* \leq \left(1 - \frac{m}{L}\right)^k \left(f(\mathbf{x}^{(0)}) - p^*\right)$$

  - Here, $\frac{m}{L} \in (0,1)$
  - This is, loosely speaking, faster than what we got using only Lipschitz continuity, which was: $f(\mathbf{x}^k) - p^* \leq \frac{\left\|x^{(0)} - \mathbf{x}^*\right\|^2}{2tk}$ (sublinear convergence)

here we needed O(1/eps) iterations

- To obtain $f(\mathbf{x}^k) - p^* \leq \epsilon$, we need

O(log(1/eps)) iterations

Linear Convergence under Strong Convexity Assumption

- We get linear convergence

$$f(\mathbf{x}^k) - p^* \leq \left(1 - \frac{m}{L}\right)^k \left(f(\mathbf{x}^{(0)}) - p^*\right)$$

  - Here, $\frac{m}{L} \in (0, 1)$
  - This is, loosely speaking, faster than what we got using only Lipschitz continuity, which was: $f(\mathbf{x}^k) - p^* \leq \frac{\left\| x^{(0)} - \mathbf{x}^* \right\|^2}{2tk}$ (sublinear convergence)

- To obtain $f(\mathbf{x}^k) - p^* \leq \epsilon$, we need $O(log(1/\epsilon))$ iterations
- Linear convergence $\Rightarrow$ plot of iterations on the x-axis, and distance of function value from p^* on the y-axis on a log scale is linear

R-convergence assuming Strong convexity and L-continuity

- Now, let us consider the convergence result we got by assuming Strong convexity with backtracking and exact line searches:

$$f(x^k) - f(x^*) \leq \left(1 - \frac{m}{M}\right)^k \left(f(x^{(0)}) - f(x^*)\right)$$

- Here, v^k can be considered $\left(1 - \frac{m}{M}\right)^k \alpha$
  - $v^* = 0$
- We get

$$\frac{v^{k+1} - v^*}{v^k - v^*} = \left(1 - \frac{m}{M}\right) \in (0, 1)$$

  - We now have an upper bound < 1, unlike before
- As $r = \left(1 - \frac{m}{M}\right) \in (0, 1)$, v^k is Q-linearly convergent
  - Thus, under strong convexity, gradient descent is R-linearly convergent

R-convergence assuming Strong convexity and L-continuity

- *Question:* Is gradient descent under Strong convexity also Q-linearly convergent?
- Recall one of the intermediate steps in getting the convergence results:
  $f(x^{k+1}) - f(x^*) \leq \left((1 - \frac{m}{M}\right) \left(f(x^k) - f(x^*)\right)$
  - $\implies \frac{f(x^{k+1}) - f(x^*)}{f(x^k) - f(x^*)} \leq \left(1 - \frac{m}{M}\right)$
- Now, $r = \left(1 - \frac{m}{M}\right) \in (0, 1)$
- Yes, gradient descent under Strong convexity is also Q-linearly convergent

Summary of Convergence Rate of Gradient Descent Method

- For the gradient method, it can be proved that if f is strongly convex,

$$f(\mathbf{x}^{(k)}) - p^* \leq \rho^k \left(f(\mathbf{x}^{(0)} - p^* \right) \tag{49}$$

  The value of the linear convergence factor $\rho \in (0,1)$ depends on the strong convexity constant c, the value of $\mathbf{x}^{(0)}$ and type of ray search employed.
- The convergence rate is $1 - m/L$, where L/m is proportional to the *condition number* of the Hessian. Large eigenvalues correspond to high curvature directions and small eigenvalues correspond to low curvature directions. Many methods (such as conjugate gradient) try to improve upon the gradient method by making the hessian better conditioned. Convergence can be very slow even for moderately well-conditioned problems, with condition number in the 100s.
- The convergence of the steepest descent method can be stated in the same form as in (49), since any norm can be bounded in terms of the Euclidean norm, *i.e.*, there exists a constant $\eta \in (0,1]$ such that $||\mathbf{x}|| \geq \eta ||\mathbf{x}||_2$ (see Section 9.4.3 of Boyd)

(Sub)Gradient Descent: Generalization of Gradient Descent

Given a convex function $f\colon \mathbb{R}^n \to \mathbb{R}$, not necessarily differentiable. Subgradient method is just like gradient descent, but replacing gradients with subgradients. *I.e.*, initialize $x^{(0)}$, then repeat

$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} - t^k \cdot \mathbf{h}^{(k-1)}, k = 1, 2, 3, \cdots$$

easy to check that the necessary descent condition holds

where $\mathbf{h}^{(k-1)}$ is **any** subgradient of f at $\mathbf{x}^{(k-1)}$. We keep track of best iterate $\mathbf{x}^k_{best}$ among $\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(k)}$:

$$f(\mathbf{x}^{(k)}_{best}) = \min_{i=1,\cdots,k} f(\mathbf{x}^{(i)})$$

important given our step sizes are fixed and given that multiple subgradients exist of which our choice might have been random

To update each $\mathbf{x}^{(i)}$, there are basically two ways to select the step size:

- Fixed step size: $t^k = t$ for all $k = 1, 2, 3 \cdots$
- Diminishing step size: choose t^k to satisfy

$$\lim_{k\to\infty} (t^k) = 0\ , \quad \sum_{k=1}^{\infty} t^k = \infty$$

https://en.wikipedia.org/wiki/Series_(mathematics)

See a more detailed derivation at https://youtu.be/1_3zQtH-w4U?list=PLsd82ngobrvcYfCdnSnqM71KLqE9qUUpX&t=4503

Subgradient Algorithm: Convergence analysis

Given the convex function $f : \mathbb{R}^n \to \mathbb{R}$ that satisfies:

- f is Lipschitz continuous with constant $l > 0$,

$$|f(\mathbf{x}) - f(\mathbf{y})| \leq l||\mathbf{x} - \mathbf{y}|| \text{ for all } \mathbf{x}, \mathbf{y}$$

- $||\mathbf{x}^{(1)} - \mathbf{x}^*|| \leq R$ which means it is bounded

Theorem

For a fixed step size t, subgradient method satisfies

$$\lim_{k\to\infty} f(x_{best}^{(k)}) \leq f(\mathbf{x}^*) + \frac{l^2 t}{2}$$

For diminishing step size such as $t^k = O\left(\frac{1}{\sqrt{k}}\right)$,

$$f(x_{best}^{(k)}) \leq f(\mathbf{x}^*) + O\left(\frac{1}{\sqrt{k}}\right)$$

much worse than O(1/k)

Subgradient Descent: Convergence Analysis (contd.)

Proof:

$$\begin{aligned} \|\mathbf{x}^{(k+1)} - \mathbf{x}^*\|^2 &= \|\mathbf{x}^{(k)} - t^k\mathbf{h}^{(k)} - \mathbf{x}^*\|^2 \\ &= \|\mathbf{x}^{(k)} - \mathbf{x}^*\|^2 - 2t^k(\mathbf{h}^{(k)})^T(\mathbf{x}^{(k)} - \mathbf{x}^*) + (t^k)^2\|\mathbf{h}^{(k)}\|^2 \end{aligned}$$

By definition of the subgradient method, we have

$$f(\mathbf{x}^*) \geq f(\mathbf{x}^{(k)}) + (\mathbf{h}^{(k)})^T(\mathbf{x}^* - \mathbf{x}^{(k)})$$

$$-(\mathbf{h}^{(k)})^T(\mathbf{x}^* - \mathbf{x}^{(k)}) \leq -(f(\mathbf{x}^{(k)}) - f(\mathbf{x}^*))$$

Using this inequality, for $k, k-1, \dots i, i-1, \dots 0$ we have

Subgradient Descent: Convergence Analysis (contd.)

Proof:

$$
\begin{aligned}
||\mathbf{x}^{(k+1)} - \mathbf{x}^*||^2 &= ||\mathbf{x}^{(k)} - t^k\mathbf{h}^{(k)} - \mathbf{x}^*||^2 \\
&= ||\mathbf{x}^{(k)} - \mathbf{x}^*||^2 - 2t^k(\mathbf{h}^{(k)})^T(\mathbf{x}^{(k)} - \mathbf{x}^*) + (t^k)^2||\mathbf{h}^{(k)}||^2
\end{aligned}
$$

By definition of the subgradient method, we have

$$
f(\mathbf{x}^*) \geq f(\mathbf{x}^{(k)}) + (\mathbf{h}^{(k)})^T(\mathbf{x}^* - \mathbf{x}^{(k)})
$$
$$
-(\mathbf{h}^{(k)})^T(\mathbf{x}^* - \mathbf{x}^{(k)}) \leq -(f(\mathbf{x}^{(k)}) - f(\mathbf{x}^*))
$$

Using this inequality, for $k, k-1, \ldots i, i-1, \ldots 0$ we have

$$
\begin{aligned}
||\mathbf{x}^{(k+1)} - \mathbf{x}^*||^2 &\leq ||\mathbf{x}^{(k)} - \mathbf{x}^*||^2 - 2t^k(f(\mathbf{x}^{(k)}) - f(\mathbf{x}^*)) + (t^k)^2||\mathbf{h}^{(k)}||^2 \\
&\leq ||\mathbf{x}^{(1)} - \mathbf{x}^*||^2 - 2\sum_{i=1}^{k} t^i(f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*)) + \sum_{i=1}^{k}(t^i)^2||\mathbf{h}^{(i)}||^2
\end{aligned}
$$

Subgradient Descent: Convergence Analysis (contd.)

And since this is lower bounded by 0, we have

$$0 \leq ||\mathbf{x}^{(k+1)} - \mathbf{x}^*||^2 \leq R^2 - 2\sum_{i=1}^{k} t^i(f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*)) + \sum_{i=1}^{k}(t^i)^2 l^2$$

$$\Rightarrow 2\sum_{i=1}^{k} t^i(f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*)) \leq R^2 + \sum_{i=1}^{k}(t^i)^2 l^2$$

$$\Rightarrow 2(\sum_{i=1}^{k} t^i)(f(\mathbf{x}_{best}^{(k)}) - f(\mathbf{x}^*)) \leq R^2 + \sum_{i=1}^{k}(t^i)^2 l^2$$

Subgradient Descent: Convergence Analysis (contd.)

For a constant step size $t^i = t$:

$$\frac{R^2 + l^2 t^2 k}{2tk} \rightarrow \frac{l^2 t}{2}, \text{ as } k \rightarrow \infty,$$

and for diminishing step size, we have:

Subgradient Descent: Convergence Analysis (contd.)

For a constant step size $t^i = t$:

$$\frac{R^2 + l^2 t^2 k}{2tk} \to \frac{l^2 t}{2}, \text{ as } k \to \infty,$$

and for diminishing step size, we have:

$$\sum_{i=0}^{k} (t^i)^2 \leq 0, \sum_{i=0}^{k} t^i = \infty$$

therefore,

$$\frac{R^2 + l^2 \sum_{i=0}^{k} (t^i)^2}{2 \sum_{i=0}^{k} t^i} \to 0, \text{ as } k \to \infty,$$

□

Subgradient Descent: Convergence Analysis (contd.)

Consider taking $t^i = R/(l\sqrt{k})$, for all $i = 1, ..., k$. Then we can obtain the following tendency:

Subgradient Descent: Convergence Analysis (contd.)

Consider taking $t^i = R/(l\sqrt{k})$, for all $i = 1, ..., k$. Then we can obtain the following tendency:

$$\frac{R^2 + l^2 \sum_{i=0}^{k}(t^i)^2}{2\sum_{i=0}^{k} t^i} = \frac{Rl}{\sqrt{k}}.$$

That is, subgradient method has convergence rate of $O(\frac{1}{\sqrt{k}})$, and to get $f(x_{best}^{(k)}) - f(x^*) \leq \epsilon$, needs $O(\frac{1}{\epsilon^2})$ iterations.

This is a much worse convergence rate than even $O\left(\frac{1}{k}\right)$ obtained for gradient descent under Lipschitz continuity alone.

Optimization: Subgradient Descent and Constrained Optimization

Instructor: Prof. Ganesh Ramakrishnan

Constrained Optimization in $\Re$: Recap

Global Extrema on Closed Intervals

Recall the extreme value theorem. A consequence is that:

- if either of c or d lies in (a, b), then it is a critical number of f;
- else each of c and d must lie on one of the boundaries of $[a, b]$.

This gives us a procedure for finding the maximum and minimum of a continuous function f on a closed bounded interval $\mathcal{I}$:

Procedure

[Finding extreme values on closed, bounded intervals]:

1. *Find the critical points in int$(\mathcal{I})$.*
2. *Compute the values of f at the critical points and at the endpoints of the interval.*
3. *Select the least and greatest of the computed values.*

Global Extrema on Closed Intervals (contd)

- To compute the maximum and minimum values of $f(x) = 4x^3 - 8x^2 + 5x$ on the interval $[0, 1]$,

Global Extrema on Closed Intervals (contd)

- To compute the maximum and minimum values of $f(x) = 4x^3 - 8x^2 + 5x$ on the interval $[0, 1]$,
  - We first compute $f'(x) = 12x^2 - 16x + 5$ which is 0 at $x = \frac{1}{2}, \frac{5}{6}$.
  - Values at the critical points are $f(\frac{1}{2}) = 1$, $f(\frac{5}{6}) = \frac{25}{27}$.
  - The values at the end points are $f(0) = 0$ and $f(1) = 1$.
  - Therefore, the minimum value is $f(0) = 0$ and the maximum value is $f(1) = f(\frac{1}{2}) = 1$.

Global Extrema on Closed Intervals (contd)

- To compute the maximum and minimum values of $f(x) = 4x^3 - 8x^2 + 5x$ on the interval $[0, 1]$,
  - We first compute $f'(x) = 12x^2 - 16x + 5$ which is 0 at $x = \frac{1}{2}, \frac{5}{6}$.
  - Values at the critical points are $f(\frac{1}{2}) = 1$, $f(\frac{5}{6}) = \frac{25}{27}$.
  - The values at the end points are $f(0) = 0$ and $f(1) = 1$.
  - Therefore, the minimum value is $f(0) = 0$ and the maximum value is $f(1) = f(\frac{1}{2}) = 1$.
- In this context, it is relevant to discuss the one-sided derivatives of a function at the endpoints of the closed interval on which it is defined.

Global Extrema on Closed Intervals (contd)

Definition

[One-sided derivatives at endpoints]: *Let f be defined on a closed bounded interval* $[a, b]$. *The (right-sided) derivative of f at* $x = a$ *is defined as*

$$f'(a) = \lim_{h \to 0^+} \frac{f(a+h) - f(a)}{h}$$

Similarly, the (left-sided) derivative of f at $x = b$ *is defined as*

$$f'(b) = \lim_{h \to 0^-} \frac{f(b+h) - f(b)}{h}$$

Essentially, each of the one-sided derivatives defines one-sided slopes at the endpoints.

Global Extrema on Closed Intervals (contd)

Based on these definitions, the following result can be derived.

Claim

If f is continuous on $[a, b]$ and $f'(a)$ exists as a real number or as $\pm\infty$, then we have the following necessary conditions for extremum at a.

- *If $f(a)$ is the maximum value of f on $[a, b]$, then $f'(a) \leq 0$ or $f'(a) = -\infty$.*
- *If $f(a)$ is the minimum value of f on $[a, b]$, then $f'(a) \geq 0$ or $f'(a) = \infty$.*

If f is continuous on $[a, b]$ and $f'(b)$ exists as a real number or as $\pm\infty$, then we have the following necessary conditions for extremum at b

Global Extrema on Closed Intervals (contd)

Based on these definitions, the following result can be derived.

Claim

If f is continuous on $[a, b]$ and $f'(a)$ exists as a real number or as $\pm\infty$, then we have the following necessary conditions for extremum at a.

- *If $f(a)$ is the maximum value of f on $[a, b]$, then $f'(a) \leq 0$ or $f'(a) = -\infty$.*
- *If $f(a)$ is the minimum value of f on $[a, b]$, then $f'(a) \geq 0$ or $f'(a) = \infty$.*

If f is continuous on $[a, b]$ and $f'(b)$ exists as a real number or as $\pm\infty$, then we have the following necessary conditions for extremum at b

- *If $f(b)$ is the maximum value of f on $[a, b]$, then $f'(b) \geq 0$ or $f'(b) = \infty$.*
- *If $f(b)$ is the minimum value of f on $[a, b]$, then $f'(b) \leq 0$ or $f'(b) = -\infty$.*

Global Extrema on Closed Intervals (contd)

The following result gives a useful procedure for finding **extrema on closed intervals**.

Claim

If f is continuous on $[a, b]$ and $f''(x)$ exists for all $x \in (a, b)$. Then,

- *If $f''(x) \leq 0, \ \forall x \in (a, b)$, then the minimum value of f on $[a, b]$ is either $f(a)$ or $f(b)$. If, in addition, f has a critical point $c \in (a, b)$, then $f(c)$ is the maximum value of f on $[a, b]$.*
- *If $f''(x) \geq 0, \ \forall x \in (a, b)$, then the maximum value of f on $[a, b]$ is either $f(a)$ or $f(b)$. If, in addition, f has a critical point $c \in (a, b)$, then $f(c)$ is the minimum value of f on $[a, b]$.*

Global Extrema on Open Intervals

The next result is very useful for finding **extrema on open intervals.**

Claim

Let $\mathcal{I}$ be an open interval and let $f'(x)$ exist $\forall x \in \mathcal{I}$.

- *If $f'(x) \geq 0$, $\forall x \in \mathcal{I}$, and if there is a number $c \in \mathcal{I}$ where $f'(c) = 0$, then $f(c)$ is the global minimum value of f on $\mathcal{I}$.*
- *If $f'(x) \leq 0$, $\forall x \in \mathcal{I}$, and if there is a number $c \in \mathcal{I}$ where $f'(c) = 0$, then $f(c)$ is the global maximum value of f on $\mathcal{I}$.*

For example, let $f(x) = \frac{2}{3}x - \sec x$ and
$\mathcal{I} = \left(\frac{-\pi}{2}, \frac{\pi}{2}\right)$.

Global Extrema on Open Intervals

The next result is very useful for finding **extrema on open intervals.**

Claim

Let $\mathcal{I}$ be an open interval and let $f'(x)$ exist $\forall x \in \mathcal{I}$.

- *If $f''(x) \geq 0,\ \forall x \in \mathcal{I}$, and if there is a number $c \in \mathcal{I}$ where $f'(c) = 0$, then $f(c)$ is the global minimum value of f on $\mathcal{I}$.*
- *If $f''(x) \leq 0,\ \forall x \in \mathcal{I}$, and if there is a number $c \in \mathcal{I}$ where $f'(c) = 0$, then $f(c)$ is the global maximum value of f on $\mathcal{I}$.*

For example, let $f(x) = \frac{2}{3}x - \sec x$ and $\mathcal{I} = (\frac{-\pi}{2}, \frac{\pi}{2})$.$f'(x) = \frac{2}{3} - \sec x \tan x = \frac{2}{3} - \frac{\sin x}{\cos^2 x} = 0 \Rightarrow x = \frac{\pi}{6}$. Further, $f''(x) = -\sec x(\tan^2 x + \sec^2 x) < 0$ on $(\frac{-\pi}{2}, \frac{\pi}{2})$. Therefore, f attains the maximum value $f(\frac{\pi}{6}) = \frac{\pi}{9} - \frac{2}{\sqrt{3}}$ on $\mathcal{I}$.

Global Extrema on Open Intervals (contd)

As another example, let us find the dimensions of the cone with minimum volume that can contain a sphere with radius R. Let h be the height of the cone and r the radius of its base. The objective to be minimized is the volume $f(r,h) = \frac{1}{3}\pi r^2 h$. The constraint betwen r and h is shown in Figure 10. The traingle AEF is similar to traingle ADB and therefore, $\frac{h-R}{R} = \frac{\sqrt{h^2+r^2}}{r}$.

Global Extrema on Open Intervals (contd)

Our first step is to reduce the volume formula to involve only one of r[8] or h.
The algebra involved will be the simplest if we solved for h.
The constraint gives us $r^2 = \frac{R^2 h}{h-2R}$. Substituting this expression for r^2 into the volume formula, we get $g(h) = \frac{\pi R^2}{3}\frac{h^2}{(h-2R)}$ with the domain given by $\mathcal{D} = \left\{h|2R < h < \infty\right\}$.
Note that $\mathcal{D}$ is an open interval.
$g' = \frac{\pi R^2}{3}\frac{2h(h-2R)-h^2}{(h-2R)^2} = \frac{\pi R^2}{3}\frac{h(h-4R)}{(h-2R)^2}$ which is 0 in its domain $\mathcal{D}$ if and only if $h = 4R$.
$g'' = \frac{\pi R^2}{3}\frac{2(h-2R)^3-2h(h-4R)(h-2R)^2}{(h-2R)^4} = \frac{\pi R^2}{3}\frac{2(h^2-4Rh+4R^2-h^2+4Rh)}{(h-2R)^3} = \frac{\pi R^2}{3}\frac{8R^2}{(h-2R)^3}$, which is greater than 0 in $\mathcal{D}$.
Therefore, g (and consequently f) has a unique minimum at $h = 4R$ and correspondingly, $r^2 = \frac{R^2 h}{h-2R} = 2R^2$.

[8] Since r appears in the volume formula only in terms of r^2.