PAGES 216 TO 231 OF http://www.cse.iitb.ac.in/~cs709/notes/BasicsOfConvexOptimization.pdf, interspersed with pages between 239 and 253 and summary of material thereafter, which extend univariate concepts to generic spaces

Maximum and Minimum values of univariate functions

Let f be a function with domain $\mathcal{D}$. Then f has an *absolute maximum* (or global maximum) value at point $c \in \mathcal{D}$ if

$$f(x) \leq f(c), \ \forall x \in \mathcal{D}$$

and an *absolute minimum* (or global minimum) value at $c \in \mathcal{D}$ if

$$f(x) \geq f(c), \ \forall x \in \mathcal{D}$$

If there is an open interval $\mathcal{I}$ containing c in which $f(c) \geq f(x)$, $\forall x \in \mathcal{I}$, then we say that $f(c)$ is a *local maximum value* of f. On the other hand, if there is an open interval $\mathcal{I}$ containing c in which $f(c) \leq f(x)$, $\forall x \in \mathcal{I}$, then we say that $f(c)$ is a *local minimum value* of f. If $f(c)$ is either a local maximum or local minimum value of f in an open interval $\mathcal{I}$ with $c \in \mathcal{I}$, the $f(c)$ is called a *local extreme value* of f.

Theorem 39 *If $f(c)$ is a local extreme value and if f is differentiable at $x = c$, then $f'(c) = 0$.* → If all p.d.s of f exist at $x = c \in \mathcal{D} \subseteq \mathbb{R}^n$ & if $f(c)$ is local extreme, $\nabla f(c) = 0$

Theorem 40 *A continuous function $f(x)$ on a closed and bounded interval $[a, b]$ attains a minimum value $f(c)$ for some $c \in [a, b]$ and a maximum value $f(d)$ for some $d \in [a, b]$. That is, a continuous function on a closed, bounded interval attains a minimum and a maximum value.*

Note: $[a, \infty)$ is closed but NOT bounded
So both conditions are needed

replace with sets for $\mathbb{R}^n$

FOR $\mathbb{R}^n$

Theorem 60 *If $f(\mathbf{x})$ defined on a domain $\mathcal{D} \subseteq \Re^n$ has a local maximum or minimum at $\mathbf{x}^*$ and if the first-order partial derivatives exist at $\mathbf{x}^*$, then $f_{x_i}(\mathbf{x}^*) = 0$ for all $1 \leq i \leq n$.*

ie $\nabla f(x^*) = 0$

Definition 27 [Critical point]: *A point $\mathbf{x}^*$ is called a critical point of a function $f(\mathbf{x})$ defined on $\mathcal{D} \subseteq \Re^n$ if*

1. *If $f_{x_i}(\mathbf{x}^*) = 0$, for $1 \leq i \leq n$.*
2. *OR $f_{x_i}(\mathbf{x}^*)$ fails to exist for any $1 \leq i \leq n$.*

A procedure for computing all critical points of a function f is:

1. Compute f_{x_i} for $1 \leq i \leq n$.

2. Determine if there are any points where any one of f_{x_i} fails to exist. Add such points (if any) to the list of critical points.

3. Solve the system of equations $f_{x_i} = 0$ simultaneously. Add the solution points to the list of saddle points.

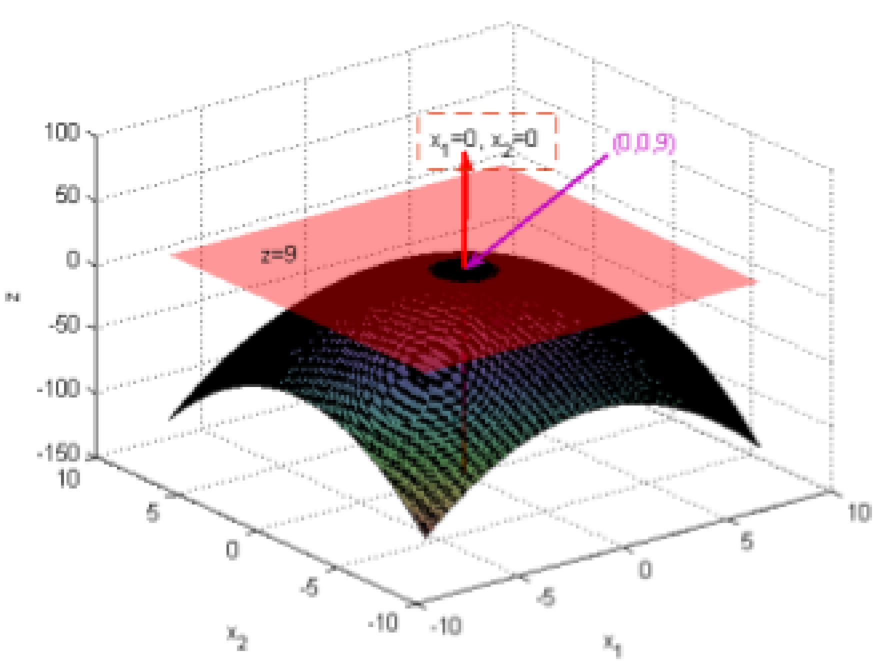


Figure 4.17: The paraboloid $f(x_1, x_2) = 9 - x_1^2 - x_2^2$ attains its maximum at $(0, 0)$. The tangent plane to the surface at $(0, 0, f(0, 0))$ is also shown, and so is

$|x_1|$

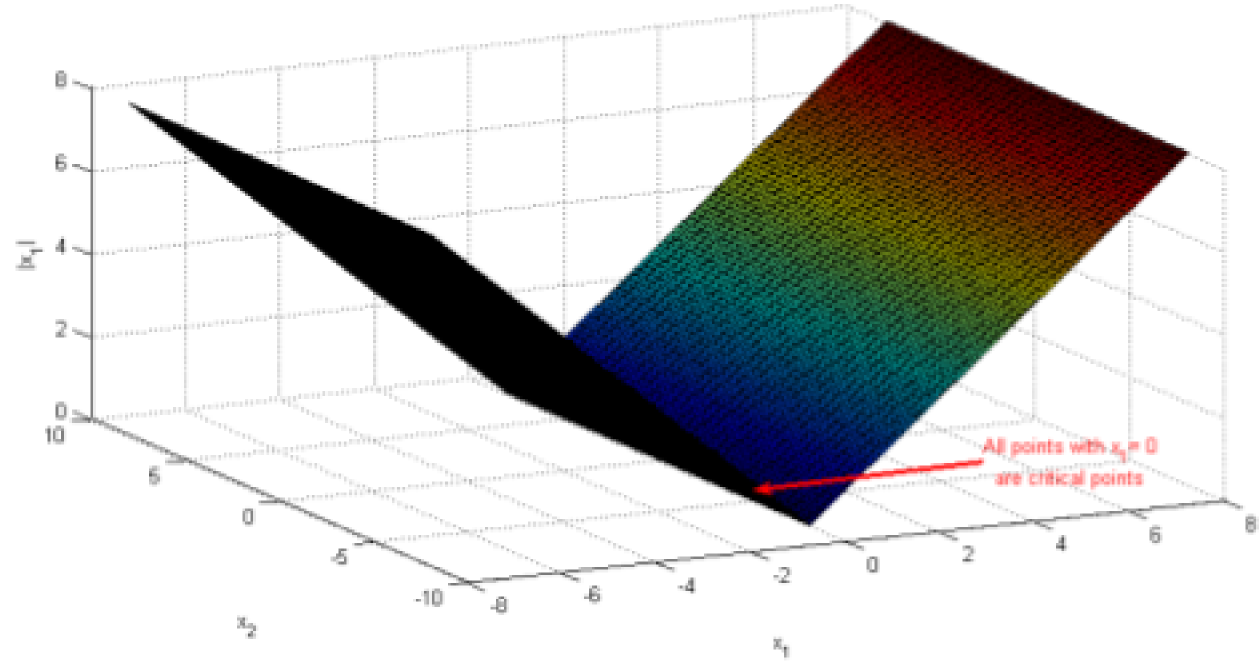


Figure 4.18: Plot illustrating critical points where derivative fails to exist.

Definition 28 [Saddle point]: *A point* $\mathbf{x}^*$ *is called a saddle point of a function* $f(\mathbf{x})$ *defined on* $\mathcal{D} \subseteq \Re^n$ *if* $\mathbf{x}^*$ *is a critical point of* f *but* $\mathbf{x}^*$ *does not correspond to a local maximum or minimum of the function.*

$x_1^2 - x_2^2 \rightarrow$

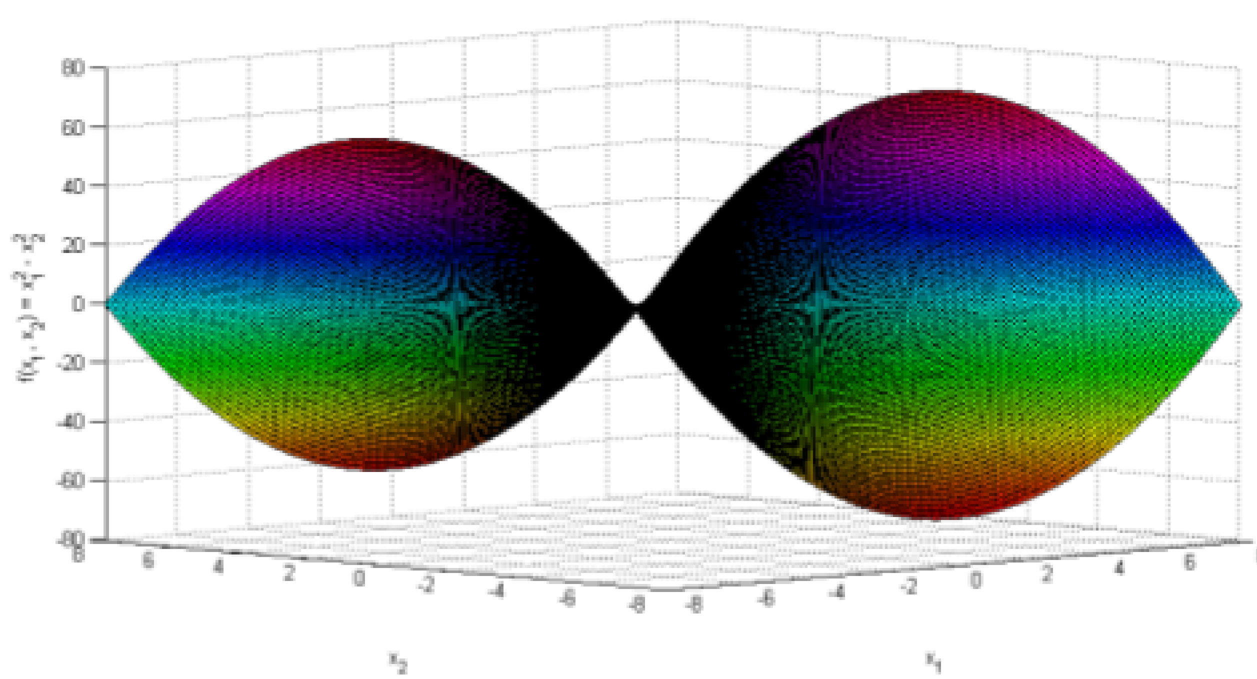


Figure 4.19: The hyperbolic paraboloid $f(x_1, x_2) = x_1^2 - x_2^2$, which has a saddle point at $(0, 0)$.

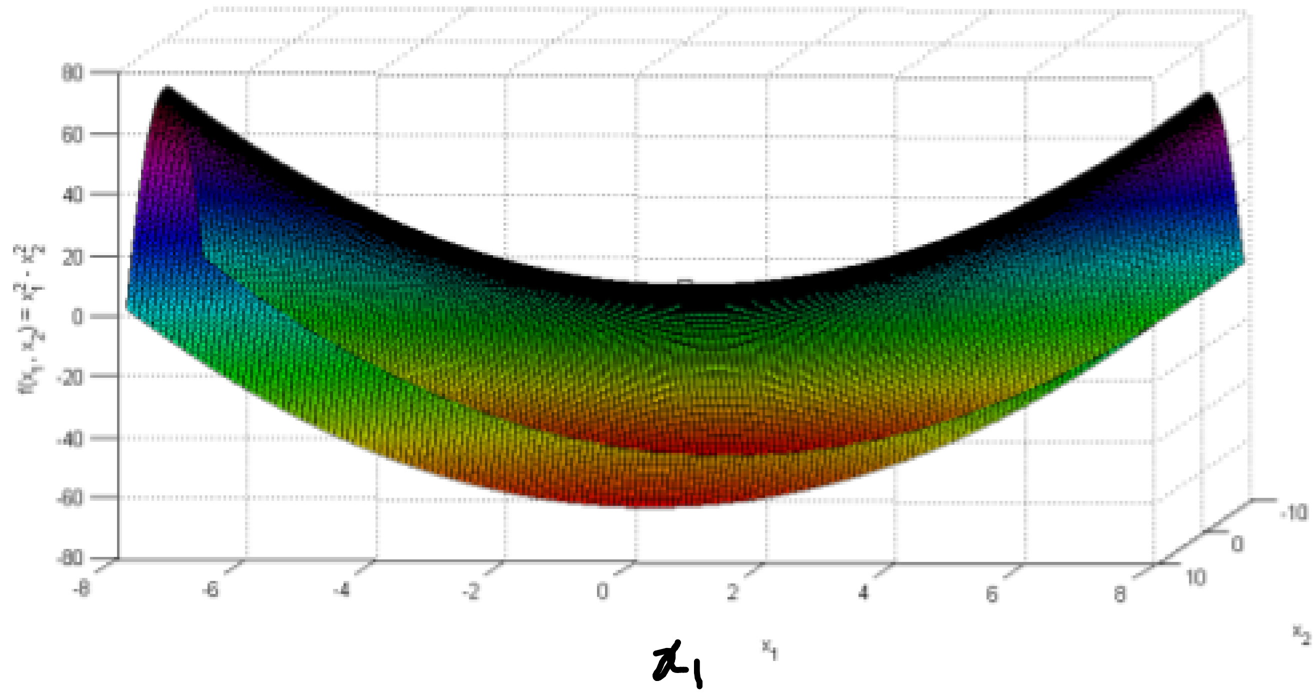


Figure 4.20: The hyperbolic paraboloid $f(x_1, x_2) = x_1^2 - x_2^2$, when viewed from the x_1 axis is concave up.

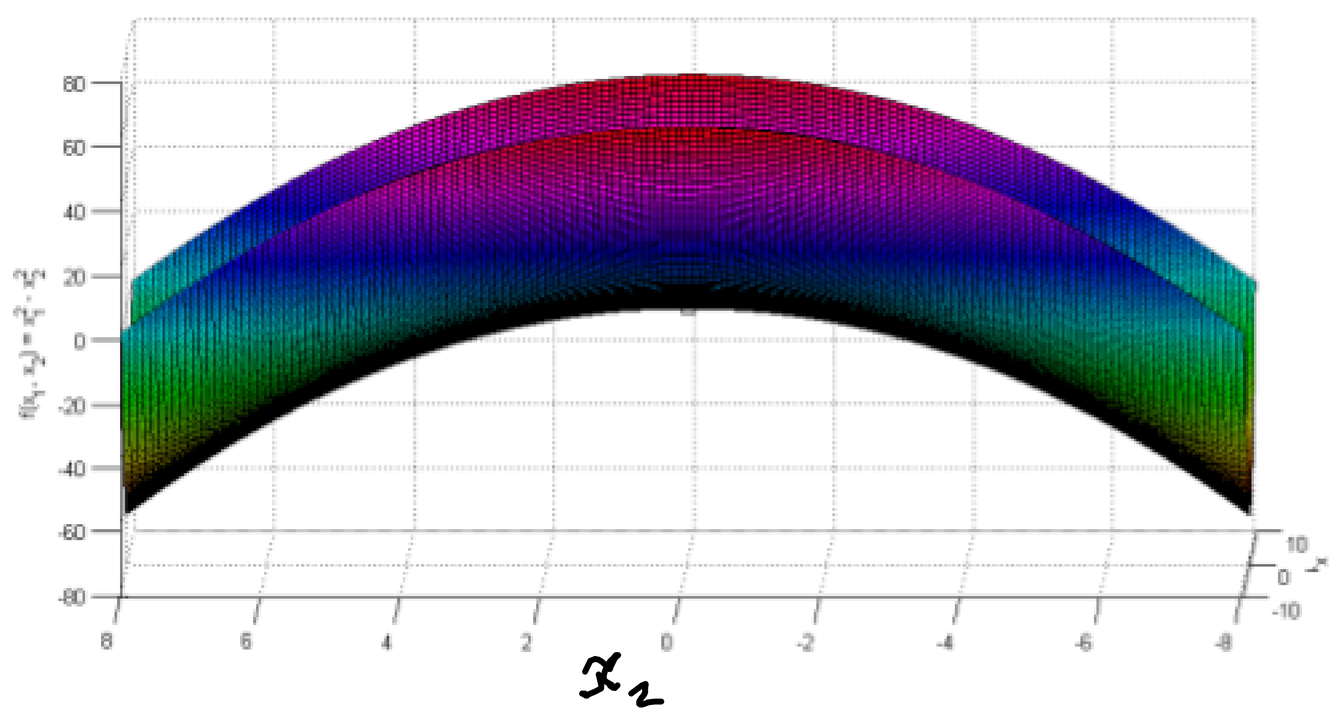
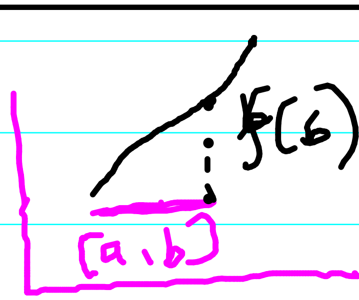


Figure 4.21: The hyperbolic paraboloid $f(x_1, x_2) = x_1^2 - x_2^2$, when viewed from the x_2 axis is concave down.



Note: For LP's, $Ax \geq b$ is closed and bounded D & $f(x) = c^T x$ attains global max/min on bdry of D. ∴ This thm not applicable

Theorem 41 *A continuous function $f(x)$ on a closed and bounded interval $[a, b]$ attains a minimum value $f(c)$ for some $c \in [a, b]$ and a maximum value $f(d)$ for some $d \in [a, b]$. If $a < c < b$ and $f'(c)$ exists, then $f'(c) = 0$. If $a < d < b$ and $f'(d)$ exists, then $f'(d) = 0$.* : If $D \subseteq \mathbb{R}^n$ is closed & bounded & f is cts on D & if global max/min is attained at $c \in \mathrm{Int}(D)$ & f is differentiable at c then $\nabla f(c) = 0$

Theorem 42 *If f is continuous on $[a, b]$ and differentiable at all $x \in (a, b)$ and if $f(a) = f(b)$, then $f'(c) = 0$ for some $c \in (a, b)$.*

Figure 4.1 illustrates Rolle's theorem with an example function $f(x) = 9 - x^2$ on the interval $[-3, +3]$.

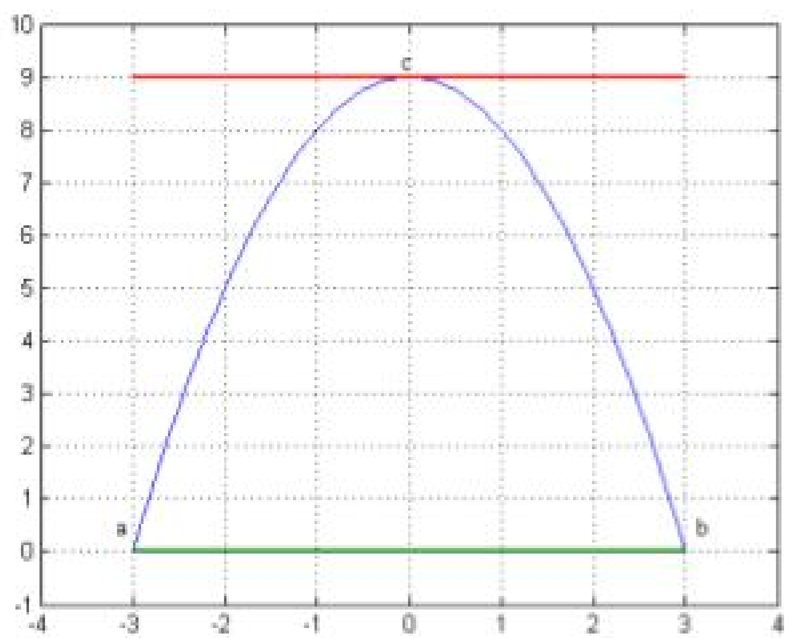


Figure 4.1: Illustration of Rolle's theorem with $f(x) = 9 - x^2$ on the interval $[-3, +3]$. We see that $f'(0) = 0$.

Q: What is a more general version of Rolle's thm?

Ans: Mean value thm

Theorem 43 *If f is continuous on $[a, b]$ and differentiable at all $x \in (a, b)$, then there is some $c \in (a, b)$ such that, $f'(c) = \frac{f(b)-f(a)}{b-a}$.*

If $D \subseteq \mathbb{R}^n$ is closed & bounded & f is cts on D & diff on int(D) then: ?

Turn to next page

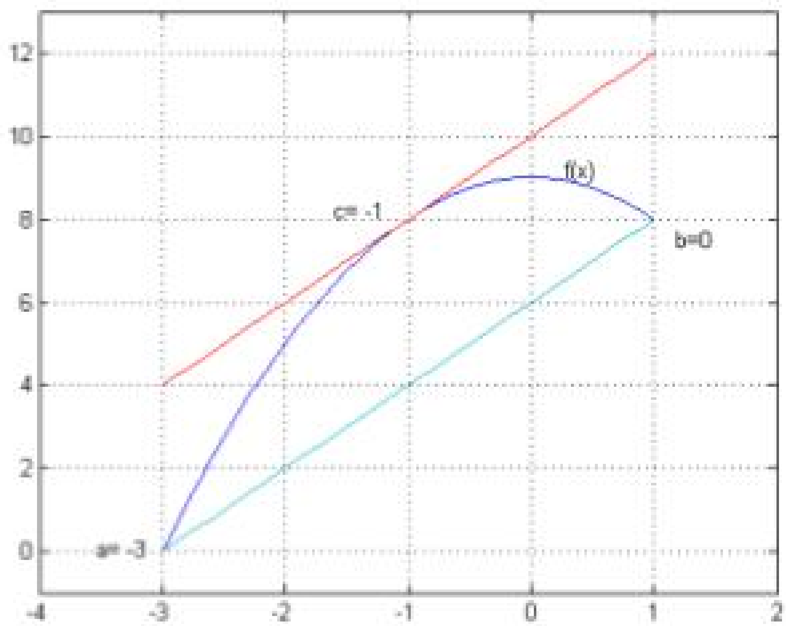


Figure 4.2: Illustration of mean value theorem with $f(x) = 9-x^2$ on the interval $[-3, 1]$. We see that $f'(-1) = \frac{f(1)-f(-3)}{4}$.

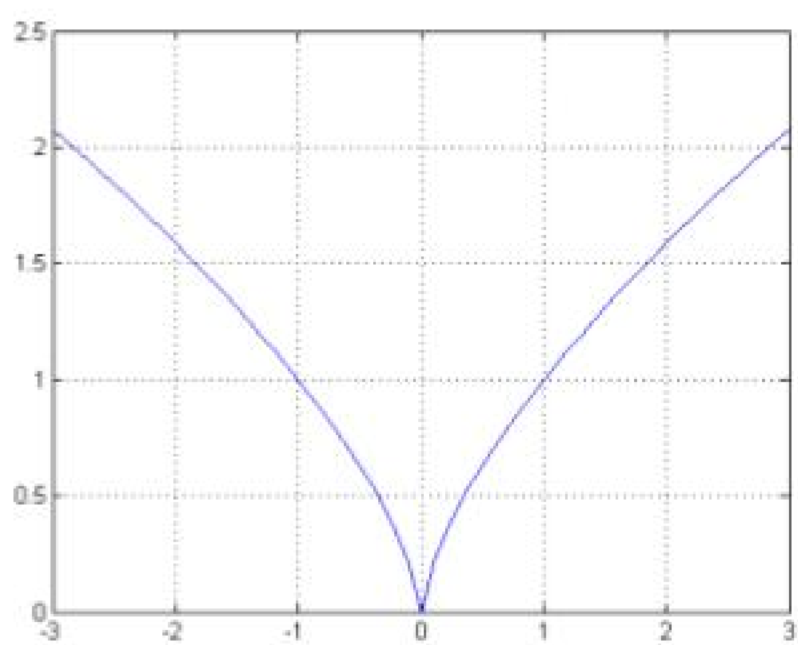


Figure 4.4: The mean value theorem can be violated if $f(x)$ is not differentiable at even a single point of the interval. Illustration on $f(x) = x^{2/3}$ with the

The mean value theorem in one variable generalizes to several variables by applying the theorem in one variable via parametrization. Let G be an open subset of $\mathbf{R}^n$, and let $f : G \to \mathbf{R}$ be a differentiable function. Fix points $x, y \in G$ such that the interval $x\,y$ lies in G, and define $g(t) = f((1 - t)x + ty)$. Since g is a differentiable function in one variable, the mean value theorem gives:

$$g(1) - g(0) = g'(c)$$

for some c between 0 and 1. But since $g(1) = f(y)$ and $g(0) = f(x)$, computing $g'(c)$ explicitly we have:

$$f(y) - f(x) = \nabla f((1 - c)x + cy) \cdot (y - x)$$

Convexity of the domain is fundamental since $\forall t \in [0,1]$, $\boxed{x(1-t) + ty \in \text{Domain}}$

That is, we require convexity of set in some sense

Corollary 44 *Let f be continuous on $[a,b]$ and differentiable on (a,b) with $m \le f'(x) \le M$, $\forall x \in (a,b)$. Then, $m(x-t) \le f(x) - f(t) \le M(x-t)$, if $a \le t \le x \le b$.*

Applying mean value thm & substituting inequality

Let $\mathcal{D}$ be the domain of function f. We define

1. the linear approximation of a differentiable function $f(x)$ as $L_a(x) = f(a) + f'(a)(x-a)$ for some $a \in \mathcal{D}$. We note that $L_a(x)$ and its first derivative at a agree with $f(a)$ and $f'(a)$ respectively.

   $f(a) + f'(c)(x-a)$ for $(c-a)$ is MVT

   $f'(a) = \text{slope}$

2. the quadratic approximatin of a twice differentiable function $f(x)$ as the parabola $Q_a(x) = f(a) + f'(a)(x-a) + \frac{1}{2}f''(a)(x-a)^2$. We note that $Q_a(x)$ and its first and second derivatives at a agree with $f(a)$, $f'(a)$ and $f''(a)$ respectively.

   $P_a(x) = c_1 + c_2 x + c_3 x^2$ s.t. $P_a(a) = f(a)$ $P'_a(a) = f'(a)$ $P''_a(a) = f''(a)$

3. the cubic approximation of a thrice differentiable function $f(x)$ is $C_a(x) = f(a) + f'(a)(x-a) + \frac{1}{2}f''(a)(x-a)^2 + \frac{1}{6}f'''(a)(x-a)^3$. $C_a(x)$ and its first, second and third derivatives at a agree with $f(a)$, $f'(a)$, $f''(a)$ and $f'''(a)$ respectively.

   $R_a(x) = c_1 + c_2 x + c_3 x^2 + c_4 x^3$ s.t. $R_a(a) = f(a)$ $R'_a(a) = f'(a)$, $R''(a) = f''(a)$ $R'''(a) = f'''(a)$

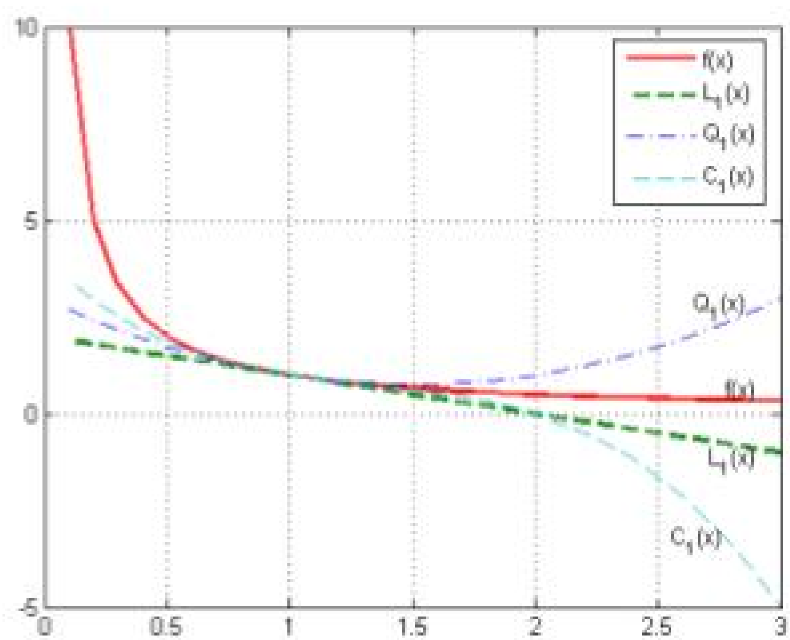


Figure 4.3: Plot of $f(x) = \frac{1}{x}$, and its linear, quadratic and cubic approximations.

can be thought of as general n^{th} order representation of $f(b)$

Theorem 45 *The Taylor's theorem states that if f and its first n derivatives $f', f'', \ldots, f^{(n)}$ are continuous on the closed interval $[a, b]$, and differentiable on (a, b), then there exists a number $c \in (a, b)$ such that*

$$f(b) = \underbrace{f(a) + f'(a)(b-a)}_{\text{MVT is special case}} + \frac{1}{2!}f''(a)(b-a)^2 + \ldots + \frac{1}{n!}f^{(n)}(a)(b-a)^n + \frac{1}{(n+1)!}f^{(n+1)}(c)(b-$$

MVT! $\exists\ c \in (a,b)$ s.t. $f(b) = f(a) + f'(c)(b-a)$ No c in the approximations

To prove use MVT successively on $f(\cdot), f'(\cdot), \ldots f^n(\cdot)$

Consider the function $\phi(t) = f(\mathbf{x} + t\mathbf{h})$ considered in theorem 71, defined on the domain $\mathcal{D}_\phi = [0, 1]$. Using the chain rule,

$$\phi'(t) = \sum_{i=1}^{n} f_{x_i}(\mathbf{x} + t\mathbf{h})\frac{dx_i}{dt} = \mathbf{h}^T.\nabla f(\mathbf{x} + t\mathbf{h})$$

Since f has partial and mixed partial derivatives, ϕ' is a differentiable function of t on $\mathcal{D}_\phi$ and

$$\phi''(t) = \mathbf{h}^T\nabla^2 f(\mathbf{x} + t\mathbf{h})\mathbf{h}$$

Since ϕ and ϕ' are continous on $\mathcal{D}_\phi$ and ϕ' is differentiable on $int(\mathcal{D}_\phi)$, we can make use of the Taylor's theorem (45) with $n = 3$ to obtain:

$$\phi(t) = \phi(0) + t.\phi'(0) + t^2.\frac{1}{2}\phi''(0) + O(t^3)$$

Writing this equation in terms of f gives

is neglected for second order approx

$$f(\underbrace{\mathbf{x} + t\mathbf{h}}_{y}) = f(\mathbf{x}) + t\mathbf{h}^T\nabla f(\mathbf{x}) + t^2\frac{1}{2}h^T\nabla^2 f(\mathbf{x})\mathbf{h} + O(t^3)$$

For 2nd order Taylor expansion

replace $\nabla^2 f(x)$ by $\nabla^2 f(x+ch)$ for $c \in (0,t)$

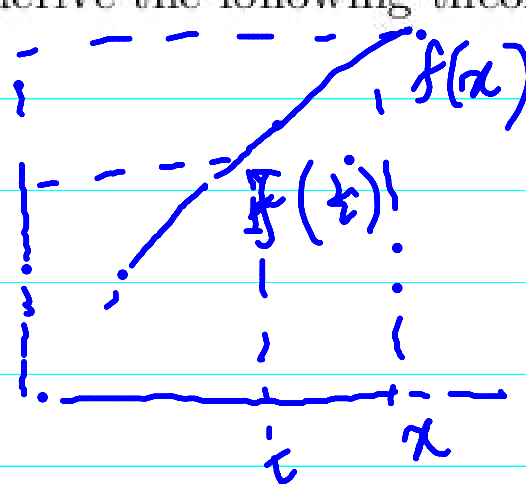
We discussed in class, derivation of the second order Taylor expression.

We ppp also discussed that the matrix $\nabla^2 f$ of mixed partial derivatives is symmetric if f has continuous mixed partial derivatives

We will introduce some definitions at this point:

- A function f is said to be strictly *increasing* on an interval $\mathcal{I}$ in its domain $\mathcal{D}$ if $f(t) < f(x)$ whenever $t < x$.
- The function f is said to be strictly *decreasing* on an interval $\mathcal{I} \in \mathcal{D}$ if $f(t) > f(x)$ whenever $t < x$.

These definitions help us derive the following theorem:



Theorem 46 *Let $\mathcal{I}$ be an interval and suppose f is continuous on $\mathcal{I}$ and differentiable on $int(\mathcal{I})$. Then:*

1. *if $f'(x) > 0$ for all $x \in int(\mathcal{I})$, then f is increasing on $\mathcal{I}$;* → Sufficient
2. *if $f'(x) < 0$ for all $x \in int(\mathcal{I})$, then f is decreasing on $\mathcal{I}$;*
3. *if $f'(x) = 0$ for all $x \in int(\mathcal{I})$, iff, f is constant on $\mathcal{I}$.* → Necessary & sufficient

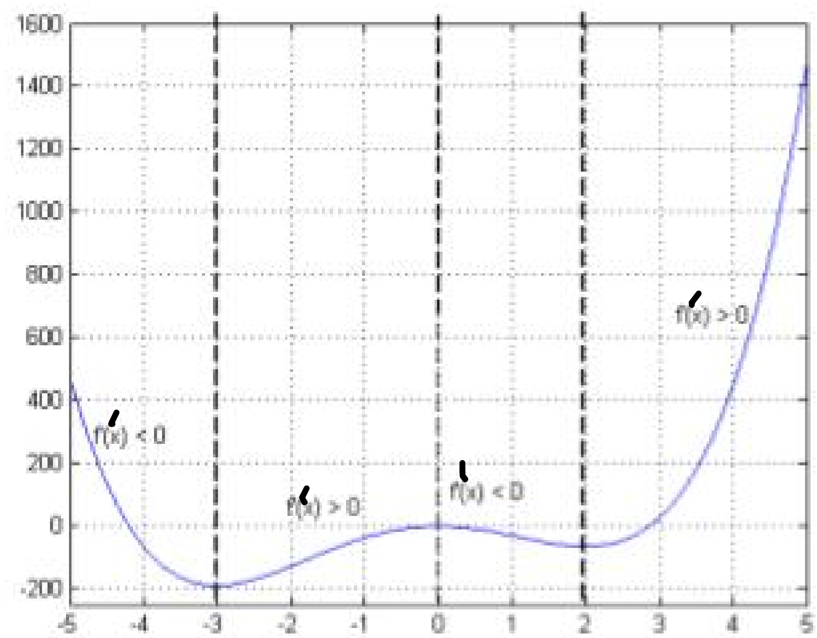


Figure 4.5: Illustration of the increasing and decreasing regions of ~~a function~~ $f(x) = 3x^4 + 4x^3 - 36x^2$

Theorem 47 *Let $\mathcal{I}$ be an interval and suppose f is continuous on $\mathcal{I}$ and differentiable on $int(\mathcal{I})$. Then:*

1. *if $f'(x) \geq 0$ for all $x \in int(\mathcal{I})$, and if $f'(x) = 0$ at only finitely many $x \in \mathcal{I}$, then f is increasing on $\mathcal{I}$;* Necessary
2. *if $f'(x) \leq 0$ for all $x \in int(\mathcal{I})$, and if $f'(x) = 0$ at only finitely many $x \in \mathcal{I}$, then f is decreasing on $\mathcal{I}$.* Necessary

Theorem 48 *Let $\mathcal{I}$ be an interval, and suppose f is continuous on $\mathcal{I}$ and differentiable in $int(\mathcal{I})$. Then:*

1. *if f is increasing on $\mathcal{I}$, then $f'(x) \geq 0$ for all $x \in int(\mathcal{I})$;*
2. *if f is decreasing on $\mathcal{I}$, then $f'(x) \leq 0$ for all $x \in int(\mathcal{I})$.*

Necessary condition for increasing function

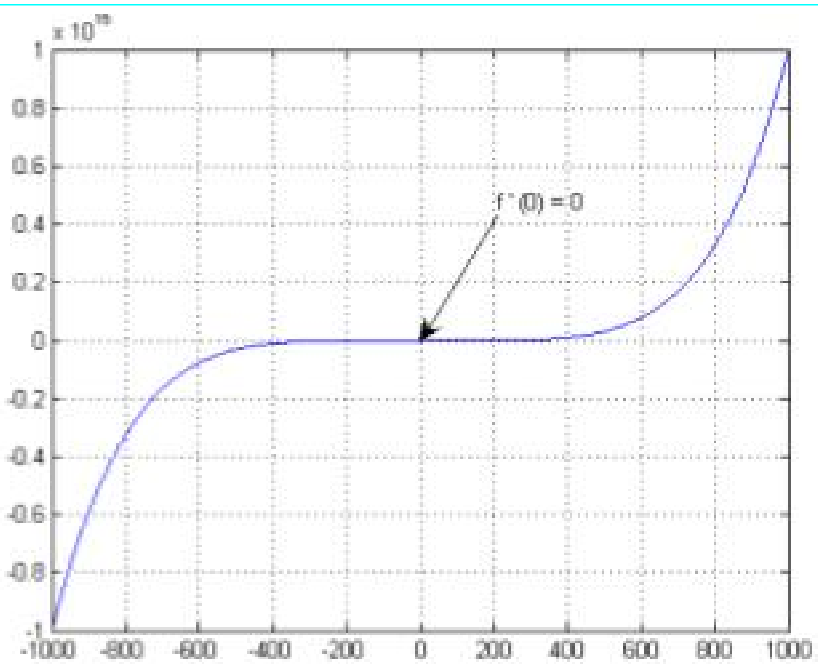


Figure 4.6: Plot of $f(x) = x^5$, illustrating that though the function is increasing on $(-\infty, \infty)$, $f'(0) = 0$.

In summary: $f'(x) \geq 0 \Leftrightarrow f$ is increasing

$f'(x) \geq 0$ & $f'(x) = 0$ at countable # pts $\Leftrightarrow f$ is strictly increasing

Analogous to the definition of increasing functions introduced on page number 220, we next introduce the concept of monotonic functions. This concept is very useful for characterization of a convex function.

The simple case

Definition 39 *Let* $f : \mathcal{D} \to \Re^n$ *and* $\mathcal{D} \subseteq \Re^n$*. Then*

eg: $\nabla f : R^n \to R^n$

if $n=1$ $(f(x_1)-f(x_2))(x_1-x_2) \geq 0$

1. f *is monotone on* $\mathcal{D}$ *if for any* $x_1, x_2 \in \mathcal{D}$,

extension of increasing fn to $f: R^n \to R^n$

$$(f(x_1) - f(x_2))^T (x_1 - x_2) \geq 0 \tag{4.41}$$

2. f *is strictly monotone on* $\mathcal{D}$ *if for any* $x_1, x_2 \in \mathcal{D}$ *with* $x_1 \neq x_2$,

$$(f(x_1) - f(x_2))^T (x_1 - x_2) > 0 \tag{4.42}$$

3. f *is uniformly or strongly monotone on* $\mathcal{D}$ *if for any* $x_1, x_2 \in \mathcal{D}$*, there is a constant* $c > 0$ *such that*

$$\|f(x_1) - f(x_2)\| \, \|x_1 - x_2\| \geq (f(x_1) - f(x_2))^T (x_1 - x_2) \geq c\|x_1 - x_2\|^2 \tag{4.43}$$

For $n=1$, and $D=(a,b)$, this implies (by mean value theorem) that $f'(t) \geq c \;\forall\; t \in (a,b)$ ...

For $n>1$, norm of every row of the Jacobian ($n \times n$ matrix) should be $\geq c$ (verify)

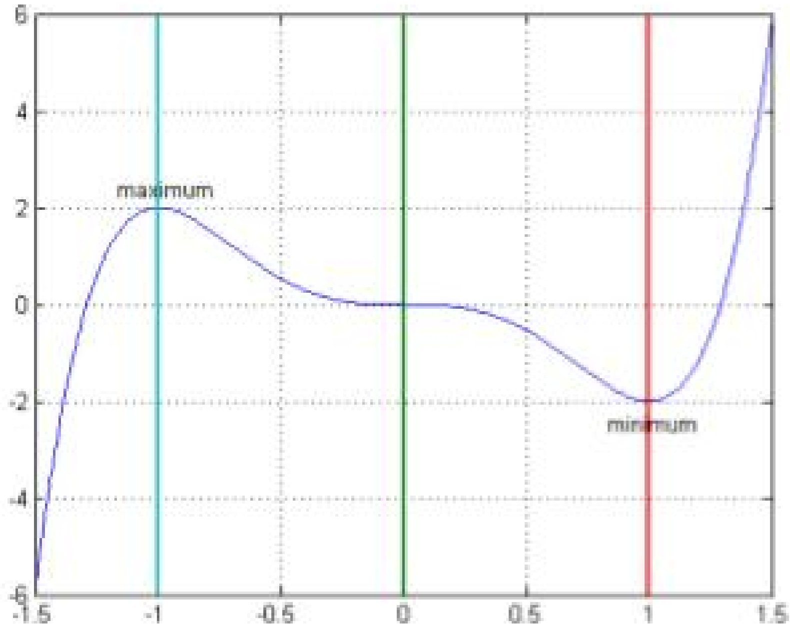


Figure 4.7: Example illustrating the derivative test for function $f(x) = 3x^5 - 5x^3$.

Procedure 1 [First derivative test]: *Let c be an isolated critical number of f. Then,*

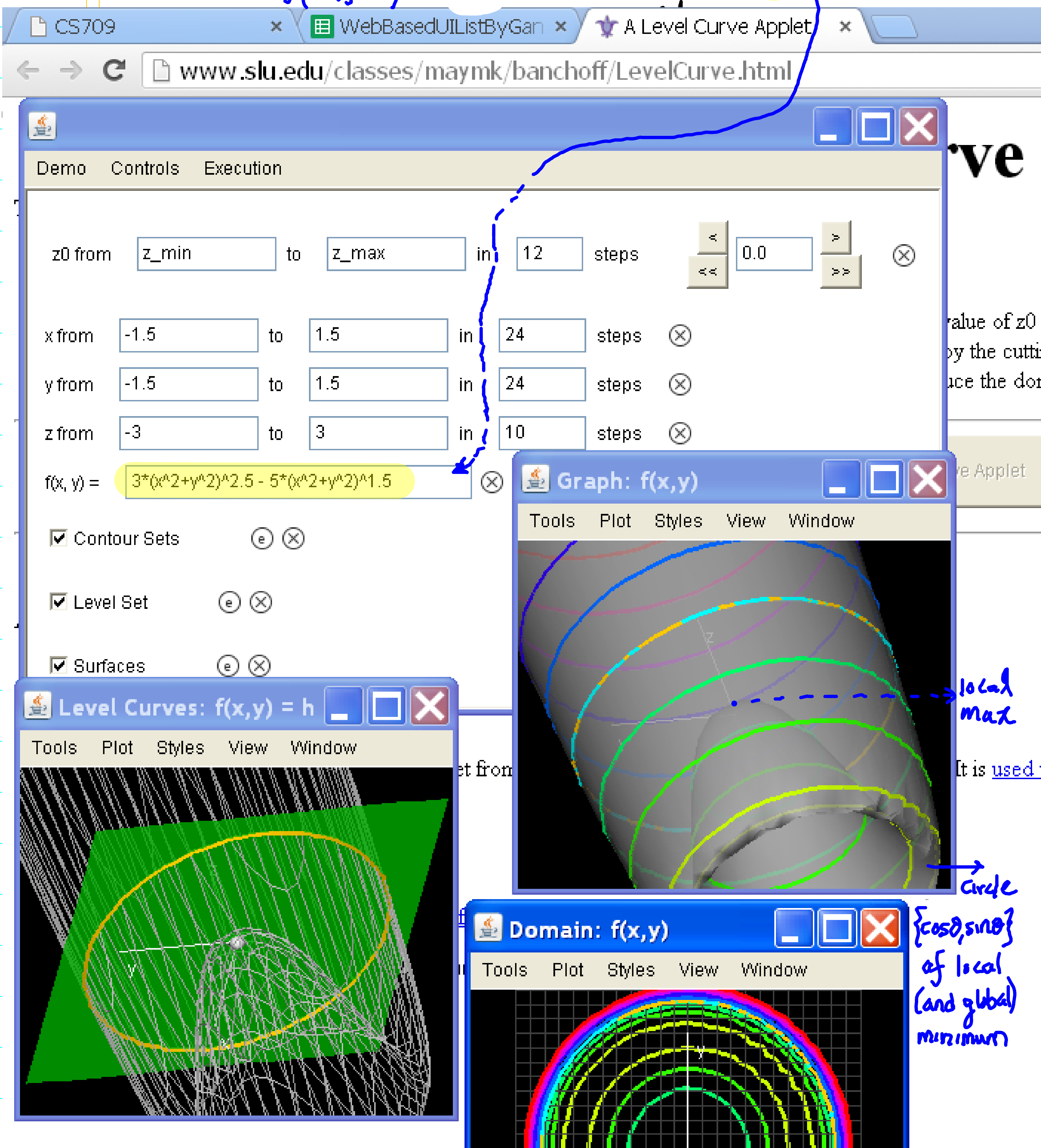
1. *$f(c)$ is a local minimum if $f(x)$ is decreasing in an interval $[c-\epsilon_1, c]$ and increasing in an interval $[c, c+\epsilon_2]$ with $\epsilon_1, \epsilon_2 > 0$, or (but not equivalently), the sign of $f'(x)$ changes from negative in $[c-\epsilon_1, c]$ to positive in $[c, c+\epsilon_2]$ with $\epsilon_1, \epsilon_2 > 0$.*
2. *$f(c)$ is a local maximum if $f(x)$ is increasing in an interval $[c-\epsilon_1, c]$ and decreasing in an interval $[c, c+\epsilon_2]$ with $\epsilon_1, \epsilon_2 > 0$, or (but not equivalently), the sign of $f'(x)$ changes from positive in $[c-\epsilon_1, c]$ to negative in $[c, c+\epsilon_2]$ with $\epsilon_1, \epsilon_2 > 0$.*
3. *If $f'(x)$ is positive in an interval $[c-\epsilon_1, c]$ and also positive in an interval $[c, c-\epsilon_2]$, or $f'(x)$ is negative in an interval $[c-\epsilon_1, c]$ and also negative in an interval $[c, c-\epsilon_2]$ with $\epsilon_1, \epsilon_2 > 0$, then $f(c)$ is not a local extremum.*

Refer to claims on reln betwn increasing/decreasing fn & $f'(x)$

As an example, the function $f(x) = 3x^5 - 5x^3$ has the derivative $f'(x) = 15x^2(x+1)(x-1)$. The critical points are 0, 1 and -1. Of the three, the sign of $f'(x)$ changes at 1 and -1, which are local minimum and maximum respectively.

Extending $f(x) = 3x^5 - 5x^3$ to 2 dimensional i/p space

$$f(x_1, x_2) = 3(x_1^2 + x_2^2)^{5/2} - 5(x_1^2 + x_2^2)^{3/2}$$



local max

circle {cosθ, sinθ} of local (and global) minimum

Procedure 3 [Second derivative test]: *Let c be a critical number of f where $f'(c) = 0$ and $f''(c)$ exists.*

1. *If $f''(c) > 0$ then $f(c)$ is a local minimum.*
2. *If $f''(c) < 0$ then $f(c)$ is a local maximum.*
3. *If $f''(c) = 0$ then $f(c)$ could be a local maximum, a local minimum, neither or both. That is, the test fails.*

For example,

- If $f(x) = x^4$, then $f'(0) = 0$ and $f''(0) = 0$ and we can see that $f(0)$ is a local minimum.
- If $f(x) = -x^4$, then $f'(0) = 0$ and $f''(0) = 0$ and we can see that $f(0)$ is a local maximum.
- If $f(x) = x^3$, then $f'(0) = 0$ and $f''(0) = 0$ and we can see that $f(0)$ is neither a local minimum nor a local maximum. $(0,0)$ is an inflection point in this case.

- If $f(x) = x + 2\sin x$, then $f'(x) = 1 + 2\cos x$. $f'(x) = 0$ for $x = \frac{2\pi}{3}, \frac{4\pi}{3}$, which are the critical numbers. $f''\left(\frac{2\pi}{3}\right) = -2\sin\frac{2\pi}{3} = -\sqrt{3} < 0 \Rightarrow f\left(\frac{2\pi}{3}\right) = \frac{2\pi}{3} + \sqrt{3}$ is a local maximum value. On the other hand, $f''\left(\frac{4\pi}{3}\right) = \sqrt{3} > 0 \Rightarrow f\left(\frac{4\pi}{3}\right) = \frac{4\pi}{3} - \sqrt{3}$ is a local minimum value.
- If $f(x) = x + \frac{1}{x}$, then $f'(x) = 1 - \frac{1}{x^2}$. The critical numbers are $x = \pm 1$. Note that $x = 0$ is not a critical number, even though $f'(0)$ does not exist, because 0 is not in the domain of f. $f''(x) = \frac{2}{x^3}$. $f''(-1) = -2 < 0$ and therefore $f(-1) = -2$ is a local maximum. $f''(1) = 2 > 0$ and therefore $f(1) = 2$ is a local minimum.

Convexity of a function $f : R \to R$

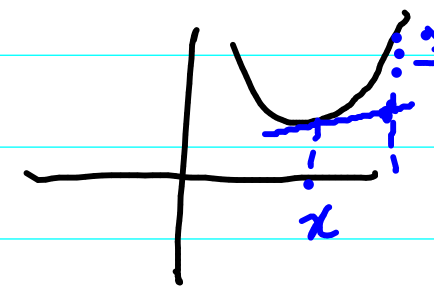
I. f is (strictly) convex iff $\underbrace{f(\theta x_1 + (1-\theta) x_2)}$ $\le$ ($<$) $\theta f(x_1) + (1-\theta) f(x_2)$

$\forall\, x_1, x_2$ in domain $D \subseteq R$

$\forall\, \theta \in [0,1]$

for $\theta x_1 + (1-\theta) x_2 \in D$,

D should be convex



II. f is (strictly) convex iff $f'(x)$ (strictly?) increasing in D

$$\left(f'(x_1) - f'(x_2)\right)(x_1 - x_2) \ge 0 \quad (>)$$

III. $f(y) \ge f(x) + f'(x)(y-x)$ ($>$)

$\underbrace{f(x) + f'(x)(y-x)}_{\text{Linear approximation to } y \text{ using } x}$

IV. $\because f'(x)$ is increasing $\Rightarrow f''(x) \ge 0$ ($>$)

Need to prove that I $\equiv$ II $\equiv$ III $\equiv$ IV

1] Extending these equivalences for a general $f: D \to R$ where D is a convex set

• $\boxed{\text{I}}$ $\forall x_1, x_2 \in D$ & $\forall \theta \in [0,1]$

$$f(\theta x_1 + (1-\theta) x_2) \le \theta f(x_1) + (1-\theta) f(x_2)$$
$$(<)$$

if ∇f exists $\forall x \in D$

$\boxed{\text{II}}$ ∇f is monotone in D i.e $\forall x_1, x_2 \in D$

ie $\langle \nabla f(x_1) - \nabla f(x_2), x_1 - x_2 \rangle \ge 0$ $(>)$

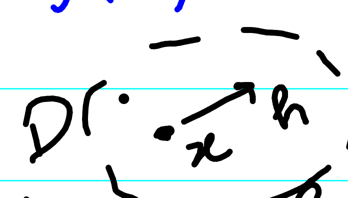
$\boxed{\text{III}}$ $\forall x_1, x_2 \in D$

$$f(x_2) \ge f(x_1) + \langle \nabla f(x_1), x_2 - x_1 \rangle$$
$$(>)$$

$\nabla^2 f$ exists at all $x \in D$

$\boxed{\text{IV}}$ $\forall x \in D$

$$\nabla^2 f(x) \succeq 0 \quad [\text{ie } \nabla^2 f(x) \text{ is psd}]$$
$$(\succ)$$



Another way of looking at extension of concept of convexity to $f: D \to R$ ~~is~~ to look at convexity of $\phi(t) = f(x + ht)$ for a pt x & any dirn h & $t \in [0,1]$

1. A differentiable function f is *strictly convex* (or *strictly concave up*) on an open interval $\mathcal{I}$, *iff*, $f'(x)$ is increasing on $\mathcal{I}$. Recall from theorem 46, the graphical interpretation of the first derivative $f'(x)$; $f'(x) > 0$ implies that $f(x)$ is increasing at x. Similarly, $f'(x)$ is increasing when $f''(x) > 0$. This gives us a sufficient condition for the strict convexity of a function:

Theorem 50 *If at all points in an open interval $\mathcal{I}$, $f(x)$ is doubly differentiable and if $f''(x) > 0$, $\forall x \in \mathcal{I}$, then the slope of the function is always increasing with x and the graph is strictly convex. This is illustrated in Figure 4.8.*

On the other hand, if the function is strictly convex and doubly differentiable in $\mathcal{I}$, then $f''(x) \geq 0$, $\forall x \in \mathcal{I}$.

There is also a slopeless interpretation of strict convexity as stated in the following theorem:

Theorem 51 *A differentiable function f is (strictly) convex on an open interval $\mathcal{I}$, iff*

$$f(ax_1 + (1-a)x_2) < af(x_1) + (1-a)f(x_2) \tag{4.2}$$

whenver $x_1, x_2 \in \mathcal{I}$, $x_1 \neq x_2$ and $0 < a < 1$.

IS EQUIVALENT TO SAYING THAT

A differentiable function f is (strictly) convex on $\mathcal{I}$ iff f' is strictly increasing on $\mathcal{I}$

Proof: First we will prove the necessity. Suppose f' is increasing on $\mathcal{I}$. Let $0 < a < 1$, $x_1, x_2 \in \mathcal{I}$ and $x_1 \neq x_2$. Without loss of generality assume that $x_1 < x_2$[3]. Then, $x_1 < ax_1 + (1-a)x_2 < x_2$ and therefore $ax_1 + (1-a)x_2 \in \mathcal{I}$. By the mean value theorem, there exist s and t with $x_1 < s < ax_1 + (1-a)x_2 < t < x_2$, such that $f(ax_1 + (1-a)x_2) - f(x_1) = f'(s)(x_2 - x_1)(1-a)$ and $f(x_2) - f(ax_1 + (1-a)x_2) = f'(t)(x_2 - x_1)a$. Therefore,

$$\begin{aligned} (1-a)f(x_1) - f(ax_1 + (1-a)x_2) + af(x_2) &= \\ a\left[f(x_2) - f(ax_1 + (1-a)x_2)\right] - (1-a)\left[f(ax_1 + (1-a)x_2) - f(x_1)\right] &= \\ a(1-a)(x_2 - x_1)\left[f'(t) - f'(s)\right] \end{aligned}$$

Since $f(x)$ is strictly convex on $\mathcal{I}$, $f'(x)$ is increasing $\mathcal{I}$ and therefore, $f'(t) - f'(s) > 0$. Moreover, $x_2 - x_1 > 0$ and $0 < a < 1$. This implies that $(1-a)f(x_1) - f(ax_1 + (1-a)x_2) + af(x_2) > 0$, or equivalently, $f(ax_1 + (1-a)x_2) < af(x_1) + (1-a)f(x_2)$, which is what we wanted to prove in 4.2.

Next, we prove the sufficiency. Suppose the inequality in 4.2 holds. Therefore,

$$\lim_{a \to 0} \frac{f(x_2 + a(x_1 - x_2)) - f(x_2)}{a} \leq f(x_1) - f(x_2)$$

that is,

$$f'(x_2)(x_1 - x_2) \leq f(x_1) - f(x_2) \tag{4.3}$$

Similarly, we can show that

$$f'(x_1)(x_2 - x_1) \leq f(x_2) - f(x_1) \tag{4.4}$$

Adding the left and right hand sides of inequalities in (4.3) and (4.4), and multiplying the resultant inequality by -1 gives us

$$\left(f'(x_2) - f'(x_1)\right)(x_2 - x_1) \geq 0 \tag{4.5}$$

Using the mean value theorem, $\exists z = x_1 + t(x_2 - x_1)$ for $t \in (0,1)$ such that

$$f(x_2) - f(x_1) = f'(z)(x_2 - x_1) \tag{4.6}$$

Since 4.5 holds for any $x_1, x_2 \in \mathcal{I}$, it also hold for $x_2 = z$. Therefore,

$$(f'(z) - f'(x_1))(x_2 - x_1) = \frac{1}{t}(f'(z) - f'(x_1))(z - x_1) \geq 0$$

Additionally using 4.6, we get

Proves III

$$f(x_2)-f(x_1) = (f'(z)-f'(x_1))(x_2-x_1)+f'(x_1)(x_2-x_1) \geq f'(x_1)(x_2-x_1) \tag{4.7}$$

Suppose equality holds in 4.5 for some $x_1 \neq x_2$. Then equality holds in 4.7 for the same x_1 and x_2. That is,

$$f(x_2) - f(x_1) = f'(x_1)(x_2 - x_1) \tag{4.8}$$

Applying 4.7 we can conclude that

$$f(x_1) + af'(x_1)(x_2 - x_1) \leq f(x_1 + a(x_2 - x_1)) \tag{4.9}$$

From 4.2 and 4.8, we can derive that

$$f(x_1 + a(x_2 - x_1)) < (1 - a)f(x_1) + af(x_2) = f(x_1) + af'(x_1)(x_2 - x_1) \tag{4.10}$$

However, equations 4.9 and 4.10 contradict each other. Therefore, equality in 4.5 cannot hold for any $x_1 \neq x_2$, implying that

$$(f'(x_2) - f'(x_1))(x_2 - x_1) > 0$$

that is, $f'(x)$ is increasing and therefore f is convex on $\mathcal{I}$. $\square$

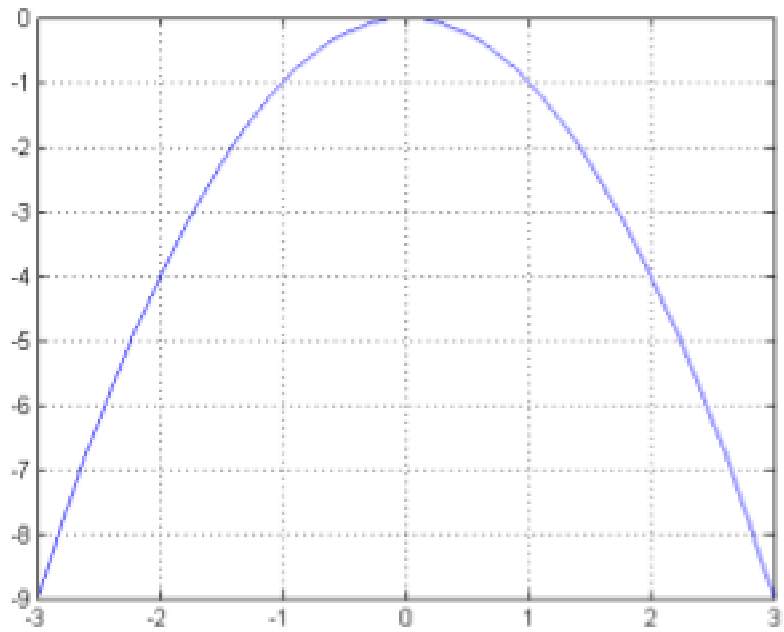


Figure 4.9: Plot for the strictly concave function $f(x) = -x^2$ which has $f''(x) = -2 < 0, \ \forall x$.

A differentiable function f is said to be *strictly concave* on an open interval $\mathcal{I}$, *iff*, $f'(x)$ is decreasing on $\mathcal{I}$. Recall from theorem 46, the graphical interpretation of the first derivative $f'(x)$; $f'(x) < 0$ implies that $f(x)$ is decreasing at x. Similarly, $f'(x)$ is monotonically decreasing when $f''(x) > 0$. This gives us a sufficient condition for the concavity of a function:

Theorem 52 *If at all points in an open interval $\mathcal{I}$, $f(x)$ is doubly differentiable and if $f''(x) < 0$, $\forall x \in \mathcal{I}$, then the slope of the function is always decreasing with x and the graph is strictly concave. This is illustrated in Figure 4.9.*

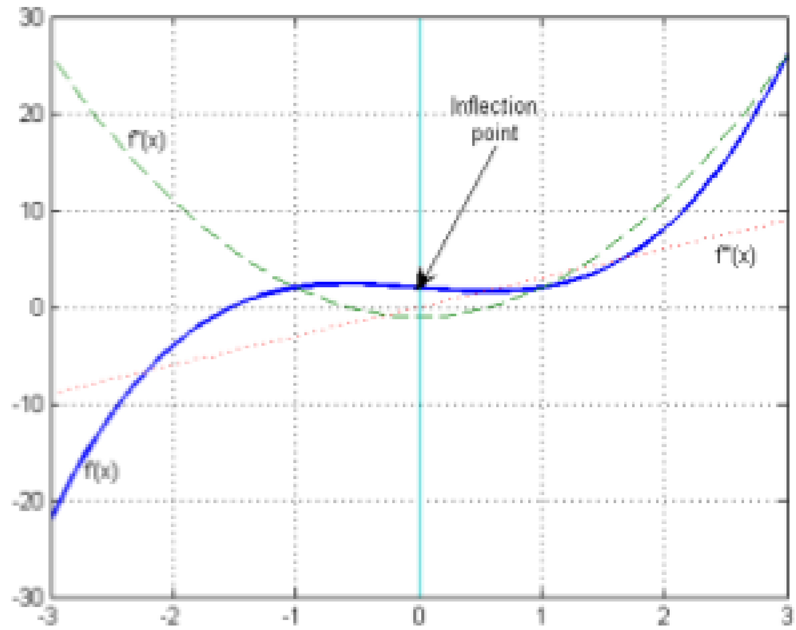
On the other hand, if the function is strictly concave and doubly differentiable in $\mathcal{I}$, then $f''(x) \leq 0$, $\forall x \in \mathcal{I}$.

There is also a slopeless interpretation of concavity as stated in the following theorem:

Theorem 53 *A differentiable function f is strictly concave on an open interval $\mathcal{I}$, iff*

$$f(ax_1 + (1-a)x_2) > af(x_1) + (1-a)f(x_2) \tag{4.11}$$

whenver $x_1, x_2 \in \mathcal{I}$, $x_1 \neq x_2$ and $0 < a < 1$.



concave region convex region

Figure 4.10: Plot for $f(x) = x^3 + x + 2$, which has an inflection point $x = 0$, along with plots for $f'(x)$ and $f''(x)$.

Procedure 2 [First derivative test in terms of strict convexity]: *Let c be a critical number of f and $f'(c) = 0$. Then,*

1. *$f(c)$ is a local minimum if the graph of $f(x)$ is strictly convex on an open interval containing c.*
2. *$f(c)$ is a local maximum if the graph of $f(x)$ is strictly concave on an open interval containing c.*

3. Convex functions

- basic properties and examples
- operations that preserve convexity
- the conjugate function
- quasiconvex functions
- log-concave and log-convex functions
- convexity with respect to generalized inequalities

Replace all $\mathbb{R}^n$ with D in definitions & claims made

Footnote: Several proofs will be in the setting of D being an inner prod space or D = $\mathbb{R}^n$

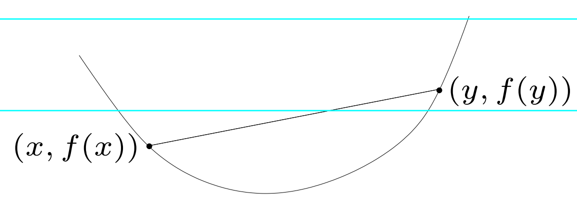
for each proof, note assumptions we make on topology

Definition

$f : \mathbf{R}^n \to \mathbf{R}$ is convex if $\mathbf{dom}\, f$ is a convex set and

$$f(\theta x + (1-\theta)y) \le \theta f(x) + (1-\theta) f(y)$$

for all $x, y \in \mathbf{dom}\, f$, $0 \le \theta \le 1$



- f is concave if $-f$ is convex
- f is strictly convex if $\mathbf{dom}\, f$ is convex and

$$f(\theta x + (1-\theta)y) < \theta f(x) + (1-\theta) f(y)$$

for $x, y \in \mathbf{dom}\, f$, $x \ne y$, $0 < \theta < 1$

Definition 35 [Convex Function]: *A function $f : \mathcal{D} \rightarrow \Re$ is convex if $\mathcal{D}$ is a convex set and*

implies

$$f(\theta\mathbf{x} + (1-\theta)\mathbf{y}) \leq \theta f(\mathbf{x}) + (1-\theta)f(\mathbf{y}) \quad \forall\, \mathbf{x},\mathbf{y} \in \mathcal{D} \quad 0 \leq \theta \leq 1 \tag{4.31}$$

Figure 4.37 illustrates an example convex function. A function $f : \mathcal{D} \rightarrow \Re$ is strictly convex if $\mathcal{D}$ is convex and

implies

$$f(\theta\mathbf{x} + (1-\theta)\mathbf{y}) < \theta f(\mathbf{x}) + (1-\theta)f(\mathbf{y})) \quad \forall\, \mathbf{x},\mathbf{y} \in \mathcal{D} \quad 0<\theta<1 \tag{4.32}$$

A function $f : \mathcal{D} \rightarrow \Re$ is called uniformly or strongly convex if $\mathcal{D}$ is convex and there exists a constant $c > 0$ such that

$$f(\theta\mathbf{x} + (1-\theta)\mathbf{y}) \leq \theta f(\mathbf{x}) + (1-\theta)f(\mathbf{y})) - \overbrace{\underbrace{\tfrac{1}{2}c\theta(1-\theta)||\mathbf{x}-\mathbf{y}||}_{\text{if } \theta\in(0,1) \ \& \ x\neq y}}^{>0}{}^{2} \quad \forall\, \mathbf{x},\mathbf{y} \in \mathcal{D} \quad 0 \leq \theta \leq 1($$

An inner prod norm

Prove

> **Theorem 69** *Let $f : \mathcal{D} \rightarrow \Re$ be a convex function on a convex domain $\mathcal{D}$. Any point of locally minimum solution for f is also a point of its globally minimum solution.*

Prove

> **Theorem 70** *Let $f : \mathcal{D} \rightarrow \Re$ be a strictly convex function on a convex domain $\mathcal{D}$. Then f has a unique point corresponding to its global minimum.*

> **Theorem 71** *A function $f : \mathcal{D} \rightarrow \Re$ is (strictly) convex if and only if the function $\phi : \mathcal{D}_\phi \rightarrow \Re$ defined below, is (strictly) convex in t for every $\mathbf{x} \in \mathcal{D}$ and for every $\mathbf{h}$* s.t $\forall\, t\in[0,1]$ $x+ht\in D$
>
> $$\phi(t) = f(\mathbf{x} + t\mathbf{h})$$
>
> *with the domain of ϕ given by $\mathcal{D}_\phi = \{t | \mathbf{x} + t\mathbf{h} \in \mathcal{D}\}$.*

Theorem 69 *Let $f : \mathcal{D} \rightarrow \Re$ be a convex function on a convex domain $\mathcal{D}$. Any point of locally minimum solution for f is also a point of its globally minimum solution.*

Proof: Suppose $\mathbf{x} \in \mathcal{D}$ is a point of local minimum and let $\mathbf{y} \in \mathcal{D}$ be a point of global minimum. Thus, $f(\mathbf{y}) < f(\mathbf{x})$. Since $\mathbf{x}$ corresponds to a local minimum, there exists an $\epsilon > 0$ such that

Assuming normed space

$$\forall\, \mathbf{z} \in \mathcal{D},\ ||\mathbf{z} - \mathbf{x}|| \leq \epsilon \Rightarrow f(\mathbf{z}) \geq f(\mathbf{x})$$

Consider a point $\mathbf{z} = \theta\mathbf{y} + (1-\theta)\mathbf{x}$ with $\theta = \frac{\epsilon}{2||\mathbf{y}-\mathbf{x}||}$. Since $\mathbf{x}$ is a point of local minimum (in a ball of radius ϵ), and since $f(\mathbf{y}) < f(\mathbf{x})$, it must be that $||\mathbf{y} - \mathbf{x}|| > \epsilon$. Thus, $0 < \theta < \frac{1}{2}$ and $\mathbf{z} \in \mathcal{D}$. Furthermore, $||\mathbf{z} - \mathbf{x}|| = \frac{\epsilon}{2}$. Since f is a convex function

$$f(\mathbf{z}) \leq \theta f(\mathbf{x}) + (1-\theta) f(\mathbf{y})$$

Since $f(\mathbf{y}) < f(\mathbf{x})$, we also have

$$\theta f(\mathbf{x}) + (1-\theta) f(\mathbf{y}) < f(\mathbf{x})$$

The two equations imply that $f(\mathbf{z}) < f(\mathbf{x})$, which contradicts our assumption that $\mathbf{x}$ corresponds to a point of local minimum. That is f cannot have a point of local minimum, which does not coincide with the point $\mathbf{y}$ of global minimum. □

Theorem 70 *Let $f : \mathcal{D} \rightarrow \Re$ be a strictly convex function on a convex domain $\mathcal{D}$. Then f has a unique point corresponding to its global minimum.*

Proof: Suppose $\mathbf{x} \in \mathcal{D}$ and $\mathbf{y} \in \mathcal{D}$ with $\mathbf{y} \neq \mathbf{x}$ are two points of global minimum. That is $f(\mathbf{x}) = f(\mathbf{y})$ for $\mathbf{y} \neq \mathbf{x}$. The point $\frac{\mathbf{x}+\mathbf{y}}{2}$ also belongs to the convex set $\mathcal{D}$ and since f is strictly convex, we must have

$$f\left(\frac{\mathbf{x}+\mathbf{y}}{2}\right) < \frac{1}{2} f(\mathbf{x}) + \frac{1}{2} f(\mathbf{y}) = f(\mathbf{x})$$

which is a contradiction. Thus, the point corresponding to the minimum of f must be unique. □

Theorem 71 *A function $f : \mathcal{D} \to \Re$ is (strictly) convex if and only if the function $\phi : \mathcal{D}_\phi \to \Re$ defined below, is (strictly) convex in t for every $\mathbf{x} \in \Re^n$ and for every $\mathbf{h} \in \Re^n$*

$$\phi(t) = f(\mathbf{x} + t\mathbf{h})$$

with the domain of ϕ given by $\mathcal{D}_\phi = \{t | \mathbf{x} + t\mathbf{h} \in \mathcal{D}\}$.

Proof: We will prove the necessity and sufficiency of the convexity of ϕ for a convex function f. The proof for necessity and sufficiency of the strict convexity of ϕ for a strictly convex f is very similar and is left as an exercise.

Proof of Necessity: Assume that f is convex. And we need to prove that $\phi(t) = f(\mathbf{x} + t\mathbf{h})$ is also convex. Let $t_1, t_2 \in \mathcal{D}_\phi$ and $\theta \in [0, 1]$. Then,

$$\begin{aligned}
&\phi(\theta t_1 + (1-\theta)t_2) = f\left(\theta(\mathbf{x} + t_1\mathbf{h}) + (1-\theta)(\mathbf{x} + t_2\mathbf{h})\right) \\
&\leq \theta f\left((\mathbf{x} + t_1\mathbf{h})\right) + (1-\theta) f\left((\mathbf{x} + t_2\mathbf{h})\right) = \theta\phi(t_1) + (1-\theta)\phi(t_2)
\end{aligned} \quad (4.35)$$

Thus, ϕ is convex.

Proof of Sufficiency: Assume that for every $\mathbf{h} \in \Re^n$ and every $\mathbf{x} \in \Re^n$, $\phi(t) = f(\mathbf{x} + t\mathbf{h})$ is convex. We will prove that f is convex. Let $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{D}$. Take, $\mathbf{x} = \mathbf{x}_1$ and $\mathbf{h} = \mathbf{x}_2 - \mathbf{x}_1$. We know that $\phi(t) = f\left(\mathbf{x}_1 + t(\mathbf{x}_2 - \mathbf{x}_1)\right)$ is convex, with $\phi(1) = f(\mathbf{x}_2)$ and $\phi(0) = f(\mathbf{x}_1)$. Therefore, for any $\theta \in [0, 1]$

$$\begin{aligned}
&f\left(\theta\mathbf{x}_2 + (1-\theta)\mathbf{x}_1\right) = \phi(\theta) \\
&\leq \theta\phi(1) + (1-\theta)\phi(0) \leq \theta f(\mathbf{x}_2) + (1-\theta) f(\mathbf{x}_1)
\end{aligned} \quad (4.36)$$

This implies that f is convex. $\square$

First-order condition

f is **differentiable** if $\mathbf{dom}\, f$ is open and the gradient

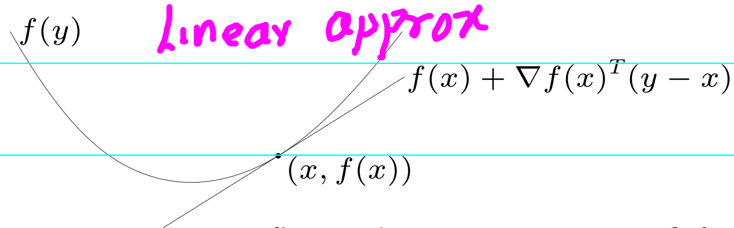
$$\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1}, \frac{\partial f(x)}{\partial x_2}, \ldots, \frac{\partial f(x)}{\partial x_n} \right)$$

exists at each $x \in \mathbf{dom}\, f$

1st-order condition: differentiable f with convex domain is convex iff

(>) for strict

$$f(y) \geq \underbrace{f(x) + \nabla f(x)^T (y - x)}_{\text{Linear approx}} \quad \text{for all } x, y \in \mathbf{dom}\, f$$



first-order approximation of f is global underestimator

Theorem 75 *Let $f : \mathcal{D} \to \Re$ be a differentiable convex function on an open convex set $\mathcal{D}$. Then:*

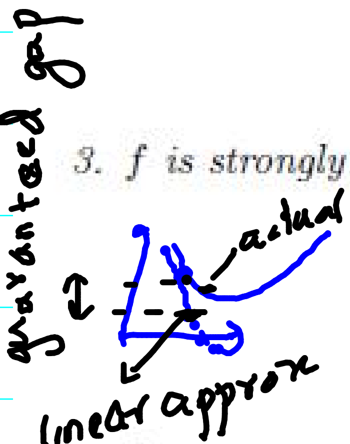
1. *f is convex if and only if, for any $\mathbf{x}, \mathbf{y} \in \mathcal{D}$,*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla^T f(\mathbf{x})(\mathbf{y} - \mathbf{x}) \tag{4.44}$$

2. *f is strictly convex on $\mathcal{D}$ if and only if, for any $\mathbf{x}, \mathbf{y} \in \mathcal{D}$, with $\mathbf{x} \neq \mathbf{y}$,*

$$f(\mathbf{y}) > f(\mathbf{x}) + \nabla^T f(\mathbf{x})(\mathbf{y} - \mathbf{x}) \tag{4.45}$$

3. *f is strongly convex on $\mathcal{D}$ if and only if, for any $\mathbf{x}, \mathbf{y} \in \mathcal{D}$,*



[Turn to page 32]

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla^T f(\mathbf{x})(\mathbf{y} - \mathbf{x}) + \frac{1}{2}c\|\mathbf{y} - \mathbf{x}\|^2 \tag{4.46}$$

for some constant $c > 0$.

Q: For a fixed x, what is minimum value RHS can take?

Proof:

Sufficiency: The proof of sufficiency is very similar for all the three statements of the theorem. So we will prove only for statement (4.44). Suppose (4.44) holds. Consider $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{D}$ and any $\theta \in (0, 1)$. Let $\mathbf{x} = \theta\mathbf{x}_1 + (1 - \theta)\mathbf{x}_2$. Then,

$$\begin{aligned} f(\mathbf{x}_1) &\geq f(\mathbf{x}) + \nabla^T f(\mathbf{x})(\mathbf{x}_1 - \mathbf{x}) \\ f(\mathbf{x}_2) &\geq f(\mathbf{x}) + \nabla^T f(\mathbf{x})(\mathbf{x}_2 - \mathbf{x}) \end{aligned} \tag{4.47}$$

Adding $(1 - \theta)$ times the second inequality to θ times the first, we get,

$$\theta f(\mathbf{x}_1) + (1-\theta) f(\mathbf{x}_2) \geq f(\mathbf{x})$$

which proves that $f(\mathbf{x})$ is a convex function. In the case of strict convexity, strict inequality holds in (4.47) and it follows through. In the case of strong convexity, we need to additionally prove that

$$\theta \frac{1}{2} c ||\mathbf{x} - \mathbf{x}_1||^2 + (1-\theta) \frac{1}{2} c ||\mathbf{x} - \mathbf{x}_2||^2 = \frac{1}{2} c\theta(1-\theta)||\mathbf{x}_2 - \mathbf{x}_1||^2$$

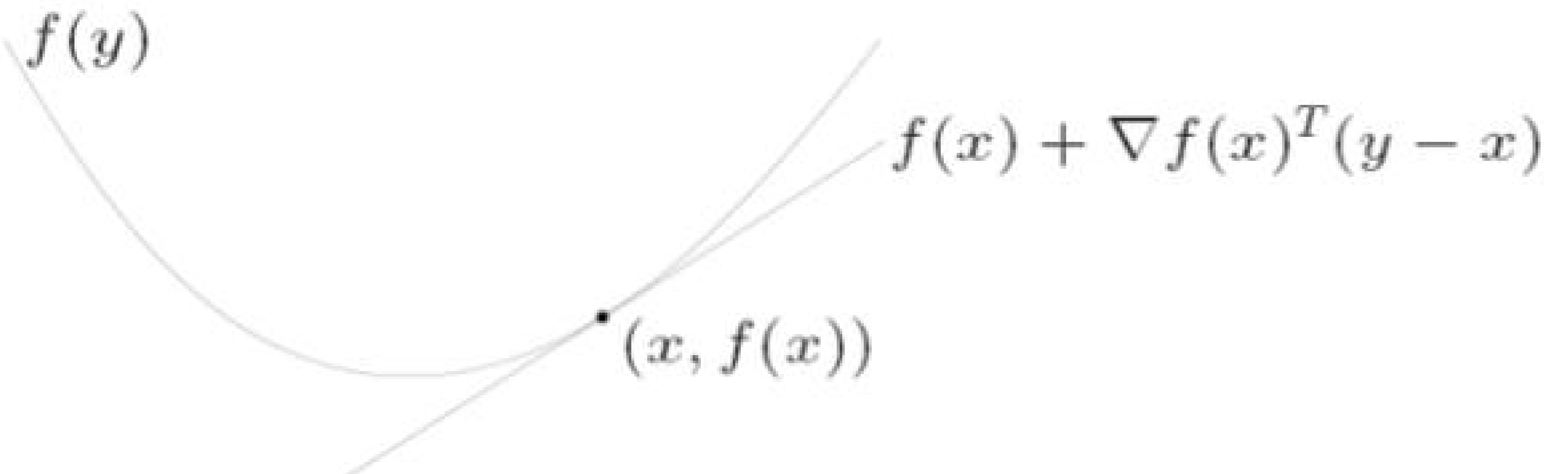


Figure 4.38: Figure illustrating Theorem 75.

Necessity: Suppose f is convex. Then for all $\theta \in (0,1)$ and $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{D}$, we must have

$$f(\theta \mathbf{x}_2 + (1-\theta)\mathbf{x}_1) \leq \theta f(\mathbf{x}_2) + (1-\theta) f(\mathbf{x}_1)$$

Thus,

Directional derivative along $x_2 - x_1$

$$\nabla^T f(\mathbf{x}_1)(\mathbf{x}_2 - \mathbf{x}_1) = \lim_{\theta \to 0} \frac{f\left(\mathbf{x}_1 + \theta(\mathbf{x}_2 - \mathbf{x}_1)\right) - f\left(\mathbf{x}_1\right)}{\theta} \leq f(\mathbf{x}_2) - f(\mathbf{x}_1)$$

This proves necessity for (4.44). The necessity proofs for (4.45) and (4.46) are very similar, except for a small difference for the case of strict convexity; the strict inequality is not preserved when we take limits. Suppose equality does hold in the case of strict convexity, that is for a strictly convex function f, let

$$f(\mathbf{x}_2) = f(\mathbf{x}_1) + \nabla^T f(\mathbf{x}_1)(\mathbf{x}_2 - \mathbf{x}_1) \tag{4.48}$$

succint representation for dir deriv

for some $\mathbf{x}_2 \neq \mathbf{x}_1$. Because f is stricly convex, for any $\theta \in (0,1)$ we can write

$$f\left(\theta \mathbf{x}_1 + (1-\theta)\mathbf{x}_2\right) = f\left(\mathbf{x}_2 + \theta(\mathbf{x}_1 - \mathbf{x}_2)\right) < \theta f(\mathbf{x}_1) + (1-\theta) f(\mathbf{x}_2) \quad (4.49)$$

Since (4.44) is already proved for convex functions, we use it in conjunction with (4.48), and (4.49), to get

$$f(\mathbf{x}_2) + \theta \nabla^T f(\mathbf{x}_2)(\mathbf{x}_1 - \mathbf{x}_2) \leq f\left(\mathbf{x}_2 + \theta(\mathbf{x}_1 - \mathbf{x}_2)\right) < f(\mathbf{x}_2) + \theta \nabla^T f(\mathbf{x}_2)(\mathbf{x}_1 - \mathbf{x}_2)$$

which is a contradiction. Thus, equality can never hold in (4.44) for any $\mathbf{x}_1 \neq \mathbf{x}_2$. This proves the necessity of (4.45). □

Definition 40 [Some corollaries of theorem 75 for strongly convex fun

For a fixed $\mathbf{x}$, the right hand side of the inequality (4.46) is a convex quadratic function of $\mathbf{y}$. Thus, the critical point of the RHS should correspond to the minimum value that the RHS could take. This yields another lower bound on $f(\mathbf{y})$.

$$f(\mathbf{y}) \geq f(\mathbf{x}) - \frac{1}{2c}||\nabla f(\mathbf{x})||_2^2 \quad (4.50)$$

Since this holds for any $\mathbf{y} \in \mathcal{D}$, we have

$$\min_{\mathbf{y} \in \mathcal{D}} f(\mathbf{y}) \geq f(\mathbf{x}) - \frac{1}{2c}||\nabla f(\mathbf{x})||_2^2 \quad (4.51)$$

which can be used to bound the suboptimality of a point $\mathbf{x}$ in terms of $||\nabla f(\mathbf{x})||_2$. This bound comes handy in theoretically understanding the convergence of gradient methods. If $\widehat{\mathbf{y}} = \min_{\mathbf{y} \in \mathcal{D}} f(\mathbf{y})$, we can also derive a bound on the distance between any point $\mathbf{x} \in \mathcal{D}$ and the point of optimality $\widehat{\mathbf{y}}$.

should be perhaps $\frac{1}{c}$?

$$||\mathbf{x} - \widehat{\mathbf{y}}||_2 \leq \frac{2}{c}||\nabla f(\mathbf{x})||_2 \quad (4.52)$$

Theorem 78 *Let $f : \mathcal{D} \rightarrow \Re$ with $\mathcal{D} \subseteq \Re^n$ be differentiable on the convex set $\mathcal{D}$. Then,*

1. *f is convex on $\mathcal{D}$ if and only if is its gradient ∇f is monotone. That is, for all $\mathbf{x}, \mathbf{y} \in \Re$*

$$\left(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\right)^T (\mathbf{x} - \mathbf{y}) \geq 0 \tag{4.53}$$

2. *f is strictly convex on $\mathcal{D}$ if and only if is its gradient ∇f is strictly monotone. That is, for all $\mathbf{x}, \mathbf{y} \in \Re$ with $\mathbf{x} \neq \mathbf{y}$,*

$$\left(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\right)^T (\mathbf{x} - \mathbf{y}) > 0 \tag{4.54}$$

3. *f is uniformly or strongly convex on $\mathcal{D}$ if and only if is its gradient ∇f is uniformly monotone. That is, for all $\mathbf{x}, \mathbf{y} \in \Re$,*

$$\left(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\right)^T (\mathbf{x} - \mathbf{y}) \geq c||\mathbf{x} - \mathbf{y}||^2 \tag{4.55}$$

   for some constant $c > 0$.

Proof:

Necessity: Suppose f is uniformly convex on $\mathcal{D}$. Then from theorem 75, we know that for any $\mathbf{x}, \mathbf{y} \in \mathcal{D}$,

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla^T f(\mathbf{x})(\mathbf{y}-\mathbf{x}) - \frac{1}{2}c||\mathbf{y}+\mathbf{x}||^2$$
$$f(\mathbf{x}) \geq f(\mathbf{y}) + \nabla^T f(\mathbf{y})(\mathbf{x}-\mathbf{y}) - \frac{1}{2}c||\mathbf{x}+\mathbf{y}||^2$$

Adding the two inequalities, we get (4.55). If f is convex, the inequalities hold with $c = 0$, yielding (4.54). If f is strictly convex, the inequalities will be strict, yielding (4.54).

Sufficiency: Suppose ∇f is monotone. For any fixed $\mathbf{x}, \mathbf{y} \in \mathcal{D}$, consider the function $\phi(t) = f(\mathbf{x} + t(\mathbf{y}-\mathbf{x}))$. By the mean value theorem applied to $\phi(t)$, we should have for some $t \in (0, 1)$,

$$\phi(1) - \phi(0) = \phi'(t) \tag{4.56}$$

Letting $\mathbf{z} = \mathbf{x} + t(\mathbf{y}-\mathbf{x})$, (4.56) translates to

$$f(\mathbf{y}) - f(\mathbf{x}) = \nabla^T f(\mathbf{z})(\mathbf{y}-\mathbf{x}) \tag{4.57}$$

Also, by definition of monotonicity of ∇f, (from (4.53)),

$$\left(\nabla f(\mathbf{z}) - \nabla f(\mathbf{x})\right)^T (\mathbf{y}-\mathbf{x}) = \frac{1}{t}\left(\nabla f(\mathbf{z}) - \nabla f(\mathbf{x})\right)^T (\mathbf{z}-\mathbf{x}) \geq 0 \tag{4.58}$$

Combining (4.57) with (4.58), we get,

$$\begin{aligned} f(\mathbf{y}) - f(\mathbf{x}) &= (\nabla f(\mathbf{z}) - f(\mathbf{x}))^T (\mathbf{y} - \mathbf{x}) + \nabla^T f(\mathbf{x})(\mathbf{y} - \mathbf{x}) \\ &\geq \nabla^T f(\mathbf{x})(\mathbf{y} - \mathbf{x}) \end{aligned} \tag{4.59}$$

By theorem 75, this inequality proves that f is convex. Strict convexity can be similarly proved by using the strict inequality in (4.58) inherited from strict monotonicity, and letting the strict inequality follow through to (4.59). For the case of strong convexity, from (4.55), we have

$$\begin{aligned} \phi'(t) - \phi'(0) &= (\nabla f(\mathbf{z}) - f(\mathbf{x}))^T (\mathbf{y} - \mathbf{x}) \\ = \frac{1}{t} (\nabla f(\mathbf{z}) - f(\mathbf{x}))^T (\mathbf{z} - \mathbf{x}) &\geq \frac{1}{t} c||\mathbf{z} - \mathbf{x}||^2 = ct||\mathbf{y} - \mathbf{x}||^2 \end{aligned} \tag{4.60}$$

Therefore,

$$\phi(1) - \phi(0) - \phi'(0) = \int_0^1 [\phi'(t) - \phi'(0)] dt \geq \frac{1}{2} c||\mathbf{y} - \mathbf{x}||^2 \tag{4.61}$$

which translates to

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla^T f(\mathbf{x})(\mathbf{y} - \mathbf{x}) + \frac{1}{2} c||\mathbf{y} - \mathbf{x}||^2$$

By theorem 75, f must be strongly convex. $\square$

Theorem 79 *A twice differential function $f : \mathcal{D} \to \Re$ for a nonempty open convex set $\mathcal{D}$*

1. *is convex if and only if its domain is convex and its Hessian matrix is positive semidefinite at each point in $\mathcal{D}$. That is*

$$\nabla^2 f(\mathbf{x}) \succeq 0 \quad \forall\, \mathbf{x} \in \mathcal{D} \tag{4.62}$$

2. *is strictly convex if its domain is convex and its Hessian matrix is positive definite at each point in $\mathcal{D}$. That is*

$$\nabla^2 f(\mathbf{x}) \succ 0 \quad \forall\, \mathbf{x} \in \mathcal{D} \tag{4.63}$$

3. *is uniformly convex if and only if its domain is convex and its Hessian matrix is uniformly positive definite at each point in $\mathcal{D}$. That is, for any $\mathbf{v} \in \Re^n$ and any $\mathbf{x} \in \mathcal{D}$, there exists a $c > 0$ such that*

$$\mathbf{v}^T \nabla^2 f(\mathbf{x}) \mathbf{v} \geq c||\mathbf{v}||^2 \tag{4.64}$$

From discussions on generalised inequalities this means $\nabla^2 f(x) - cI \in S_+^n$

In other words

$$\nabla^2 f(\mathbf{x}) \succeq cI_{n\times n}$$

where $I_{n\times n}$ is the $n \times n$ identity matrix and $\succeq$ corresponds to the positive semidefinite inequality. That is, the function f is strongly convex iff $\nabla^2 f(\mathbf{x}) - cI_{n\times n}$ is positive semidefinite, for all $\mathbf{x} \in \mathcal{D}$ and for some constant $c > 0$, which corresponds to the positive minimum curvature of f.

H/w problem:

04/10/2013. Make sure that you understand the proofs for local minimizer=global minimizer, unique global minimizer for a strictly convex function, equivalence of different mathematical specifications (gradient free, first order, gradient monotonicity and Hessian) of convexity spanning pages 25 to 36. Now solve following problems (i) Show that the sum of a convex and a strictly/strongly convex function is strictly/strongly convex. (ii) Suppose that f(x)= x'Qx, where Q is an n x n matrix. Show conditions under which f(x) is (strictly/strongly) convex and show this using each of the 4 equivalent conditions for (strict/strong) convexity. **Deadline:** October 9 2013.

Rules for gradient (assuming gradient is column vector)

$$f(x) = \cancel{x^T} Q y \qquad f(x) = x^T Q x$$

$$\nabla f(x) = Q y \qquad \nabla f(x) = \underbrace{Q x + Q^T x}$$

$$f(x) = y^T Q x = \cancel{x^T} Q^T y \qquad \frac{d\, g(x) h(x)}{dx} = g'(x) h(x) + h'(x) g(x)$$

$$\nabla f(x) = Q^T y$$

Rough high level plan for the course from hereon
by Ganesh Ramakrishnan - Wednesday, 9 October 2013, 9:16 AM

Please give feedback (of course, I am not listing topics within each high level topic)

1] Further properties of convex functions, subgradients.

2] Algorithms for unconstrained optimisation, illustrations/comparisons, convergence analysis for some

3] Dealing with constraints: Lagrange multipliers, duality, conjugate functions, polars, etc

4] Algorithms for constrained optimisation

Ganesh

Examples on R

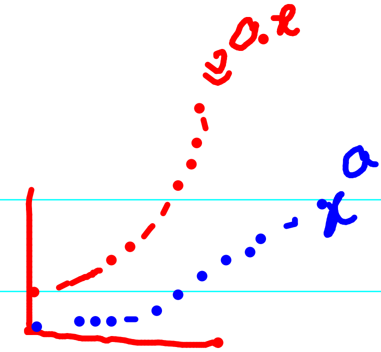
convex:

- affine: $ax + b$ on **R**, for any $a, b \in$ **R**
- exponential: e^{ax}, for any $a \in$ **R**
- powers: x^α on $\mathbf{R}_{++}$, for $\alpha \geq 1$ or $\alpha \leq 0$
- powers of absolute value: $|x|^p$ on **R**, for $p \geq 1$
- negative entropy: $x \log x$ on $\mathbf{R}_{++}$ → strictly convex

$\frac{\partial^2}{\partial x^2} e^{ax} = a^2 e^{ax}$

$\frac{\partial^2}{\partial x^2} x^{\alpha} = \alpha(\alpha-1)x^{\alpha-2}$

$e^x = 1 + x + \frac{x^2}{2!}$



concave:

- affine: $ax + b$ on **R**, for any $a, b \in$ **R**
- powers: x^α on $\mathbf{R}_{++}$, for $0 \leq \alpha \leq 1$
- logarithm: $\log x$ on $\mathbf{R}_{++}$

f is concave if
-f is convex

Examples on $\mathbf{R}^n$ and $\mathbf{R}^{m \times n}$

affine functions are convex and concave; all norms are convex

examples on $\mathbf{R}^n$

- affine function $f(x) = a^T x + b$
- norms: $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$ for $p \geq 1$; $\|x\|_\infty = \max_k |x_k|$

examples on $\mathbf{R}^{m \times n}$ ($m \times n$ matrices)

- affine function

$$f(X) = \mathbf{tr}(A^T X) + b = \sum_{i=1}^m \sum_{j=1}^n A_{ij} X_{ij} + b$$

often used representation.

- spectral (maximum singular value) norm

$$f(X) = \|X\|_2 = \sigma_{\max}(X) = (\lambda_{\max}(X^T X))^{1/2}$$

$\|X\|_{spec} = \sup_v \frac{\|Xv\|_2}{\|v\|_2}$

Recall: induced norm for $\|v\|_2$

Proof that spectral norm is convex

(In general any induced matrix norm: H/w exercise)

$$\|\theta X_1 + (1-\theta) X_2\|_{spec} = \sup_v \frac{\|\theta X_1 v + (1-\theta) X_2 v\|_2}{\|v\|_2}$$

$$\leq \sup_v \left[\frac{\theta \|X_1 v\|_2}{\|v\|_2} + \frac{(1-\theta)\|X_2 v\|_2}{\|v\|_2} \right]$$

[using Cauchy Schwarz?]

$$\leq \sup_v \theta \frac{\|X_1 v\|_2}{\|v\|_2} + \sup_v (1-\theta) \frac{\|X_2 v\|_2}{\|v\|_2}$$

[supremem or max of sums is $\leq$ sum of supremums/max]

$$= \theta \|X_1\|_{spec} + (1-\theta) \|X_2\|_{spec}$$

For proof that $\|X\|_{spec} = \sup_{v} \frac{\|Xv\|_2}{\|v\|_2}$

a) We can prove that $X^T X$ is always positive definite $\left\{ v^T X^T X v = (Xv)^T (Xv) = \|Xv\|^2 \geq 0 \right\}$
& all its eigenvalues are therefore ≥ 0

Let $\lambda_1^2 \geq \lambda_2^2 \cdots \geq \lambda_n^2 \geq 0$ (they are called the singular values of X)

Let $u_1, u_2 \ldots u_n$ be n corresponding independent orthonormal eigenvectors of $X^T X$ (H/ω)
∴ They can be used as basis vectors and any $v \in \mathbb{R}^n$ can be expressed as:

$$\boxed{v = \sum_{i=1}^{n} c_i u_i}$$

$$\therefore \|Xv\|_2^2 = (Xv)^T (Xv) = v^T \left[X^T X v\right] = \left(\sum_{i=1}^{n} c_i u_i\right)^T \left(\sum_{i=1}^{n} c_i \lambda_i^2 u_i\right)$$

$$= \sum_{i=1}^{n} c_i^2 \lambda_i^2$$

$$\|v\|_2^2 = \left(\sum_{i=1}^{n} c_i u_i\right)^T \left(\sum_{i=1}^{n} c_i u_i\right) = \sum_{i=1}^{n} c_i^2$$

$$\|X\|_{spec} = \sup_{c_1 \ldots c_n} \sqrt{\frac{\sum c_i^2 \lambda_i^2}{\sum c_i^2}} = \lambda_1 = \lambda_{max}$$

soln: $c_1 \neq 0$, $c_2 = c_3 = c_4 \ldots = 0$

{ Can be solved analytically

Restriction of a convex function to a line

$f : \mathbf{R}^n \to \mathbf{R}$ is convex if and only if the function $g : \mathbf{R} \to \mathbf{R}$,

Proved earlier [Theorem 7.1]

Restriction of $f(x)$ in any direction v gives you convex fn $g(t)$

$$g(t) = f(x + tv), \qquad \mathbf{dom}\, g = \{t \mid x + tv \in \mathbf{dom}\, f\}$$

is convex (in t) for any $x \in \mathbf{dom}\, f$, $v \in \mathbf{R}^n$

can check convexity of f by checking convexity of functions of one variable

example. $f : \mathbf{S}^n \to \mathbf{R}$ with $f(X) = \log\det X$, $\mathbf{dom}\, X = \mathbf{S}^n_{++}$

$\det(AB) = \det(A) * \det(B)$

$$\begin{aligned} g(t) = \log\det(X + tV) &= \log\det X + \log\det(I + tX^{-1/2}VX^{-1/2}) \\ &= \log\det X + \sum_{i=1}^{n} \log(1 + t\lambda_i) \end{aligned}$$

$X + tV = X(I + tX^{-1/2}VX^{1/2})$ (H/w)

where λ_i are the eigenvalues of $X^{-1/2}VX^{-1/2}$

g is concave in t (for any choice of $X \succ 0$, V); hence f is concave

http://www.proofwiki.org/wiki/Determinant_of_Matrix_Product

http://en.wikipedia.org/wiki/Matrix_determinant_lemma

Extended-value extension

extended-value extension $\tilde{f}$ of f is

$$\tilde{f}(x) = f(x), \quad x \in \mathbf{dom}\, f, \qquad \tilde{f}(x) = \infty, \quad x \notin \mathbf{dom}\, f$$

often simplifies notation; for example, the condition

$$0 \le \theta \le 1 \quad \Longrightarrow \quad \tilde{f}(\theta x + (1-\theta)y) \le \theta \tilde{f}(x) + (1-\theta)\tilde{f}(y)$$

(as an inequality in $\mathbf{R} \cup \{\infty\}$), means the same as the two conditions

- $\mathbf{dom}\, f$ is convex
- for $x, y \in \mathbf{dom}\, f$,

$$0 \le \theta \le 1 \quad \Longrightarrow \quad f(\theta x + (1-\theta)y) \le \theta f(x) + (1-\theta)f(y)$$

Examples

quadratic function: $f(x) = (1/2)x^T Px + q^T x + r$ (with $P \in \mathbf{S}^n$)

$$\nabla f(x) = Px + q, \qquad \nabla^2 f(x) = P$$

convex if $P \succeq 0$

least-squares objective: $f(x) = \|Ax - b\|_2^2$

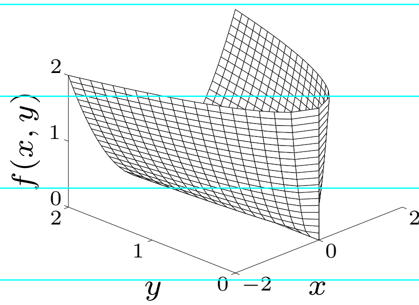
$$\nabla f(x) = 2A^T(Ax - b), \qquad \nabla^2 f(x) = 2A^T A$$

convex (for any A)

quadratic-over-linear: $f(x, y) = x^2/y$

$$\nabla^2 f(x,y) = \frac{2}{y^3}\begin{bmatrix} y \\ -x \end{bmatrix}\begin{bmatrix} y \\ -x \end{bmatrix}^T \succeq 0$$

convex for $y > 0$



log-sum-exp: $f(x) = \log \sum_{k=1}^n \exp x_k$ is convex

$$\nabla^2 f(x) = \frac{1}{\mathbf{1}^T z}\mathbf{diag}(z) - \frac{1}{(\mathbf{1}^T z)^2} z z^T \qquad (z_k = \exp x_k)$$

to show $\nabla^2 f(x) \succeq 0$, we must verify that $v^T \nabla^2 f(x) v \geq 0$ for all v:

$$v^T \nabla^2 f(x) v = \frac{(\sum_k z_k v_k^2)(\sum_k z_k) - (\sum_k v_k z_k)^2}{(\sum_k z_k)^2} \geq 0$$

since $(\sum_k v_k z_k)^2 \leq (\sum_k z_k v_k^2)(\sum_k z_k)$ (from Cauchy-Schwarz inequality)

geometric mean: $f(x) = (\prod_{k=1}^n x_k)^{1/n}$ on $\mathbf{R}_{++}^n$ is concave

(similar proof as for log-sum-exp)

Recall Level set/curve:

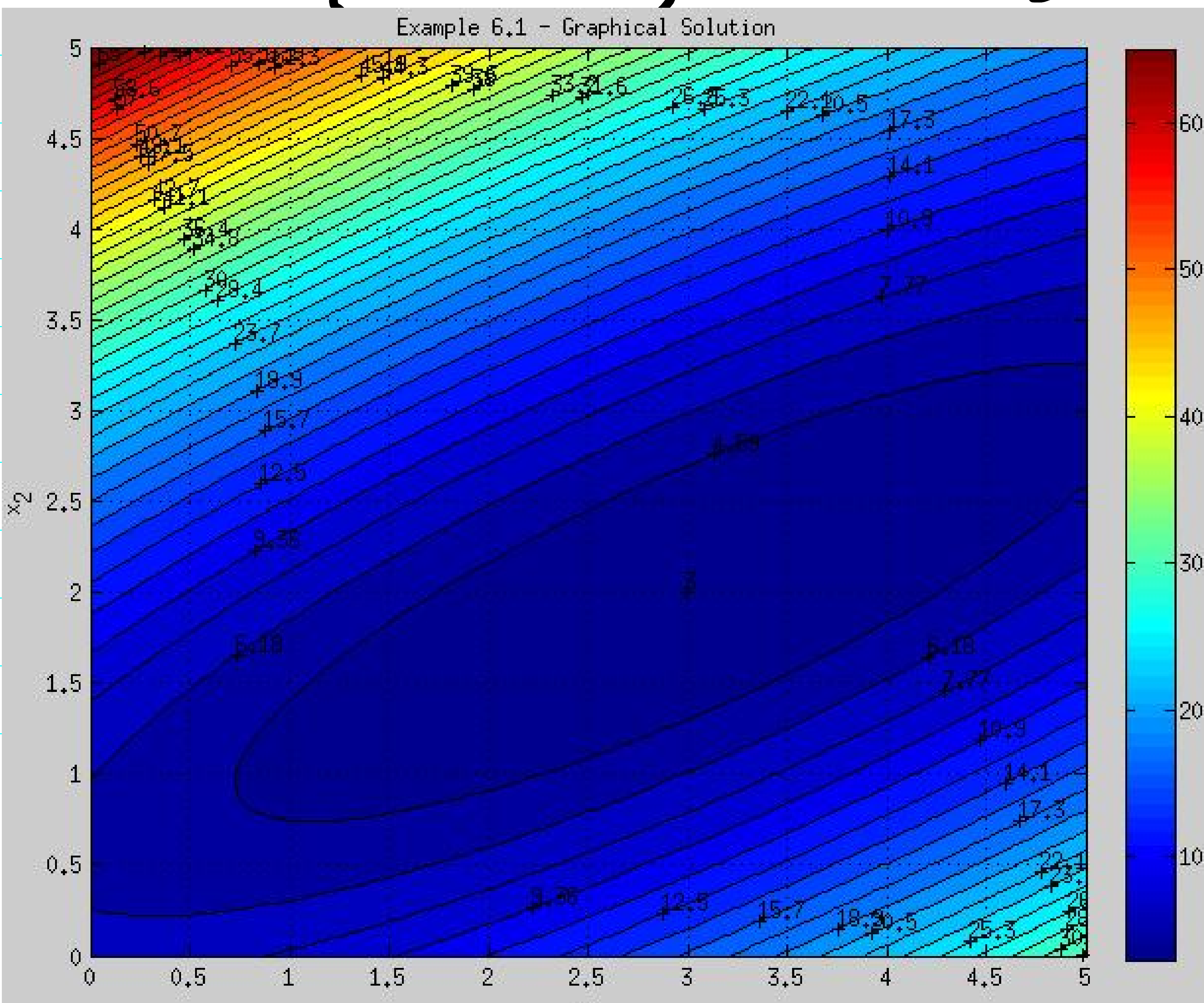
α-level set of $f: D \to \mathbb{R}$

$$L_\alpha = \{x \in D \mid f(x) = \alpha\}$$

(convex fn with convex sublevel sets & epigraph)

Consider: $f(x_1, x_2) = 3 + (x_1 - 1.5x_2)^2 + (x_2 - 2)^2$

$$D = \{0 \le x_1 \le 5,\ 0 \le x_2 \le 5\}$$



Epigraph and sublevel set

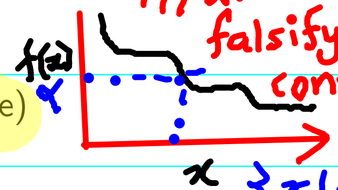
α-**sublevel set** of $f : \mathbf{R}^n \to \mathbf{R}$:

Domain has to be convex

$$C_\alpha = \{x \in \mathbf{dom}\, f \mid f(x) \le \alpha\}$$

sublevel sets of convex functions are convex (converse is false)

Illustration falsifying converse



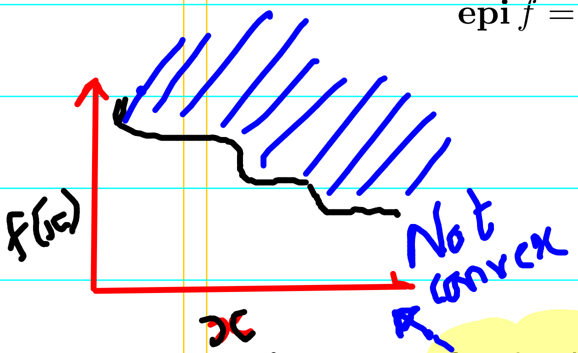
$\{x \mid f(x) \le \alpha\}$

epigraph of $f : \mathbf{R}^n \to \mathbf{R}$:

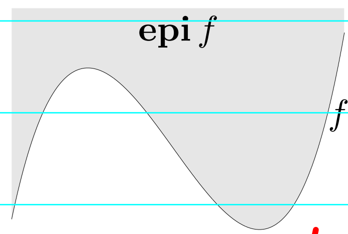
$$\mathbf{epi}\, f = \{(x,t) \in \mathbf{R}^{n+1} \mid x \in \mathbf{dom}\, f,\ f(x) \le t\}$$ + $f(y) \ge f(x) + \nabla^T f(x)(y-x)$

$\Downarrow$

$(\nabla f(x), -1)$ is normal to supporting hyperplane to epigraph of f at x



Not convex



f is convex if and only if $\mathbf{epi}\, f$ is a convex set

Jensen's inequality

basic inequality: if f is convex, then for $0 \le \theta \le 1$,

$$f(\theta x + (1-\theta)y) \le \theta f(x) + (1-\theta) f(y)$$

extension: if f is convex, then

$$f(\mathbf{E}\, z) \le \mathbf{E}\, f(z)$$

for any random variable z

basic inequality is special case with discrete distribution

$$\mathbf{prob}(z = x) = \theta, \qquad \mathbf{prob}(z = y) = 1 - \theta$$

Operations that preserve convexity

practical methods for establishing convexity of a function

1. verify definition (often simplified by restricting to a line)

2. for twice differentiable functions, show $\nabla^2 f(x) \succeq 0$

3. show that f is obtained from simple convex functions by operations that preserve convexity

   - nonnegative weighted sum
   - composition with affine function
   - pointwise maximum and supremum
   - composition
   - minimization
   - perspective

Positive weighted sum & composition with affine function

nonnegative multiple: αf is convex if f is convex, $\alpha \geq 0$

sum: $f_1 + f_2$ convex if f_1, f_2 convex (extends to infinite sums, integrals)

composition with affine function: $f(Ax + b)$ is convex if f is convex

examples

- log barrier for linear inequalities

$$f(x) = -\sum_{i=1}^{m} \log(b_i - a_i^T x), \qquad \mathbf{dom}\, f = \{x \mid a_i^T x < b_i, i = 1, \ldots, m\}$$

- (any) norm of affine function: $f(x) = \|Ax + b\|$

Pointwise maximum

if $f_1, \ldots, f_m$ are convex, then $f(x) = \max\{f_1(x), \ldots, f_m(x)\}$ is convex

Even if f_i's are differentiable, f will not be $\therefore$

need concept of subgradient

examples

- piecewise-linear function: $f(x) = \max_{i=1,\ldots,m}(a_i^T x + b_i)$ is convex
- sum of r largest components of $x \in \mathbf{R}^n$:

$$f(x) = x_{[1]} + x_{[2]} + \cdots + x_{[r]}$$

  is convex ($x_{[i]}$ is ith largest component of x)

  proof:

$$f(x) = \max\{x_{i_1} + x_{i_2} + \cdots + x_{i_r} \mid \underbrace{1 \le i_1 < i_2 < \cdots < i_r \le n}_{{}^nC_r \text{ such index choices}}\}$$

Pointwise supremum

if $f(x, y)$ is convex in x for each $y \in \mathcal{A}$, then

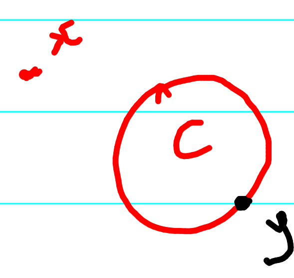
$$g(x) = \sup_{y \in \mathcal{A}} f(x, y)$$

is convex

examples

- support function of a set C: $S_C(x) = \sup_{y \in C} y^T x$ is convex
- distance to farthest point in a set C:

$$f(x) = \sup_{y \in C} \|x - y\|$$

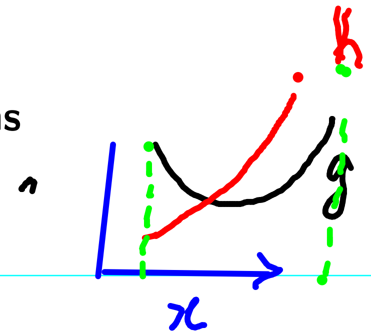


- maximum eigenvalue of symmetric matrix: for $X \in \mathbf{S}^n$,

$$\lambda_{\max}(X) = \underbrace{\sup_{\|y\|_2 = 1} y^T X y}_{(H/W)}$$

$y = \sum_i c_i u_i$ & $X u_i = \lambda_i u_i$

$u_1 .. u_n$ are orthonormal eigenvectors of X



Composition with scalar functions

composition of $g : \mathbf{R}^n \to \mathbf{R}$ and $h : \mathbf{R} \to \mathbf{R}$:

$$f(x) = h(g(x))$$

f is convex if $\begin{cases} g \text{ convex, } h \text{ convex, } \tilde{h} \text{ nondecreasing} \\ g \text{ concave, } h \text{ convex, } \tilde{h} \text{ nonincreasing} \end{cases}$

Generalisation

- proof (for $n = 1$, differentiable g, h)

Instead do from first principles

$$f''(x) = h''(g(x))g'(x)^2 + h'(g(x))g''(x)$$

- note: monotonicity must hold for extended-value extension $\tilde{h}$

examples

- $\exp g(x)$ is convex if g is convex : How?
- $1/g(x)$ is convex if g is concave and positive : How?

Vector composition

composition of $g : \mathbf{R}^n \to \mathbf{R}^k$ and $h : \mathbf{R}^k \to \mathbf{R}$:

$$f(x) = h(g(x)) = h(g_1(x), g_2(x), \ldots, g_k(x))$$

f is convex if $\begin{cases} g_i \text{ convex, } h \text{ convex, } \tilde{h} \text{ nondecreasing in each argument} \\ g_i \text{ concave, } h \text{ convex, } \tilde{h} \text{ nonincreasing in each argument} \end{cases}$

proof (for $n = 1$, differentiable g, h)

$$f''(x) = g'(x)^T \underbrace{\nabla^2 h(g(x))}_{\geq 0} g'(x) + \underbrace{\nabla h(g(x))^T}_{\geq 0} g''(x)$$

$$[\nabla^2 f]_{ij} = g_i'(x)^T \nabla^2 h(g(x)) g_j'(x) + \nabla h(g(x))^T g''_{ij}(x)$$

H/w: Evaluate expr quad $v^T \nabla^2 f v$

examples

- $\sum_{i=1}^m \log g_i(x)$ is concave if g_i are concave and positive
- $\log \sum_{i=1}^m \exp g_i(x)$ is convex if g_i are convex

Minimization

if $f(x,y)$ is convex in (x,y) and C is a convex set, then

$$g(x) = \inf_{y \in C} f(x,y)$$

is convex

examples

- $f(x,y) = x^T Ax + 2x^T By + y^T Cy$ with

$$\left[\begin{array}{cc} A & B \\ B^T & C \end{array}\right] \succeq 0, \qquad C \succ 0$$

  minimizing over y gives $g(x) = \inf_y f(x,y) = x^T(A - BC^{-1}B^T)x$

  g is convex, hence Schur complement $A - BC^{-1}B^T \succeq 0$

- distance to a set: $\mathbf{dist}(x,S) = \inf_{y\in S} \|x - y\|$ is convex if S is convex

Perspective

the **perspective** of a function $f : \mathbf{R}^n \to \mathbf{R}$ is the function $g : \mathbf{R}^n \times \mathbf{R} \to \mathbf{R}$,

$$g(x,t) = tf(x/t), \qquad \mathbf{dom}\, g = \{(x,t) \mid x/t \in \mathbf{dom}\, f,\ t > 0\}$$

g is convex if f is convex

examples

- $f(x) = x^T x$ is convex; hence $g(x,t) = x^T x/t$ is convex for $t > 0$
- negative logarithm $f(x) = -\log x$ is convex; hence relative entropy $g(x,t) = t\log t - t\log x$ is convex on $\mathbf{R}^2_{++}$
- if f is convex, then

$$g(x) = (c^T x + d) f\left((Ax+b)/(c^T x + d)\right)$$

  is convex on $\{x \mid c^T x + d > 0,\ (Ax+b)/(c^T x+d) \in \mathbf{dom}\, f\}$

Subgradients

- subgradients $: f(y) \geq f(x) + g^T(x)(y-x)$
- strong and weak subgradient calculus
- optimality conditions via subgradients
- directional derivatives

Basic inequality

recall basic inequality for convex differentiable f:

$$f(y) \geq f(x) + \nabla f(x)^T(y - x)$$

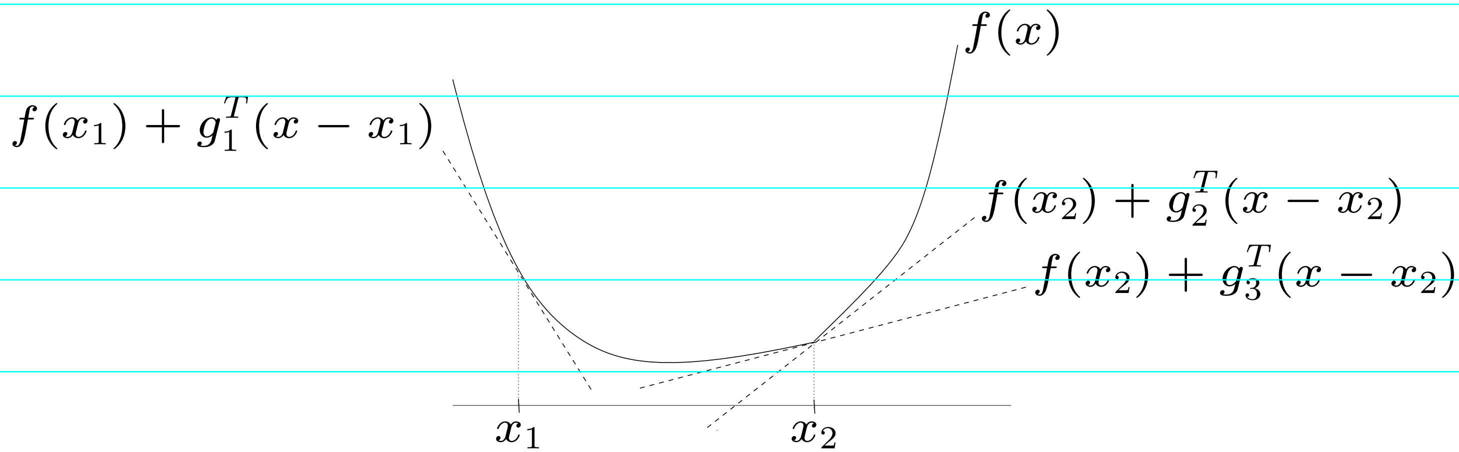
- first-order approximation of f at x is global underestimator
- $(\nabla f(x), -1)$ supports $\mathbf{epi}\, f$ at $(x, f(x))$

what if f is not differentiable?

Subgradient of a function

g is a **subgradient** of f (not necessarily convex) at x if

$$f(y) \geq f(x) + g^T(y - x) \quad \text{for all } y$$



g_2, g_3 are subgradients at x_2; g_1 is a subgradient at x_1

- g is a subgradient of f at x iff $(g, -1)$ supports $\textbf{epi}\, f$ at $(x, f(x))$ Equivalent
- g is a subgradient iff $f(x) + g^T(y - x)$ is a global (affine) underestimator of f
- if f is convex and differentiable, $\nabla f(x)$ is a subgradient of f at x

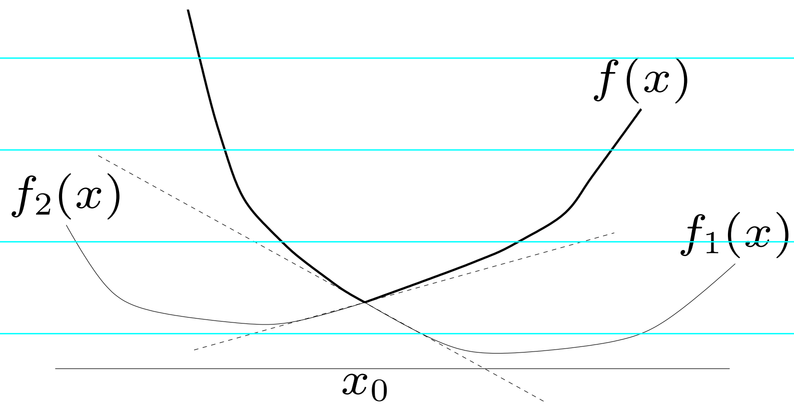
subgradients come up in several contexts:

- algorithms for nondifferentiable convex optimization
- convex analysis, *e.g.*, optimality conditions, duality for nondifferentiable problems

(if $f(y) \leq f(x) + g^T(y - x)$ for all y, then g is a **supergradient**)

Example

$f = \max\{f_1, f_2\}$, with f_1, f_2 convex and differentiable



- $f_1(x_0) > f_2(x_0)$: unique subgradient $g = \nabla f_1(x_0)$
- $f_2(x_0) > f_1(x_0)$: unique subgradient $g = \nabla f_2(x_0)$
- $f_1(x_0) = f_2(x_0)$: subgradients form a line segment $[\nabla f_1(x_0), \nabla f_2(x_0)]$

Subdifferential

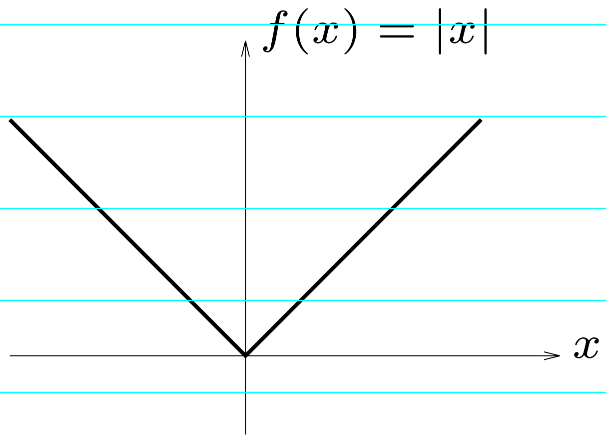
- set of all subgradients of f at x is called the **subdifferential** of f at x, denoted $\partial f(x)$
- $\partial f(x)$ is a closed convex set (can be empty)

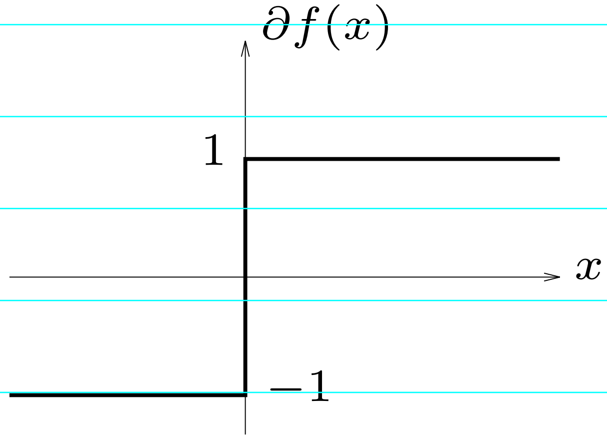
if f is convex,

- $\partial f(x)$ is nonempty, for $x \in \mathbf{relint\,dom}\, f$
- $\partial f(x) = \{\nabla f(x)\}$, if f is differentiable at x
- if $\partial f(x) = \{g\}$, then f is differentiable at x and $g = \nabla f(x)$

Example

$f(x) = |x|$





righthand plot shows $\bigcup \{(x, g) \mid x \in \mathbf{R},\ g \in \partial f(x)\}$

Subgradient calculus

- **weak subgradient calculus**: formulas for finding *one* subgradient $g \in \partial f(x)$
- **strong subgradient calculus**: formulas for finding the whole subdifferential $\partial f(x)$, $i.e.$, *all* subgradients of f at x
- many algorithms for nondifferentiable convex optimization require only *one* subgradient at each step, so weak calculus suffices
- some algorithms, optimality conditions, etc., need whole subdifferential
- roughly speaking: if you can compute $f(x)$, you can usually compute a $g \in \partial f(x)$
- we'll assume that f is convex, and $x \in \mathbf{relint\,dom}\, f$

Some basic rules

- $\partial f(x) = \{\nabla f(x)\}$ if f is differentiable at x
- **scaling:** $\partial(\alpha f) = \alpha \partial f$ (if $\alpha > 0$)
- **addition:** $\partial(f_1 + f_2) = \partial f_1 + \partial f_2$ (RHS is addition of sets)
- **affine transformation of variables:** if $g(x) = f(Ax + b)$, then $\partial g(x) = A^T \partial f(Ax + b)$
- **finite pointwise maximum:** if $f = \max_{i=1,\ldots,m} f_i$, then

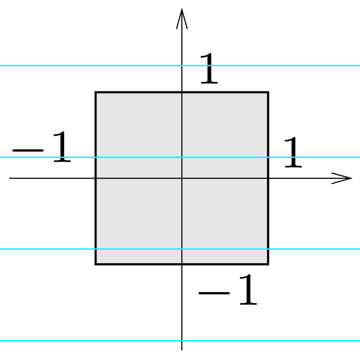
$$\partial f(x) = \mathbf{Co} \bigcup \{\partial f_i(x) \mid f_i(x) = f(x)\},$$

$i.e.$, convex hull of union of subdifferentials of 'active' functions at x

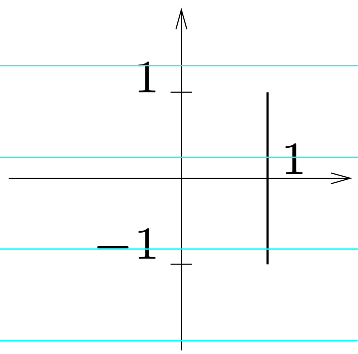
$f(x) = \max\{f_1(x), \ldots, f_m(x)\}$, with $f_1, \ldots, f_m$ differentiable

$$\partial f(x) = \mathbf{Co}\{\nabla f_i(x) \mid f_i(x) = f(x)\}$$

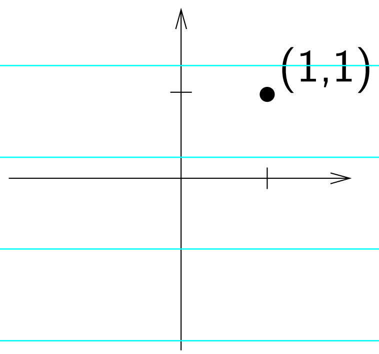
example: $f(x) = \|x\|_1 = \max\{s^T x \mid s_i \in \{-1, 1\}\}$



$\partial f(x)$ at $x = (0, 0)$



at $x = (1, 0)$



at $x = (1, 1)$

Pointwise supremum

if $f = \sup_{\alpha \in \mathcal{A}} f_\alpha$,

$$\mathbf{cl}\,\mathbf{Co} \bigcup \{\partial f_\beta(x) \mid f_\beta(x) = f(x)\} \subseteq \partial f(x)$$

(usually get equality, but requires some technical conditions to hold, *e.g.*, $\mathcal{A}$ compact, f_α cts in x and α)

roughly speaking, $\partial f(x)$ is closure of convex hull of union of subdifferentials of active functions

Weak rule for pointwise supremum

$$f = \sup_{\alpha \in \mathcal{A}} f_\alpha$$

- find *any* β for which $f_\beta(x) = f(x)$ (assuming supremum is achieved)
- choose *any* $g \in \partial f_\beta(x)$
- then, $g \in \partial f(x)$

example

$$f(x) = \lambda_{\max}(A(x)) = \sup_{\|y\|_2=1} y^T A(x) y$$

where $A(x) = A_0 + x_1 A_1 + \cdots + x_n A_n$, $A_i \in \mathbf{S}^k$

- f is pointwise supremum of $g_y(x) = y^T A(x) y$ over $\|y\|_2 = 1$
- g_y is affine in x, with $\nabla g_y(x) = (y^T A_1 y, \ldots, y^T A_n y)$
- hence, $\partial f(x) \supseteq \mathbf{Co}\,\{\nabla g_y \mid A(x)y = \lambda_{\max}(A(x))y,\ \|y\|_2 = 1\}$
  (in fact equality holds here)

to find **one** subgradient at x, can choose **any** unit eigenvector y associated with $\lambda_{\max}(A(x))$; then

$$(y^T A_1 y, \ldots, y^T A_n y) \in \partial f(x)$$

Expectation

- $f(x) = \mathbf{E}\, f(x,u)$, with f convex in x for each u, u a random variable
- for each u, choose *any* $g_u \in \partial_f(x,u)$ (so $u \mapsto g_u$ is a function)
- then, $g = \mathbf{E}\, g_u \in \partial f(x)$

Monte Carlo method for (approximately) computing $f(x)$ and *a* $g \in \partial f(x)$:

- generate independent samples $u_1, \ldots, u_K$ from distribution of u
- $f(x) \approx (1/K) \sum_{i=1}^{K} f(x, u_i)$
- for each i choose $g_i \in \partial_x f(x, u_i)$
- $g = (1/K) \sum_{i=1}^{K} g_i$ is an (approximate) subgradient
  (more on this later)

Minimization

define $g(y)$ as the optimal value of

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq y_i, \quad i = 1, \ldots, m \end{array}$$

(f_i convex; variable x)

with $\lambda^\star$ an optimal dual variable, we have

$$g(z) \geq g(y) - \sum_{i=1}^{m} \lambda_i^\star (z_i - y_i)$$

$i.e.$, $-\lambda^\star$ is a subgradient of g at y

Composition

- $f(x) = h(f_1(x), \ldots, f_k(x))$, with h convex nondecreasing, f_i convex
- find $q \in \partial h(f_1(x), \ldots, f_k(x))$, $g_i \in \partial f_i(x)$
- then, $g = q_1 g_1 + \cdots + q_k g_k \in \partial f(x)$
- reduces to standard formula for differentiable h, f_i

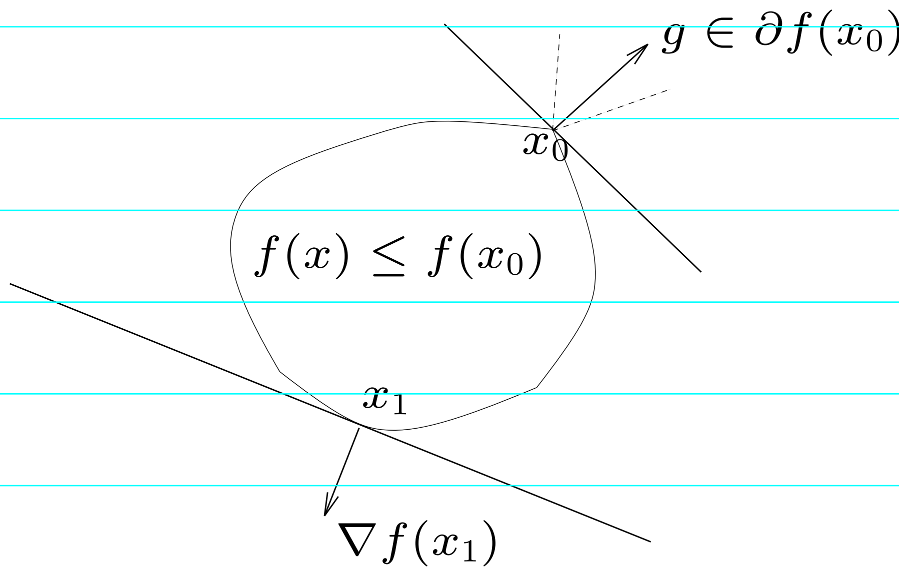
proof:

$$\begin{array}{rcl} f(y) & = & h(f_1(y), \ldots, f_k(y)) \\ & \geq & h(f_1(x) + g_1^T(y-x), \ldots, f_k(x) + g_k^T(y-x)) \\ & \geq & h(f_1(x), \ldots, f_k(x)) + q^T(g_1^T(y-x), \ldots, g_k^T(y-x)) \\ & = & f(x) + g^T(y-x) \end{array}$$

Subgradients and sublevel sets

g is a subgradient at x means $f(y) \geq f(x) + g^T(y - x)$

hence $f(y) \leq f(x) \Longrightarrow g^T(y - x) \leq 0$



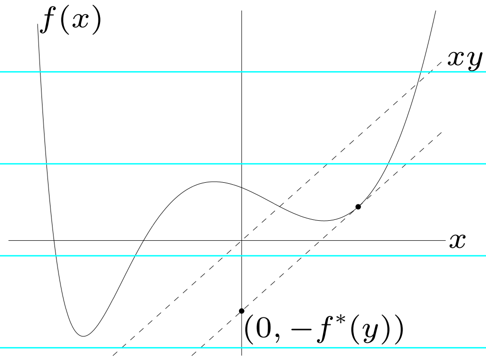
- f differentiable at x_0: $\nabla f(x_0)$ is normal to the sublevel set $\{x \mid f(x) \leq f(x_0)\}$

- f nondifferentiable at x_0: subgradient defines a supporting hyperplane to sublevel set through x_0

The conjugate function

the **conjugate** of a function f is

$$f^*(y) = \sup_{x \in \mathbf{dom}\, f} (y^T x - f(x))$$



- f^* is convex (even if f is not)
- will be useful in chapter 5

examples

- negative logarithm $f(x) = -\log x$

$$\begin{aligned} f^*(y) &= \sup_{x>0}(xy + \log x) \\ &= \begin{cases} -1 - \log(-y) & y < 0 \\ \infty & \text{otherwise} \end{cases} \end{aligned}$$

- strictly convex quadratic $f(x) = (1/2)x^T Q x$ with $Q \in \mathbf{S}^n_{++}$

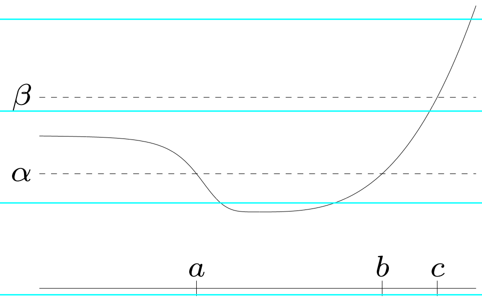
$$\begin{aligned} f^*(y) &= \sup_x (y^T x - (1/2)x^T Q x) \\ &= \frac{1}{2} y^T Q^{-1} y \end{aligned}$$

Quasiconvex functions

$f : \mathbf{R}^n \to \mathbf{R}$ is quasiconvex if $\mathbf{dom}\, f$ is convex and the sublevel sets

$$S_\alpha = \{x \in \mathbf{dom}\, f \mid f(x) \leq \alpha\}$$

are convex for all α



- f is quasiconcave if $-f$ is quasiconvex
- f is quasilinear if it is quasiconvex and quasiconcave

Examples

- $\sqrt{|x|}$ is quasiconvex on $\mathbf{R}$
- $\operatorname{ceil}(x) = \inf\{z \in \mathbf{Z} \mid z \geq x\}$ is quasilinear
- $\log x$ is quasilinear on $\mathbf{R}_{++}$
- $f(x_1, x_2) = x_1 x_2$ is quasiconcave on $\mathbf{R}^2_{++}$
- linear-fractional function

$$f(x) = \frac{a^T x + b}{c^T x + d}, \qquad \mathbf{dom}\, f = \{x \mid c^T x + d > 0\}$$

  is quasilinear

- distance ratio

$$f(x) = \frac{\|x - a\|_2}{\|x - b\|_2}, \qquad \mathbf{dom}\, f = \{x \mid \|x - a\|_2 \leq \|x - b\|_2\}$$

  is quasiconvex

internal rate of return

- cash flow $x = (x_0, \ldots, x_n)$; x_i is payment in period i (to us if $x_i > 0$)
- we assume $x_0 < 0$ and $x_0 + x_1 + \cdots + x_n > 0$
- present value of cash flow x, for interest rate r:

$$\mathrm{PV}(x, r) = \sum_{i=0}^{n} (1+r)^{-i} x_i$$

- internal rate of return is smallest interest rate for which $\mathrm{PV}(x, r) = 0$:

$$\mathrm{IRR}(x) = \inf\{r \geq 0 \mid \mathrm{PV}(x, r) = 0\}$$

IRR is quasiconcave: superlevel set is intersection of halfspaces

$$\mathrm{IRR}(x) \geq R \quad \Longleftrightarrow \quad \sum_{i=0}^{n} (1+r)^{-i} x_i \geq 0 \text{ for } 0 \leq r \leq R$$

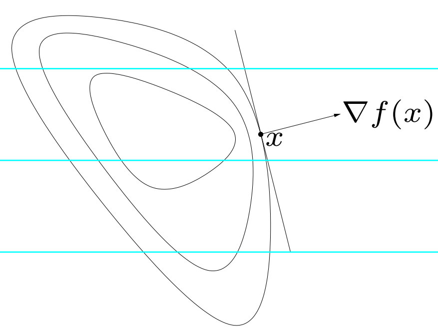
Properties

modified Jensen inequality: for quasiconvex f

$$0 \leq \theta \leq 1 \quad \Longrightarrow \quad f(\theta x + (1-\theta) y) \leq \max\{f(x), f(y)\}$$

first-order condition: differentiable f with cvx domain is quasiconvex iff

$$f(y) \leq f(x) \quad \Longrightarrow \quad \nabla f(x)^T (y - x) \leq 0$$



sums of quasiconvex functions are not necessarily quasiconvex

Log-concave and log-convex functions

a positive function f is log-concave if $\log f$ is concave:

$$f(\theta x + (1-\theta)y) \geq f(x)^\theta f(y)^{1-\theta} \quad \text{for } 0 \leq \theta \leq 1$$

f is log-convex if $\log f$ is convex

- powers: x^a on $\mathbf{R}_{++}$ is log-convex for $a \leq 0$, log-concave for $a \geq 0$
- many common probability densities are log-concave, *e.g.*, normal:

$$f(x) = \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} e^{-\frac{1}{2}(x-\bar{x})^T \Sigma^{-1}(x-\bar{x})}$$

- cumulative Gaussian distribution function Φ is log-concave

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2}\, du$$

Properties of log-concave functions

- twice differentiable f with convex domain is log-concave if and only if

$$f(x)\nabla^2 f(x) \preceq \nabla f(x) \nabla f(x)^T$$

  for all $x \in \mathbf{dom}\, f$

- product of log-concave functions is log-concave
- sum of log-concave functions is not always log-concave
- integration: if $f : \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}$ is log-concave, then

$$g(x) = \int f(x, y)\, dy$$

  is log-concave (not easy to show)

consequences of integration property

- convolution $f * g$ of log-concave functions f, g is log-concave

$$(f * g)(x) = \int f(x-y)g(y)dy$$

- if $C \subseteq \mathbf{R}^n$ convex and y is a random variable with log-concave pdf then

$$f(x) = \mathbf{prob}(x + y \in C)$$

  is log-concave

  proof: write $f(x)$ as integral of product of log-concave functions

$$f(x) = \int g(x+y)p(y)\,dy, \qquad g(u) = \begin{cases} 1 & u \in C \\ 0 & u \notin C, \end{cases}$$

  p is pdf of y

example: yield function

$$Y(x) = \mathbf{prob}(x + w \in S)$$

- $x \in \mathbf{R}^n$: nominal parameter values for product
- $w \in \mathbf{R}^n$: random variations of parameters in manufactured product
- S: set of acceptable values

if S is convex and w has a log-concave pdf, then

- Y is log-concave
- yield regions $\{x \mid Y(x) \geq \alpha\}$ are convex

Convexity with respect to generalized inequalities

$f : \mathbf{R}^n \to \mathbf{R}^m$ is K-convex if $\mathbf{dom}\, f$ is convex and

$$f(\theta x + (1-\theta)y) \preceq_K \theta f(x) + (1-\theta) f(y)$$

for x, $y \in \mathbf{dom}\, f$, $0 \le \theta \le 1$

example $f : \mathbf{S}^m \to \mathbf{S}^m$, $f(X) = X^2$ is $\mathbf{S}^m_+$-convex

proof: for fixed $z \in \mathbf{R}^m$, $z^T X^2 z = \|Xz\|_2^2$ is convex in X, *i.e.*,

$$z^T(\theta X + (1-\theta)Y)^2 z \le \theta z^T X^2 z + (1-\theta) z^T Y^2 z$$

for $X, Y \in \mathbf{S}^m$, $0 \le \theta \le 1$

therefore $(\theta X + (1-\theta)Y)^2 \preceq \theta X^2 + (1-\theta) Y^2$

Procedure 4 [Finding extreme values on closed, bounded intervals]:
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.*

For example, 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 $f(1) = f(\frac{1}{2}) = 1$.

Definition 21 [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

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

Theorem 54 *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 $f'(b) = -\infty$.*

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

Theorem 55 *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 number $c \in (a,b)$, 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 number $c \in (a,b)$, then $f(c)$ is the minimum value of f on $[a,b]$.*

Theorem 56 *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 $\mathcal{I}$.

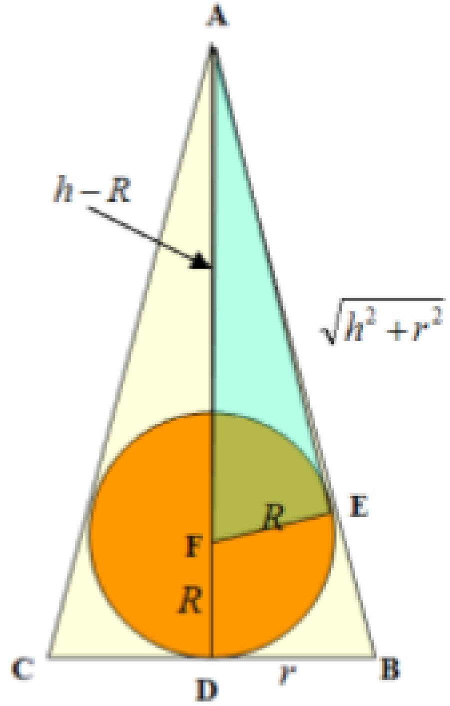


Figure 4.11: Illustrating the constraints for the optimization problem of finding the cone with minimum volume that can contain a sphere of radius R.

Theorem 61 *Let $f : \mathcal{D} \to \Re$ where $\mathcal{D} \subseteq \Re^n$. Let $f(\mathbf{x})$ have continuous partial derivatives and continuous mixed partial derivatives in an open ball $\mathcal{R}$ containing a point $\mathbf{x}^*$ where $\nabla f(\mathbf{x}^*) = 0$. Let $\nabla^2 f(\mathbf{x})$ denote an $n \times n$ matrix of mixed partial derivatives of f evaluated at the point $\mathbf{x}$, such that the ij^{th} entry of the matrix is $f_{x_i x_j}$. The matrix $\nabla^2 f(\mathbf{x})$ is called the Hessian matrix. The Hessian matrix is symmetric[6]. Then,*

- *If $\nabla^2 f(\mathbf{x}^*)$ is positive definite, $\mathbf{x}^*$ is a local minimum.*
- *If $\nabla^2 f(\mathbf{x}^*)$ is negative definite (that is if $-\nabla^2 f(\mathbf{x}^*)$ is positive definite), $\mathbf{x}^*$ is a local maximum.*

Theorem 62 *Let $f : \mathcal{D} \to \Re$ where $\mathcal{D} \subseteq \Re^n$. Let $f(\mathbf{x})$ have continuous partial derivatives and continuous mixed partial derivatives in an open region $\mathcal{R}$ containing a point $\mathbf{x}^*$ where $\nabla f(\mathbf{x}^*) = 0$. Then,*

- *If $\mathbf{x}^*$ is a point of local minimum, $\nabla^2 f(\mathbf{x}^*)$ must be positive semi-definite.*
- *If $\mathbf{x}^*$ is a point of local maximum, $\nabla^2 f(\mathbf{x}^*)$ must be negative semi-definite (that is, $-\nabla^2 f(\mathbf{x}^*)$ must be positive semi-definite).*

Corollary 63 *Let $f : \mathcal{D} \to \Re$ where $\mathcal{D} \subseteq \Re^n$. Let $f(\mathbf{x})$ have continuous partial derivatives and continuous mixed partial derivatives in an open region $\mathcal{R}$ containing a point $\mathbf{x}^*$ where $\nabla f(\mathbf{x}^*) = 0$. If $\nabla^2 f(\mathbf{x}^*)$ is neither positive semi-definite nor negative semi-definite (that is, some of its eigenvalues are positive and some negative), then $\mathbf{x}^*$ is a saddle point.*

Theorem 64 *Let the partial and second partial derivatives of $f(x_1, x_2)$ be continuous on a disk with center (a, b) and suppose $f_{x_1}(a, b) = 0$ and $f_{x_2}(a, b) = 0$ so that (a, b) is a critical point of f. Let $D(a, b) = f_{x_1x_1}(a, b)f_{x_2x_2}(a, b) - [f_{x_1x_2}(a, b)]^2$. Then*[7]*,*

- *If $D > 0$ and $f_{x_1x_1}(a, b) > 0$, then $f(a, b)$ is a local minimum.*
- *Else if $D > 0$ and $f_{x_1x_1}(a, b) < 0$, then $f(a, b)$ is a local maximum.*
- *Else if $D < 0$ then (a, b) is a saddle point.*

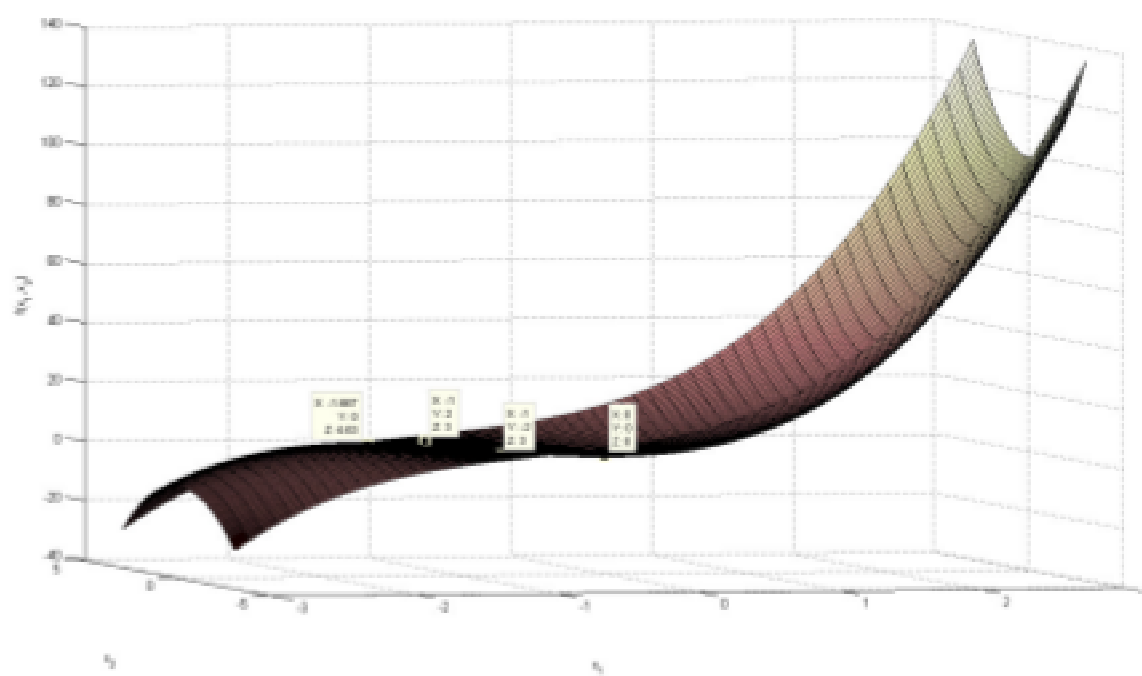


Figure 4.22: Plot of the function $2x_1^3 + x_1x_2^2 + 5x_1^2 + x_2^2$ showing the four critical points.

We saw earlier that the critical points for $f(x_1, x_2) = 2x_1^3+x_1x_2^2+5x_1^2+x_2^2$ are $(0, 0)$, $(-\frac{5}{3}, 0)$, $(-1, 2)$ and $(-1, -2)$. To determine which of these correspond to local extrema and which are saddle, we first compute compute the partial derivatives of f:

$f_{x_1x_1}(x_1, x_2) = 12x_1 + 10$
$f_{x_2x_2}(x_1, x_2) = 2x_1 + 2$
$f_{x_1x_2}(x_1, x_2) = 2x_2$

Using theorem 64, we can verify that $(0, 0)$ corresponds to a local minimum, $(-\frac{5}{3}, 0)$ corresponds to a local maximum while $(-1, 2)$ and $(-1, -2)$ correspond to saddle points. Figure 4.22 shows the plot of the function while pointing ou the four critical points.

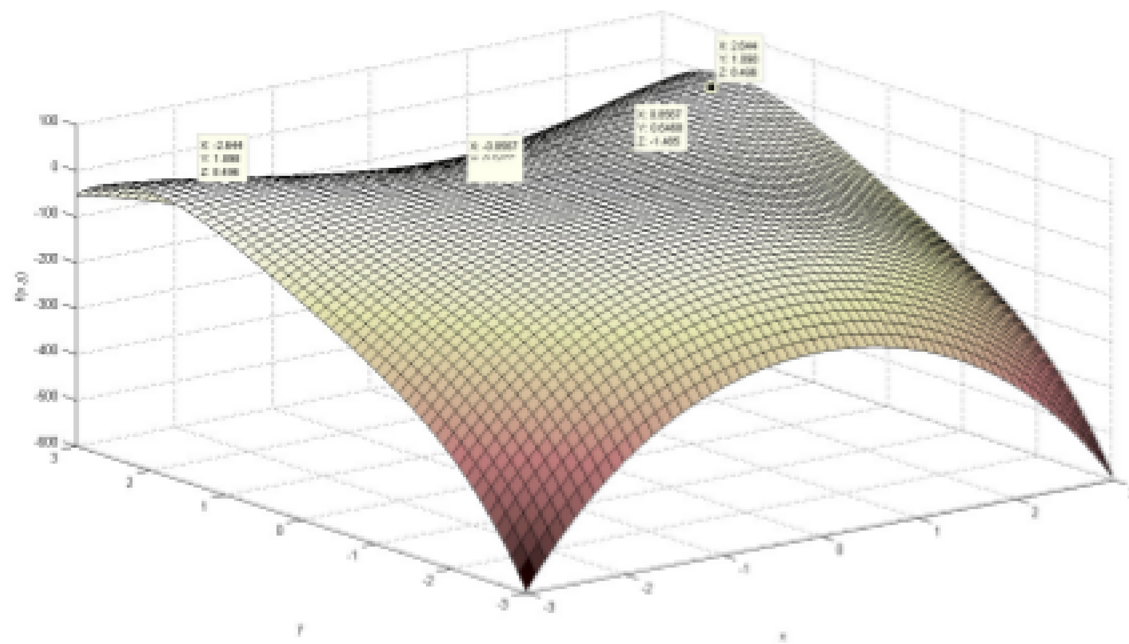


Figure 4.23: Plot of the function $10x^2y - 5x^2 - 4y^2 - x^4 - 2y^4$ showing the four critical points.

Consider a significantly harder function $f(x, y) = 10x^2y - 5x^2 - 4y^2 - x^4 - 2y^4$. Let us find and classify its critical points. The gradient vector is $\nabla f(x, y) = [20xy - 10x - 4x^3, \quad 10x^2 - 8y - 8y^3]$. The critical points correspond to solutions of the simultaneous set of equations

$$\begin{aligned} 20xy - 10x - 4x^3 &= 0 \\ 10x^2 - 8y - 8y^3 &= 0 \end{aligned} \tag{4.15}$$

One of the solutions corresponds to solving the system $-8y^3 + 42y - 25 = 0$[8] and $10x^2 = 50y - 25$, which have four real solutions[9], *viz.*, $(0.8567, 0.646772)$, $(-0.8567, 0.646772)$, $(2.6442, 1.898384)$, and $(-2.6442, 1.898384)$. Another real solution is $(0, 0)$. The mixed partial derivatives of the function are

$$\begin{aligned} f_{xx} &= 20y - 10 - 12x^2 \\ f_{xy} &= 20x \\ f_{yy} &= -8 - 24y^2 \end{aligned} \tag{4.16}$$

Using theorem 64, we can verify that $(2.6442, 1.898384)$ and $(-2.6442, 1.898384)$ correspond to local maxima whereas $(0.8567, 0.646772)$ and $(-0.8567, 0.646772)$ correspond to saddle points. This is illustrated in Figure 4.23