H/W: Subgradient of $\|x\|_1 = f(x)$ $\quad x \in \mathbb{R}^n$

$$f(x) = \|x\|_1 = \max_{i=1\dots N} \{ f_1(x), f_2(x) \dots f_i(x) \dots f_N(x) \}$$

$f_1(x) = S_1^T x$, $\quad f_2(x) = S_2^T x$, $\quad f_N(x) = S_N^T x$ $\qquad N = 2^n$

$$S_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \end{bmatrix} \quad S_2 = \begin{bmatrix} -1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \quad S_N = \begin{bmatrix} -1 \\ -1 \\ \vdots \\ -1 \end{bmatrix}$$

if No component of $x = 0$ then $S = \begin{bmatrix} \text{sgn}(x_1) \\ \text{sgn}(x_2) \\ \vdots \\ \text{sgn}(x_n) \end{bmatrix}$

In general if $f(x) = S_1'^T x = S_2'^T x = \dots = S_k'^T x$

then $\partial f(x) = \text{conv}\{S_1', S_2' \dots S_k'\}$ $\dots$ $(\partial f(x))_i = \begin{cases} +1 & \text{if } x_i > 0 \\ -1 & \text{if } x_i < 0 \\ \theta(-1) + (1-\theta)(1) & \text{if } x_i = 0 \\ \theta \in [0,1] & \end{cases}$

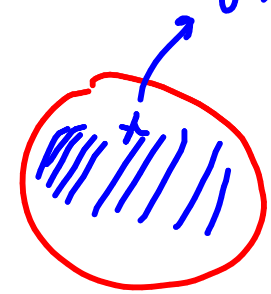
ie $\partial f(x) = \{ d \mid \|d\|_\infty \le 1, \ d^T x = \|x\|_1 \}$

Subdifferential

$f(y) \geq f(x) + g_x^T(y-x) \quad \forall y \in \mathbf{dom} f$

- 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)

$\partial f(x) \neq \phi$



D

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 $g = \nabla f(x)$

Consider supporting hyperplane at $(x, f(x))$ to epi(f) ? H/W

3 why $x \in$ relint ?

$\forall (y,z) \in \text{epi}(f) \quad a^T \begin{bmatrix} y \\ z \end{bmatrix} + \cancel{b} \leq a^T \begin{bmatrix} x \\ f(x) \end{bmatrix} + \cancel{b}$

1 How do I get $f(y)$ into inequality

2 $g_x^T y - f(y) \leq g_x^T x - f(x)$

Completed solution

Let $a = \begin{bmatrix} c \\ d \end{bmatrix}$ s.t $\begin{bmatrix} c \\ d \end{bmatrix}^T \begin{bmatrix} y \\ z \end{bmatrix} \le \begin{bmatrix} c \\ d \end{bmatrix}^T \begin{bmatrix} x \\ f(x) \end{bmatrix} \quad \forall\ (y,z) \in \operatorname{epi}(f)$

Not both c & d are 0

$\Downarrow$ (if $y=x$ $\ z \ge f(x) \Rightarrow d \le 0$)

$$d \le 0 \ \& \ c^T(y-x) + d\big(z - f(x)\big) \le \underbrace{c^T(y-x) + d\big(f(y) - f(x)\big) \le 0}_{\forall y \in \operatorname{dmn} f}$$

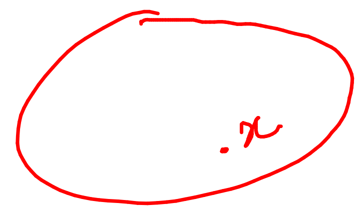
If $d \ne 0$, we can divide by d (while reversing the inequality)

$$f(y) \ge f(x) - \left(\tfrac{c}{d}\right)^T (y - x) \quad \forall y \in \operatorname{dmn} f$$

$\Downarrow$

$c/d \in \partial f$

If $d=0$ then $c^T(y-x) \le 0 \ \forall y \in \operatorname{dmn} f$



which is impossible since $x \in \operatorname{relint} \operatorname{dmn} f$ (no single vector c can make an obtuse angle with $(y-x)\ \forall\ y \in \operatorname{dmn} f$)

$g_x \in \partial f(x)$ iff $\forall\, y \in \mathrm{dmn}\, f \quad f(y) \geq f(x) + g_x^T(y-x)$

$$f(x) - g_x^T x \leq f(y) - g_x^T y \quad \forall y$$

$$g_x^T x - f(x) \geq g_x^T y - f(y) \quad \forall y$$

$$\Updownarrow$$

$$g_x^T x - f(x) \geq \max_y \; g_x^T y - f(y) = f^*(g_x)$$

If $\partial f(x) \neq \emptyset$ (ie g_x exists)

then

$g_x^T x - f(x) = f^*(g_x)$

(since $\max_y$ includes max over x)

If f is differentiable: $g_x = \nabla f(x)$

then $f^*(\nabla f(x)) = \nabla^T f(x)\, x - f(x)$

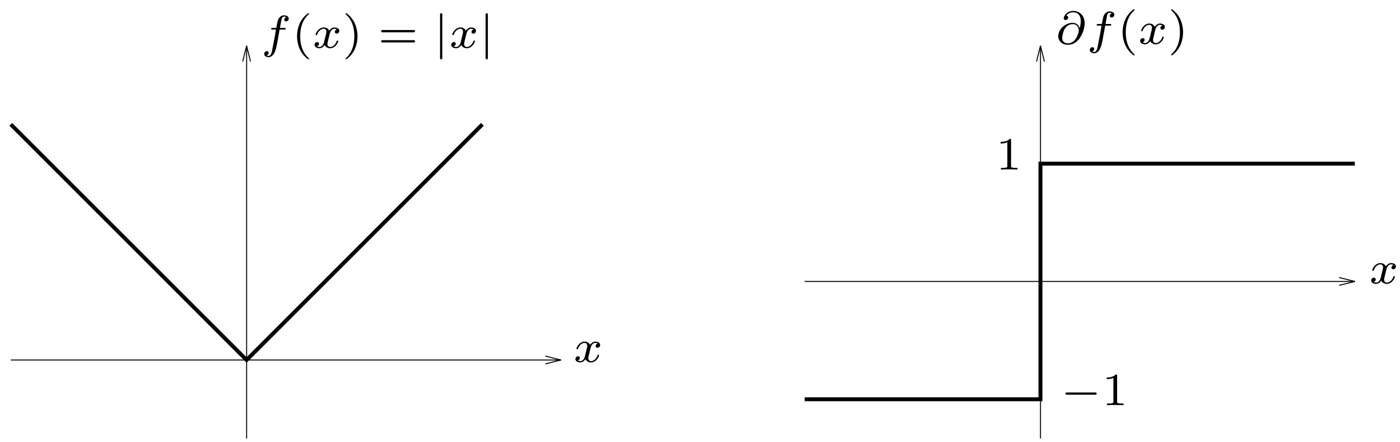
convex conjugate

If f is differentiable: $g_x = \nabla f(x)$

& $\nabla^T f(x)\, x - f(x) \geq f^*(\nabla f(x))$

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 → as in case of Lasso (we will see)
- 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 $x \in \mathbf{relint\,dom}\, f$

Gradient $\lim_{h \to 0} \frac{f(x+h)-f(x)}{h}$

Some basic rules

- $\partial f(x) = \{\nabla f(x)\}$ if f is differentiable at x → H/W: Prove by contradiction $f(y) \geq f(x) + g_x^T(y-x)$
- **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 $\partial g(x) = A^T \partial f(Ax + b)$

Prove all by invoking $f(y) \geq f(x) + g_x^T(y-x)$

- **finite pointwise maximum:** if $f = \max_{i=1,\ldots,m} f_i$, then

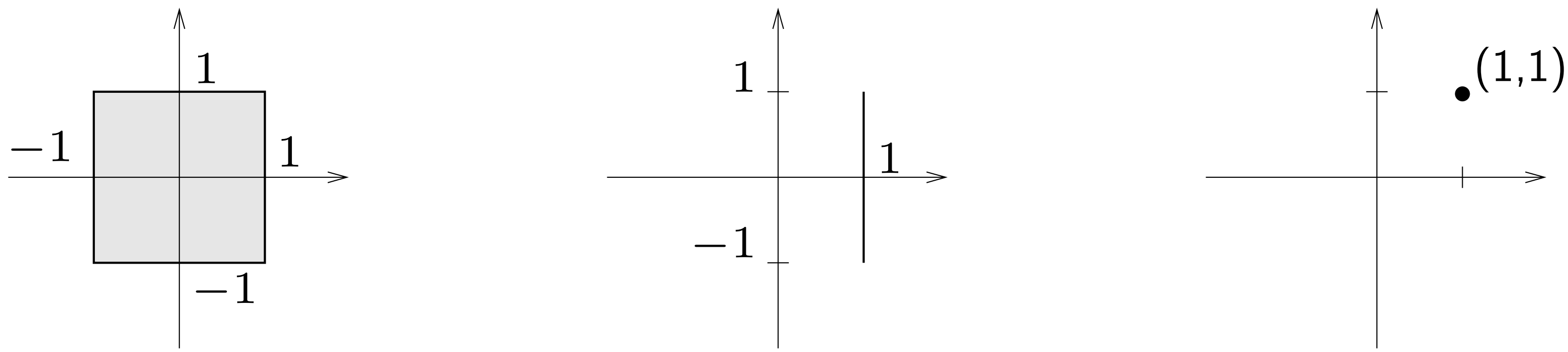
$$\partial f(x) = \mathbf{Co} \bigcup \{\partial f_i(x) \mid f_i(x) = f(x)\},$$

$i.e.$, 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: $f(x) = \|x\|_1 = \max\{s^T x \mid s_i \in \{-1, 1\}\}$ $\rightarrow g_x = \begin{bmatrix} \mathrm{sgn}(x_1) \\ \vdots \\ \mathrm{sgn}(x_n) \end{bmatrix}$



$\partial f(x)$ at $x = (0, 0)$ at $x = (1, 0)$ at $x = (1, 1)$

What abt local maxima/minima & subgradient?

① $\nabla f(x) = 0$ & f is convex then x is global min

What if $g_x = 0$?

$$f(y) \geq f(x) + g_x^T(y-x) \quad \forall y$$

if $g_x = 0$ then $f(y) \geq f(x) \Rightarrow x$ is pt of global min

Regularizer

Eg: $\min_x \frac{1}{2}\|y-x\|^2 + \lambda\|x\|_1$ $\quad \left(\underset{x}{\operatorname{argmin}} \|y-x\|^2 + \lambda\|x\|_1 = x^*\right)$

$\lambda \geq 0$

I will suggest a soln by setting "some" $g_x = 0$

$$x_i^* = \begin{cases} -\lambda + y_i & \text{if } y_i > \lambda \\ 0 & \text{if } -\lambda \leq y_i \leq \lambda \\ \lambda + y_i & \text{if } y_i < -\lambda \end{cases} \quad ✱$$

Higher $\lambda \Rightarrow$ more x_i's are zeros

lots of zeros esp if several $|y_i| \leq \lambda$.. sparsity

Why should this be imp for minimization? 2 ways of answering

① $g_x = \frac{1}{2}\nabla(\|y-x\|^2) + \lambda\partial\|x\|_1$

$= (x-y) + \lambda \begin{bmatrix} \text{sign}(x_1) \\ \vdots \\ \text{sign}(x_n) \end{bmatrix}$

② $\min_{x_i} \frac{1}{2}(y_i - x_i)^2 + \lambda|x_i|$

for each i $\quad g_{x_i} = (x_i - y_i) + \lambda\,\text{sign}(x_i)$

In either case. 1 or 2, setting $g_x = 0$ or $g_{x_i} = 0$ for each i, & checking that $*$ satisfies this equation,

---

Another example

Maximum eigenvalue of a symmetric matrix

$$f(x) = \lambda_{max}(A(x)) \quad .. \quad A(x) = A_0 + x_1 A_1 + \ldots + x_n A_n$$

$$\& \; A_i \in S^m$$

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

Index set I over fns

each function is affine in x for fixed y & has gradient $\nabla f_y(x) = (y^T A_1 y, \ldots y^T A_n y)$

| http://en.wikipedia.org/wiki/Rayleigh_quotient |

Active fns $y^T A(x) y$ are the ones for which y is (normalised) eigenvector for max eigenvalue λ_{max} of $A(x)$

$$\therefore g_x = (y^T A_1 y, \ldots y^T A_n y)$$

$$\boxed{\begin{array}{l} \min\ f(x) \\ \text{st}\ \ g_i(x) \le 0 \end{array}}$$

option 1 (0|1)

$$I_{g_i}(x) \to \begin{cases} 0 & \text{if } g_i(x) \le 0 \\ \infty & \text{o/w} \end{cases}$$

option 2 (continuous)

Let $C_i = \{x \mid g_i(x) \le 0\}$ are convex sets & let

$\text{dist}(x, C_i) = \min\{\|x-u\| : u \in C\}$

If C_i is closed, convex then $\exists$ unique $u^* \in C$ that minimizes $\|x-u\|$. Let us call $u^* = P_{C_i}(x)$ so that

$\text{dist}(x, C_i) = \|x - P_{C_i}(x)\|$

We are interested in $\hat{x}$ st $g_1(x) \le 0, \ldots g_m(x) \le 0$

ie $\hat{x} \in C_1 \cap C_2 \ldots \cap C_m$

Claim: (If $\hat{x}$ exists)

$$\min_{x \in \mathbb{R}^n} \underbrace{\max_{i=1\ldots m} \text{dist}(x, C_i)}_{\text{call it } D(x)} = 0$$

$D(\hat{x}) = 0$

$$\nabla \text{dist}(x, C_i) = \frac{x - P_{C_i}(x)}{\|x - P_{C_i}(x)\|}$$

if $D(x) = \text{dist}(x, C_i) \neq 0$ then $\dfrac{x - P_{C_i}(x)}{\|x - P_{C_i}(x)\|} \in \partial D(x)$

if g_i is convex, dmn I_{g_i} is convex & $I_{g_i}(x)$ is a convex fn

$$\partial I_{g_i}(x) = \{d \in \mathbb{R}^n \mid I_{g_i}(y) \ge I_{g_i}(x) + d^T(y - x)\ \forall y\}$$

∞ if $g_i(y) > 0$ so no issues

0 if $g_i(y) \le 0$

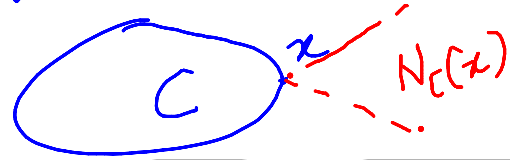
(if $g_i(x) \le 0$) $\{d \in \mathbb{R}^n \mid 0 \ge d^T(y-x)\ \forall y \text{ s.t } g_i(y) \le 0\}$

$$= \underbrace{\{d \in \mathbb{R}^n \mid d^T x \ge d^T y\ \forall y \text{ s.t } g_i(y) \le 0\}}$$

Normal cone $N_C(x)$ to C at point x.

(1) If $x \in \text{int}(\text{dmn } g_i)$ then $N_C(x) = \{0\}$ i.e no nontrivial descent possible (2) Otherwise

$$N_C(x) = \{d \in \mathbb{R}^n \mid d^T x \ge d^T y\ \forall y \in C\}$$



First-order condition

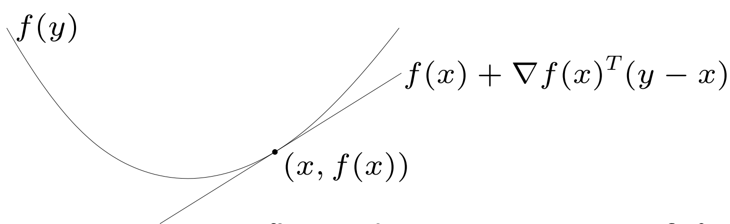
f is **differentiable** if $\mathbf{dom}\, f$ is open and the gradient

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

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

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

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



first-order approximation of f is global underestimator

Second-order conditions

f is **twice differentiable** if $\mathbf{dom}\, f$ is open and the Hessian $\nabla^2 f(x) \in \mathbf{S}^n$,

$$\nabla^2 f(x)_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}, \quad i, j = 1, \ldots, n,$$

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

2nd-order conditions: for twice differentiable f with convex domain

- f is convex if and only if

$$\nabla^2 f(x) \succeq 0 \quad \text{for all } x \in \mathbf{dom}\, f$$

- if $\nabla^2 f(x) \succ 0$ for all $x \in \mathbf{dom}\, f$, then f is strictly convex

Theorem 61 *Let $f : \mathcal{D} \rightarrow \Re$ 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.*

Proof: Since the mixed partial derivatives of f are continuous in an open ball containing $\mathcal{R}$ containing $\mathbf{x}^*$ and since $\nabla^2 f(\mathbf{x}^*) \succ 0$, it can be shown that there exists an $\epsilon > 0$, with $\mathcal{B}(\mathbf{x}^*, \epsilon) \subseteq \mathcal{R}$ such that for all $||\mathbf{h}|| < \epsilon$, $\nabla^2 f(\mathbf{x}^* + \mathbf{h}) \succ 0$. Consider an increment vector $\mathbf{h}$ such that $(\mathbf{x}^* + \mathbf{h}) \in \mathcal{B}(\mathbf{x}^*, \epsilon)$. Define $g(t) = f(\mathbf{x}^* + t\mathbf{h}) : [0, 1] \rightarrow \Re$. Using the chain rule,

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

Since f has continuous partial and mixed partial derivatives, g' is a differentiable function of t and

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

Since g and g' are continous on $[0, 1]$ and g' is differentiable on $(0, 1)$, we can make use of the Taylor's theorem (45) with $n = 1$ and $a = 0$ to obtain:

$$g(1) = g(0) + g'(0) + \frac{1}{2} g''(c)$$

for some $c \in (0,1)$. Writing this equation in terms of f gives

$$f(\mathbf{x}^* + \mathbf{h}) = f(\mathbf{x}^*) + \mathbf{h}^T \nabla f(\mathbf{x}^*) + \frac{1}{2}\mathbf{h}^T \nabla^2 f(\mathbf{x}^* + c\mathbf{h})\mathbf{h}$$

We are given that $\nabla f(\mathbf{x}^*) = 0$. Therefore,

$$f(\mathbf{x}^* + \mathbf{h}) - f(\mathbf{x}^*) = \frac{1}{2}\mathbf{h}^T \nabla^2 f(\mathbf{x}^* + c\mathbf{h})\mathbf{h}$$

The presence of an extremum of f at $\mathbf{x}^*$ is determined by the sign of $f(\mathbf{x}^* + \mathbf{h}) - f(\mathbf{x}^*)$. By virtue of the above equation, this is the same as the sign of $H(c) = \mathbf{h}^T \nabla^2 f(\mathbf{x}^* + c\mathbf{h})\mathbf{h}$. Because the partial derivatives of f are continuous in $\mathcal{R}$, if $H(0) \neq 0$, the sign of $H(c)$ will be the same as the sign of $H(0) = \mathbf{h}^T \nabla^2 f(\mathbf{x}^*)\mathbf{h}$ for $\mathbf{h}$ with sufficiently small components (*i.e.*, since the function has continuous partial and mixed partial derivatives at $(\mathbf{x}^*$, the hessian will be positive in some small neighborhood around $(\mathbf{x}^*)$. Therefore, if $\nabla^2 f(\mathbf{x}^*)$ is positive definite, we are guaranteed to have $H(0)$ positive, implying that f has a local minimum at $\mathbf{x}^*$. Similarly, if $-\nabla^2 f(\mathbf{x}^*)$ is positive definite, we are guaranteed to have $H(0)$ negative, implying that f has a local maximum at $\mathbf{x}^*$. $\square$

Theorem 61 gives sufficient conditions for local maxima and minima of functions of multiple variables. Along similar lines of the proof of theorem 61, we can prove necessary conditions for local extrema in theorem 62.

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).*

Theorem 79 *A twice differential function $f : \mathcal{D} \rightarrow \Re$ 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*

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

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.

Proof: We will prove only the first statement in the theorem; the other two statements are proved in a similar manner.

Necessity: Suppose f is a convex function, and consider a point $\mathbf{x} \in \mathcal{D}$. We will prove that for any $\mathbf{h} \in \Re^n$, $\mathbf{h}^T \nabla^2 f(\mathbf{x})\mathbf{h} \geq 0$. Since f is convex, by theorem 75, we have

$$f(\mathbf{x}+t\mathbf{h}) \geq f(\mathbf{x}) + t\nabla^T f(\mathbf{x})\mathbf{h} \tag{4.65}$$

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

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

In conjunction with (4.65), the above equation implies that

$$\frac{t^2}{2}h^T\nabla^2 f(\mathbf{x})\mathbf{h} + O(t^3) \geq 0$$

Dividing by t^2 and taking limits as $t \to 0$, we get

$$h^T\nabla^2 f(\mathbf{x})\mathbf{h} \geq 0$$

Sufficiency: Suppose that the Hessian matrix is positive semidefinite at each point $\mathbf{x} \in \mathcal{D}$. Consider the same function $\phi(t)$ defined above with $\mathbf{h} = \mathbf{y}-\mathbf{x}$ for $\mathbf{y},\mathbf{x} \in \mathcal{D}$. Applying Taylor's theorem (45) with $n=2$ and $a=0$, we obtain,

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

for some $c \in (0,1)$. Writing this equation in terms of f gives

$$f(\mathbf{x}) = f(\mathbf{y}) + (\mathbf{x}-\mathbf{y})^T\nabla f(\mathbf{y}) + \frac{1}{2}(\mathbf{x}-\mathbf{y})^T\nabla^2 f(\mathbf{z})(\mathbf{x}-\mathbf{y})$$

where $\mathbf{z} = \mathbf{y}+c(\mathbf{x}-\mathbf{y})$. Since $\mathcal{D}$ is convex, $\mathbf{z} \in \mathcal{D}$. Thus, $\nabla^2 f(\mathbf{z}) \succeq 0$. It follows that