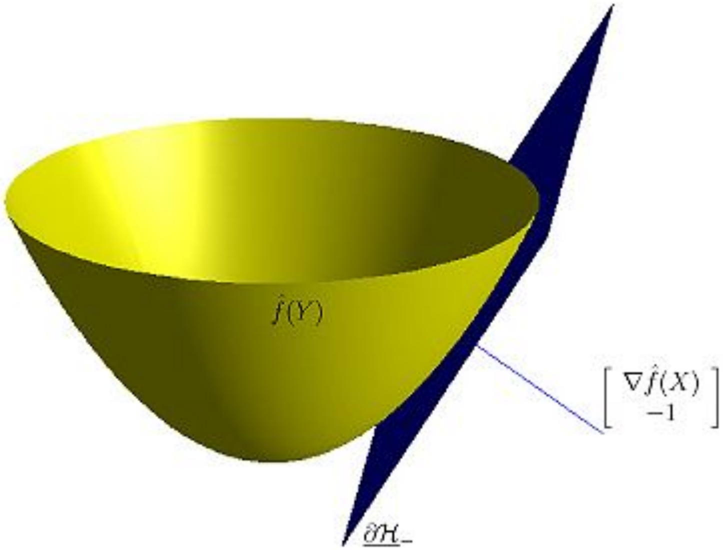
Optimization for ML: Convex Functions



Mrinmay Maharaj

Office: MB 113

mrinmay.mj@rkmvu.ac.in

Convexity in Supervised ML

- **Data**: n observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \ldots, n$, **i.i.d.**
- Prediction as a linear function $\theta^\top \Phi(x)$ of features $\Phi(x) \in \mathbb{R}^d$
- **(regularized) empirical risk minimization**: find $\hat{\theta}$ solution of

$$\min_{\theta \in \mathbb{R}^d} \quad \underbrace{\frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \theta^\top \Phi(x_i))}_{\text{convex data fitting term}} \quad + \quad \underbrace{\mu \Omega(\theta)}_{\text{regularizer}}$$

Convex functions

Loss functions in Supervised ML

$$\min_{\theta \in \mathbb{R}^d} \quad \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \theta^\top \Phi(x_i))$$

- **Regression**: $y \in \mathbb{R}$, prediction $\hat{y} = \theta^\top \Phi(x)$
  - quadratic loss $\frac{1}{2}(y - \hat{y})^2 = \frac{1}{2}(y - \theta^\top \Phi(x))^2$

- **Classification** : $y \in \{-1, 1\}$, prediction $\hat{y} = \text{sign}(\theta^\top \Phi(x))$

  "True" 0-1 loss: $\ell(y\,\theta^\top \Phi(x)) = 1_{y\,\theta^\top \Phi(x) < 0}$

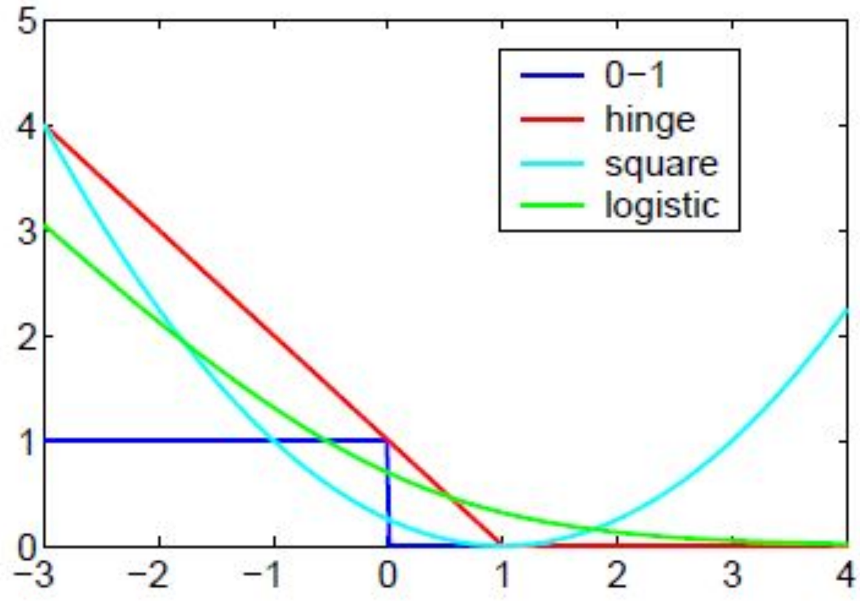
Loss functions in Supervised ML

Support vector machine (hinge loss): non-smooth

$$\ell(Y, \theta^\top \Phi(X)) = \max\{1 - Y\theta^\top \Phi(X), 0\}$$

Least-squares regression

$$\ell(Y, \theta^\top \Phi(X)) = \frac{1}{2}(Y - \theta^\top \Phi(X))^2$$



Logistic regression: smooth

$$\ell(Y, \theta^\top \Phi(X)) = \log(1 + \exp(-Y\theta^\top \Phi(X)))$$

Supervised ML

$\mu\Omega(\theta)$

regularizer

ℓ_1-norm $||\theta||_1 = \sum_{j=1}^{d} |\theta_j|$

(non smooth and convex)

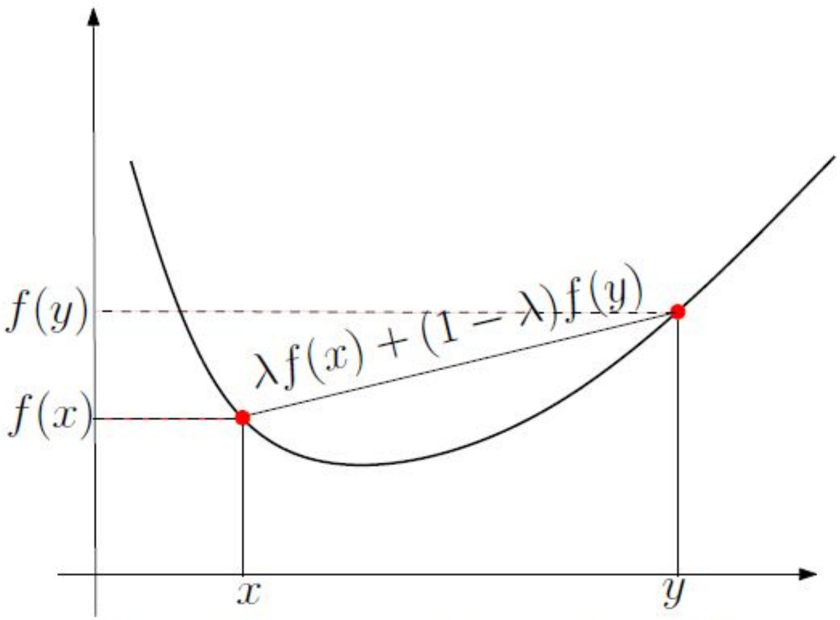
Euclidean norm: $||\theta||_2^2 = \sum_{j=1}^{d} |\theta_j|^2$

(smooth and convex)

Convex functions (general)

Def. A function $f : \mathbb{R}^n \to \mathbb{R}$ is called **convex** if its domain $\mathrm{dom}(f)$ is a convex set and for any $x, y \in \mathrm{dom}(f)$ and $\lambda \geq 0$,

$$f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y).$$



$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$$

(definition does not assume differentiability but difficult to check for all points x and y)

Convex functions: Jensen's inequality

Convex functions (general)

If $f(w) = \|w\|_p$ for a generic norm, then we have

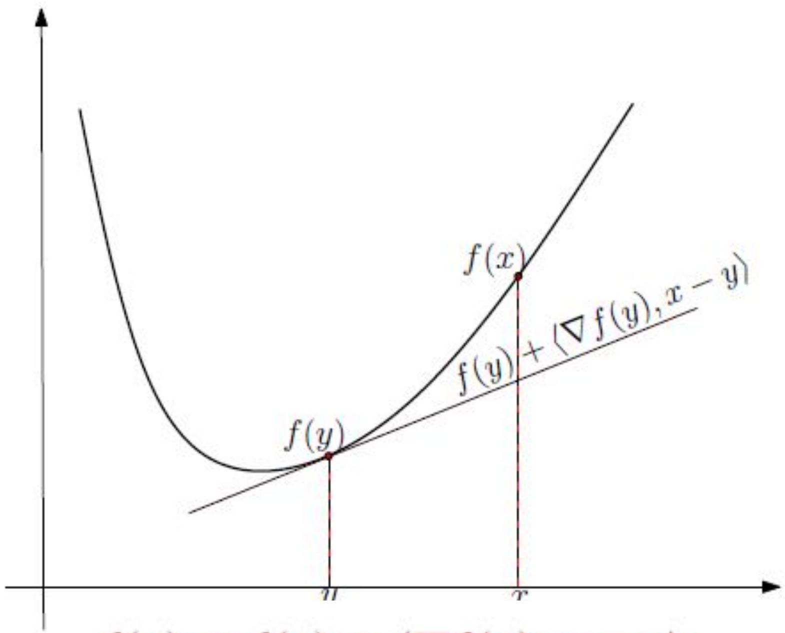
$$\begin{aligned}
f(\theta w + (1-\theta)v) &= \|\theta w + (1-\theta)v\|_p \\
&\le \|\theta w\|_p + \|(1-\theta)v\|_p && \text{(triangle inequality)} \\
&= |\theta| \cdot \|w\|_p + |1-\theta| \cdot \|v\|_p && \text{(absolute homogeneity)} \\
&= \theta\|w\|_p + (1-\theta)\|v\|_p && (0 \le \theta \le 1) \\
&= \theta f(w) + (1-\theta) f(v), && \text{(definition of } f\text{)}
\end{aligned}$$

All squared norms are convex

$$|w|, \|w\|, \|w\|_1, \|w\|^2, \|w_1\|^2, \|w\|_\infty,$$

Convex functions (Differentiable)

If f is differentiable, then f is convex *if and only if* $\operatorname{dom} f$ is convex and $f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle$ for all $x, y \in \operatorname{dom} f$.

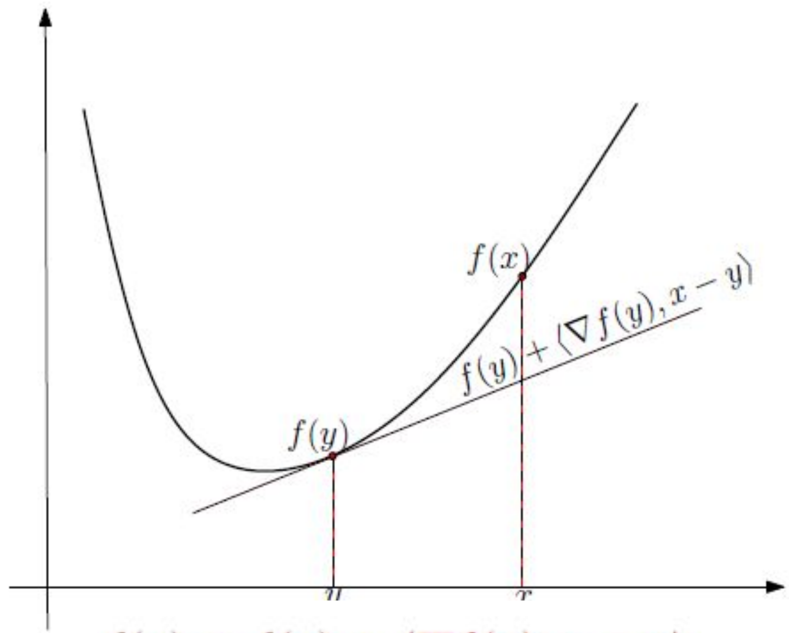


$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle$$

Convex functions: via gradients

Convex functions: Local minima=global minima

If f is differentiable, then f is convex *if and only if* $\operatorname{dom} f$ is convex and $f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle$ for all $x, y \in \operatorname{dom} f$.



$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle$$

Convex functions: via gradients

If ∇f(y) =0, then f(x) >= f(y) for all y , so x is global minimizer

Convex functions: Local minima=global minima

Proposition 1: *Let X be a convex set. If f is convex, then any local minimum of f in X is also a global minimum.*

Proof. Suppose f is convex, and let $\mathbf{x}^*$ be a local minimum of f in $\mathcal{X}$. Then for some neighborhood $N \subseteq \mathcal{X}$ about $\mathbf{x}^*$, we have $f(\mathbf{x}) \geq f(\mathbf{x}^*)$ for all $\mathbf{x} \in N$. Suppose towards a contradiction that there exists $\tilde{\mathbf{x}} \in \mathcal{X}$ such that $f(\tilde{\mathbf{x}}) < f(\mathbf{x}^*)$.

Consider the line segment $\mathbf{x}(t) = t\mathbf{x}^* + (1-t)\tilde{\mathbf{x}},\ t \in [0,1]$, noting that $\mathbf{x}(t) \in \mathcal{X}$ by the convexity of $\mathcal{X}$. Then by the convexity of f,

$$f(\mathbf{x}(t)) \leq tf(\mathbf{x}^*) + (1-t)f(\tilde{\mathbf{x}}) < tf(\mathbf{x}^*) + (1-t)f(\mathbf{x}^*) = f(\mathbf{x}^*)$$

We can pick t to be sufficiently close to 1 that $\mathbf{x}(t) \in N$; then $f(\mathbf{x}(t)) \geq f(\mathbf{x}^*)$ by the definition of N, but $f(\mathbf{x}(t)) < f(\mathbf{x}^*)$ by the above inequality, a contradiction.

It follows that $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$, so $\mathbf{x}^*$ is a global minimum of f in $\mathcal{X}$. $\square$

Subgradient of a function

Sub-gradient :

g is a **subgradient** of f (not necessarily convex) at $\boldsymbol{x}$ if

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

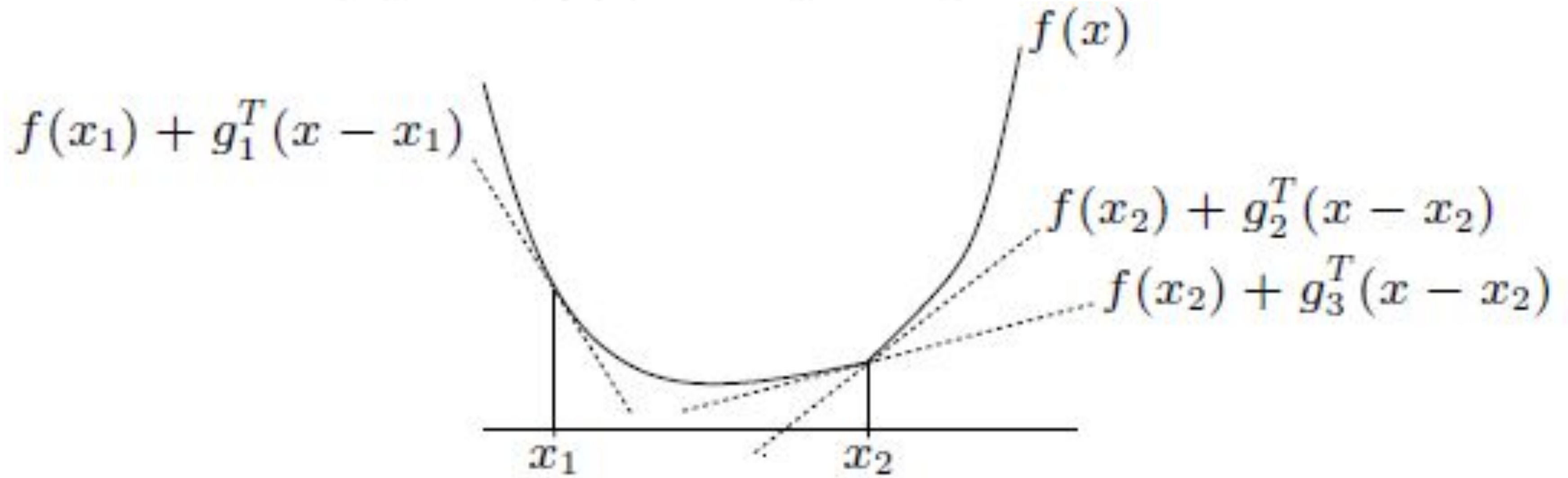


Fig: g_2, g_3 are subgradients at x_2. g_1 is a subgradient at x_1

$\partial f(\boldsymbol{x})$ is the set of all subgradients of f at $\boldsymbol{x}$ (called the subdifferential of f at $\boldsymbol{x}$)

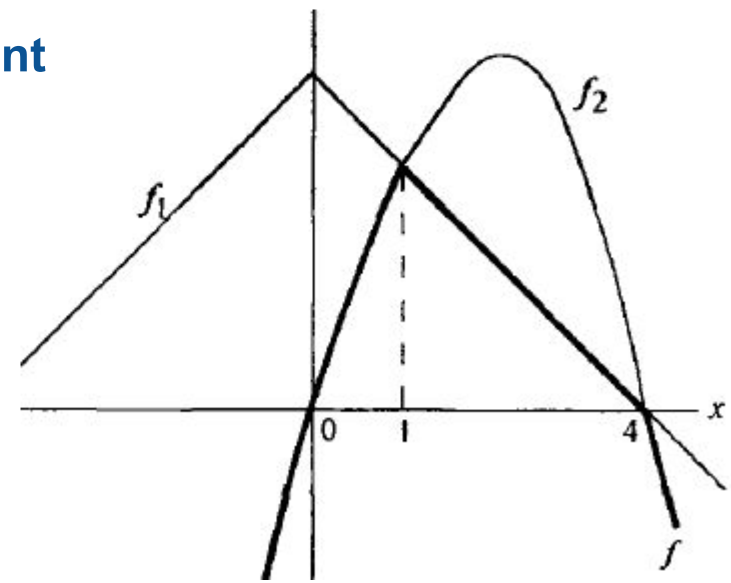
Subgradient of a function

Example 3.2.4 (Bazaraa)

$$f = min\ \{ f_1\ , f_2\ \}$$

$$f_1(x) = 4 - |x|\ ,\ x \in \mathbb{R} \qquad f_2(x) = 4 - (x-2)^2\ ,\ x \in \mathbb{R}$$

Points where subdifferential set has one element



$$x < 0 : g = \{ \nabla f_2 \} = -2x$$

$$x > 4 : g = \{ \nabla f_2 \} = -2x$$

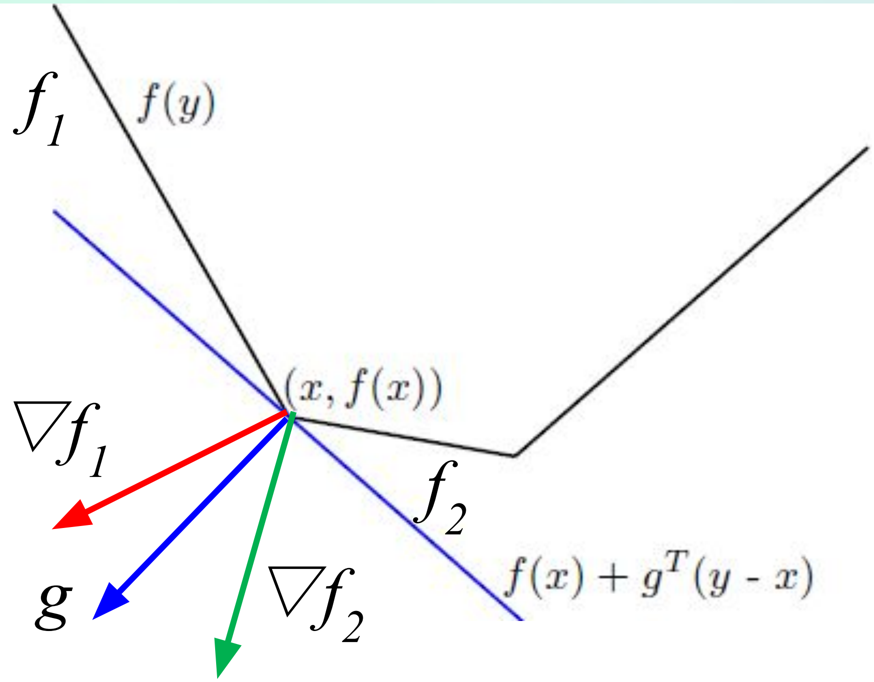
Points where subdifferential is the set of subgradients

$$0 \leq x \leq 4 : g \in \partial f$$

$$x = 4 : g = \lambda \nabla f_1 + (1-\lambda) \nabla f_2 = -4+3\lambda\ ,\ \lambda \in [0,1] \Rightarrow g \in [-1,2]$$

$$x = 1 : g = \lambda \nabla f_1 + (1-\lambda) \nabla f_2 = 2-3\lambda,\quad \lambda \in [0,1] \Rightarrow g \in [-4,-1]$$

Convex functions (non differentiable)



The *subgradient set*, or subdifferential set, $\partial f(x)$ of f at x is

$$\partial f(x) = \{g : f(y) \geq f(x) + g^T(y - x) \text{ for all } y\}.$$

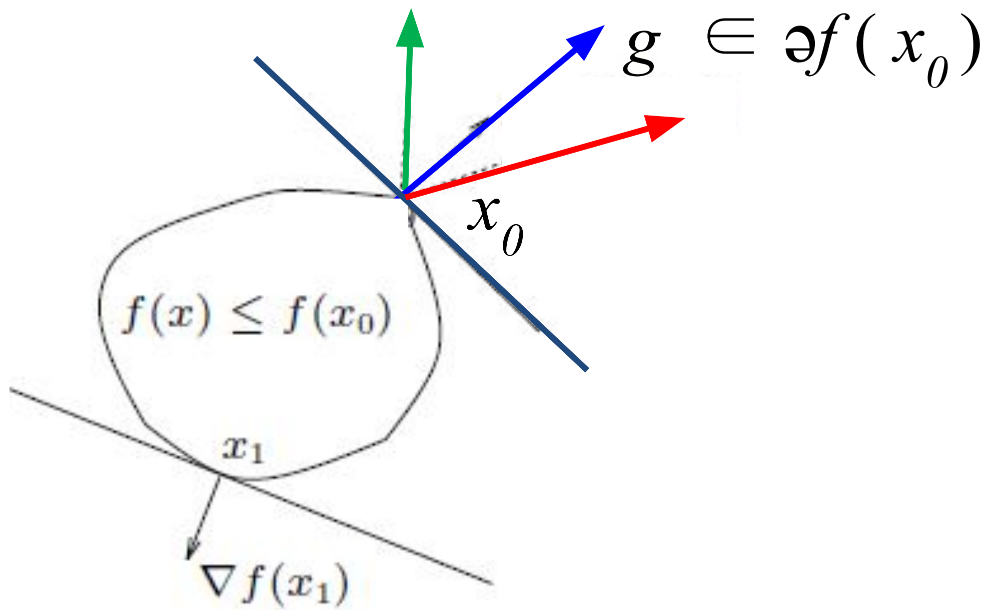
$\partial f(x) = \{\nabla f(x)\}$ if f is differentiable at x

$\partial f(x)$ is a closed convex set **(cone)**

Subgradients and sublevel sets

g is a **subgradient** of f at $\boldsymbol{x_0}$ if $f(x) \geq f(x_0) + g^T(x_0 - x)$

Given x_0 the **sublevel** set is $S = \{x \mid f(x) \leq f(x_0)\}$



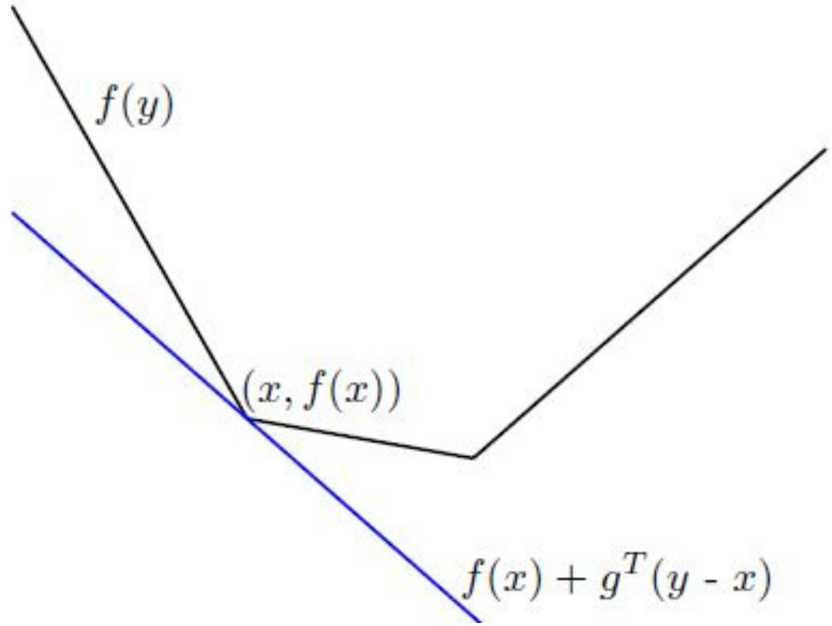
Therefore for $x \in S$, the **sugradient** g satisfies $g^T(x_0 - x) \leq 0$,i.e., g is the normal to the supporting hyperplane at the boundary point x_0

subgradients are normal to supporting hyperplanes of sublevel sets

Convex functions (non differentiable)

Convex functions via sub-gradient or sub-differential

Theorem
$f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if it has non-empty subdifferential set everywhere.



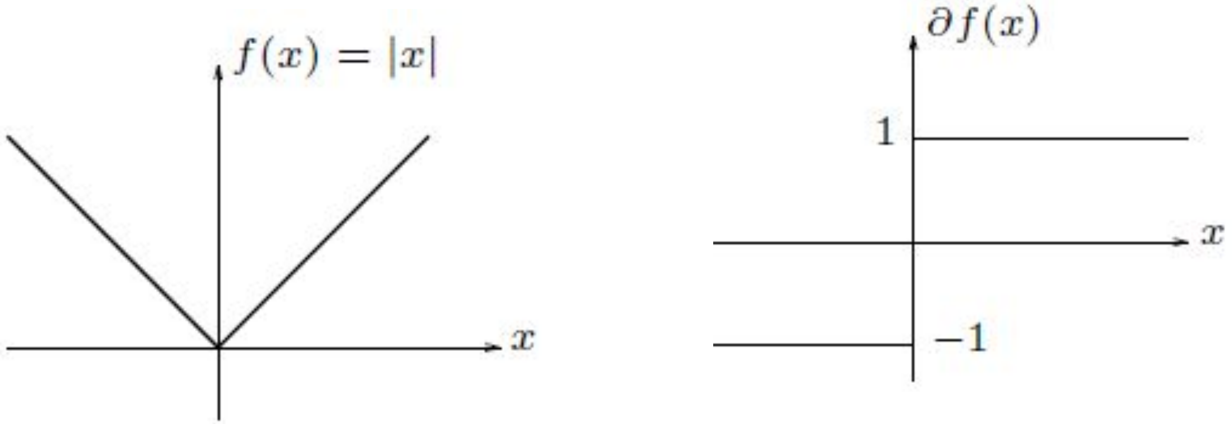
The *subgradient set*, or subdifferential set, $\partial f(x)$ of f at x is

$$\partial f(x) = \{g : f(y) \geq f(x) + g^T(y - x) \text{ for all } y\}.$$

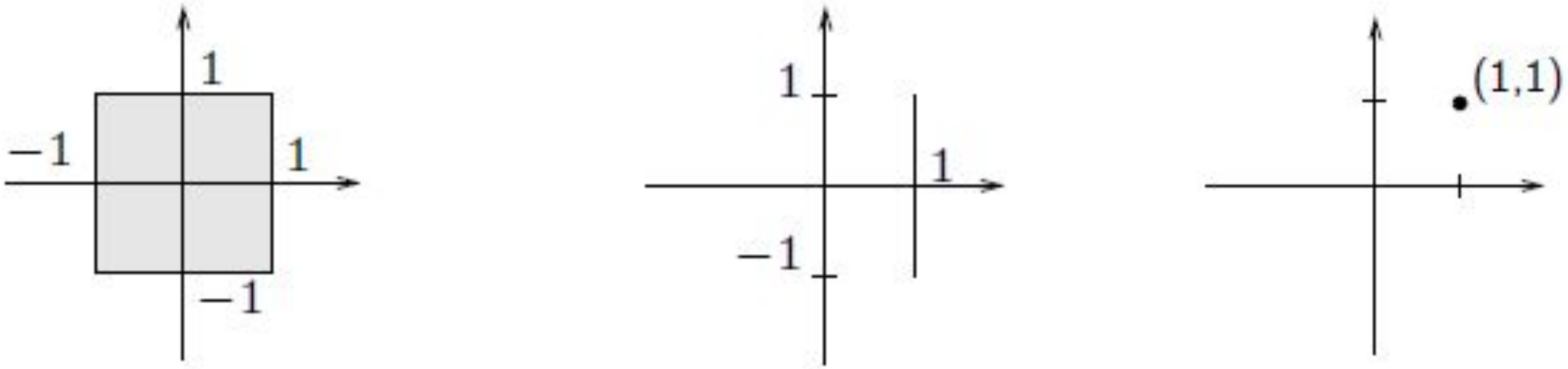
$\partial f(x) = \{\nabla f(x)\}$ if f is differentiable at x

Subdifferential set

Example:



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 $x = (1, 1)$

Convex functions (Twice differentiable)

- A C^2 function is convex iff:

$$\nabla^2 f(w) \succeq 0,$$

for all w in the domain ("curved upwards" in every direction).

- This notation $A \succeq 0$ means that A is positive semidefinite.

- Two equivalent definitions of a positive semidefinite matrix A:
  1. All eigenvalues of A are non-negative.
  2. The quadratic $v^\top A v$ is non-negative for all vectors v.

$$\nabla^2 f(w) = \begin{bmatrix} \frac{\partial}{\partial w_1 \partial w_1} f(w) & \frac{\partial}{\partial w_1 \partial w_2} f(w) & \cdots & \frac{\partial}{\partial w_1 \partial w_d} f(w) \\ \frac{\partial}{\partial w_2 \partial w_1} f(w) & \frac{\partial}{\partial w_2 \partial w_2} f(w) & \cdots & \frac{\partial}{\partial w_2 \partial w_d} f(w) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial}{\partial w_d \partial w_1} f(w) & \frac{\partial}{\partial w_d \partial w_2} f(w) & \cdots & \frac{\partial}{\partial w_d \partial w_d} f(w) \end{bmatrix}$$

Positive Semi-Definite & Positive Definite

The notation $A \succeq 0$ indicates that A is positive semi-definite.

- The eigenvalues of A are all non-negative.
- $v^\top A v \geq 0$ for all vectors v.

The notation $A \succ 0$ indicates that A is positive definite.

- The eigenvalues of A are all positive.
- $v^\top A v > 0$ for all vectors $v \neq 0$.
- This implies that A is invertible (bonus).

> If $A \succ 0$, then all the eigenvalues of A are positive.
>
> If each eigenvalue is positive, the product of the eigenvalues is positive.
>
> The product of the eigenvalues is equal to the determinant.
>
> Thus, the determinant is positive.
>
> The determinant not being 0 implies the matrix is invertible.

The notation $A \succeq B$ indicates that $A - B$ is positive semi-definite.

- The eigenvalues of $A - B$ are all non-negative.
- $v^\top A v \geq v^\top B v$ for all vectors v.

Convexity of least square loss

We can use twice-differentiable condition to show convexity of least squares,

$$f(w) = \frac{1}{2}\|Xw - y\|^2.$$

The Hessian of this objective for any w is given by

$$\nabla^2 f(w) = X^\top X.$$

So we want to show that $X^\top X \succeq 0$ or equivalently that $v^\top X^\top X v \geq 0$ for all v.
We can show this by non-negativity of norms,

$$v^\top X^\top X v = \underbrace{(v^\top X^\top)}_{(Xv)^\top} Xw = \underbrace{(Xv)^\top (Xv)}_{u^\top u} = \underbrace{\|Xv\|^2}_{\|u\|^2} \geq 0,$$

Convexity of logistic loss

We can use twice-differentiable condition to show convexity of binary logistic loss

$$f(w) = \sum_{i=1}^{n} \log(1 + \exp(-y^i w^T x^i)).$$

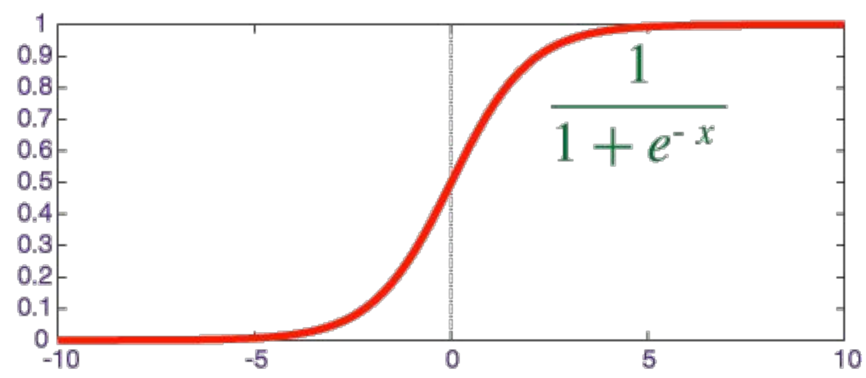
The gradient is

$$\nabla f(w) = X^T r.$$

where the vector r has elements $r_i = -y^i h(-y^i w^T x^i).$

and h is the sigmoid function

$$h(\alpha) = 1/1 + \exp(-\alpha).$$



Convexity of logistic loss

The Hessian is

$$\nabla^2 f(w) = X^T DX.$$

where D is a diagonal matrix with

$$d_{ii} = h(y_i w^T x^i) h(-y^i w^T x^i)$$

Since the sigmoid function h is non-negative, we can compute $D^{\frac{1}{2}}$, and

$$v^T X^T DXv = v^T X^T D^{\frac{1}{2}} D^{\frac{1}{2}} Xv = (D^{\frac{1}{2}} Xv)^T (D^{\frac{1}{2}} Xv) = \|XD^{\frac{1}{2}} v\|^2 \geq 0,$$

so $X^T DX$ is positive semidefinite and logistic regression is convex.

Twice differentiable Convex functions

- Show the following univariate functions are convex
  - Quadratic $w^2 + bw + c$ with $a \geq 0$.
  - Linear: $aw + b$.
  - Constant: b.
  - Exponential: $\exp(aw)$.
  - Negative logarithm: $-\log(w)$.
  - Negative entropy: $w \log w$, for $w > 0$.
  - Logistic loss: $\log(1 + \exp(-w))$.

show $f''(w) \geq 0$ for all w:

- Show the following multivariate functions are convex

$f(W) = -\log \det W$ for $W \succ 0$ (negative log-determinant).
$f(W, v) = v^\top W^{-1} v$ for $W \succ 0$.
$f(w) = \log(\sum_{j=1}^{d} \exp(w_j))$ (log-sum-exp function).

Convexity and minima

We say that a C^2 function is convex if for all w,

$$\nabla^2 f(w) \succeq 0,$$

and this implies any stationary point $(\nabla f(w) = 0)$ is a global minimum.

We say that a C^2 function is strictly convex if for all w,

$$\nabla^2 f(w) \succ 0,$$

and this implies there is at most one stationary point

- Example: $f(x)=x^2$

Strictly convex function

A function is strictly-convex if the convexity definitions hold strictly:

$$f(\theta w + (1-\theta)v) < \theta f(w) + (1-\theta) f(v), \quad 0 < \theta < 1 \qquad (C^0)$$

$$f(v) > f(w) + \nabla f(w)^\top (v - w) \qquad (C^1)$$

$$\nabla^2 f(w) \succ 0 \qquad (C^2)$$

- Function is always strictly below any chord, strictly above any tangent, and curved upwards in every direction.
- Strictly-convex function have at most one global minimum:

Strictly Convex functions: Unique global minima

Proposition 2: *Let X be a convex set. If f is strictly convex, then there exists at most one local minimum of f in X. Consequently, if it exists it is the unique global minimum of f in X.*

Proof. The second sentence follows from the first, so all we must show is that if a local minimum exists in $\mathcal{X}$ then it is unique.

Suppose $\mathbf{x}^*$ is a local minimum of f in $\mathcal{X}$, and suppose towards a contradiction that there exists a local minimum $\tilde{\mathbf{x}} \in \mathcal{X}$ such that $\tilde{\mathbf{x}} \neq \mathbf{x}^*$.

Since f is strictly convex, it is convex, so $\mathbf{x}^*$ and $\tilde{\mathbf{x}}$ are both global minima of f in $\mathcal{X}$ by the previous result. Hence $f(\mathbf{x}^*) = f(\tilde{\mathbf{x}})$. Consider the line segment $\mathbf{x}(t) = t\mathbf{x}^* + (1-t)\tilde{\mathbf{x}},\ t \in [0,1]$, which again must lie entirely in $\mathcal{X}$. By the strict convexity of f,

$$f(\mathbf{x}(t)) < tf(\mathbf{x}^*) + (1-t)f(\tilde{\mathbf{x}}) = tf(\mathbf{x}^*) + (1-t)f(\mathbf{x}^*) = f(\mathbf{x}^*)$$

for all $t \in (0,1)$. But this contradicts the fact that $\mathbf{x}^*$ is a global minimum. Therefore if $\tilde{\mathbf{x}}$ is a local minimum of f in $\mathcal{X}$, then $\tilde{\mathbf{x}} = \mathbf{x}^*$, so $\mathbf{x}^*$ is the unique minimum in $\mathcal{X}$. $\square$

Unique global minima depends on domain

Consider the function $f(x) = x^2$, $x \in X$ (*a strictly convex function*)

- If $X=\mathbb{R}$: The unique global minimum of this function in $\mathbb{R}$ is $x = 0$
- If $X = \{1\}$, which is actually convex, we still have a unique global minimum. But it is not the same as the unconstrained minimum when $X=\mathbb{R}$
- $X = \mathbb{R} \setminus \{0\}$: This set is non-convex, and we can see that f has no minima in X. For any point $x \in X$, one can find another point $y \in X$, such that $f(y) < f(x)$.
- $X =(-\infty,-1] \cup [0, \infty)$: This set is non-convex, and we can see that there is a local minimum ($x = -1$) which is distinct from the global minimum ($x = 0$)
- $X =(-\infty,-1] \cup [1, \infty)$: set is non-convex, and we can see that there are two global minima ($x \pm 1$).

Strongly convex function

- A C^0 function *f(x)* is strongly convex if the function

  $g(x) = f(x) - \frac{\mu}{2}||x||^2$ is a convex function for some $\mu > 0$.

- A C^0 function is strongly convex function if for some $\mu > 0$.

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

- **Intuitively:** strong convexity means that there exists a quadratic lower bound on the growth of the function.
- This implies that a strong convex function is strictly convex since the quadratic lower bound growth is of course strictly greater than the linear growth.

Alternately, strongly convex function is also defined as follows (for some $\mu > 0$)

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) - \frac{\alpha(1-\alpha)\mu}{2}||x - y||^2,\ \alpha \in [0,1].$$

Strongly convex function

Proposition *The following conditions are all equivalent to the condition that a differentiable function f is strongly-convex with constant $\mu > 0$.*

(i) $f(y) \geq f(x) + \nabla f(x)^T(y-x) + \frac{\mu}{2}\|y-x\|^2, \ \forall x, y.$

(ii) $g(x) = f(x) - \frac{\mu}{2}\|x\|^2$ is convex, $\forall x.$

(iii) $(\nabla f(x) - \nabla f(y))^T(x-y) \geq \mu\|x-y\|^2, \ \forall x, y.$

(iv) $f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y) - \frac{\alpha(1-\alpha)\mu}{2}\|x-y\|^2, \ \alpha \in [0,1].$

Proof:

$(i) \equiv (ii)$: It follows from the first-order condition for convexity of $g(x)$, i.e., $g(x)$ is convex if and only if $g(y) \geq g(x) + \nabla g(x)^T(y-x), \ \forall x, y.$

$(ii) \equiv (iii)$: It follows from the monotone gradient condition for convexity of $g(x)$, i.e., $g(x)$ is convex if and only if $(\nabla g(x) - \nabla g(y))^T(x-y) \geq 0, \ \forall x, y.$

$(ii) \equiv (iv)$: It simply follows from the definition of convexity, i.e., $g(x)$ is convex if $g(\alpha x + (1-\alpha)y) \leq \alpha g(x) + (1-\alpha) g(y), \ \forall x, y, \alpha \in [0,1].$

Strongly convex function

We say that a C^2 function is strongly convex if for all w.

$$\nabla^2 f(w) \succeq \mu I, \quad \text{for some } \mu > 0,$$

- μ I is a diagonal matrix and has all eigenvalues equal to μ.
- A $\gtrsim \mu$ I means eigenvalues of A are greater than μ

$$f(w) = \frac{1}{2}\|Xw - y\|^2 + \frac{\lambda}{2}\|w\|^2.$$

In L2-regularized least squares, the Hessian matrix is

$$\nabla^2 f(w) = (X^\top X + \lambda I).$$

$$v^\top \nabla^2 f(w) v = v^\top (X \top X + \lambda I) v = \underbrace{\|Xv\|^2} + v^\top (\lambda I) v \geq v^\top (\lambda I) v,$$

$$\nabla^2 f(w) \succeq \lambda I$$

eigenvalues are greater than λ

Strong convexity ⇒Strict convexity ⇒ Convexity.

<u>Proof:</u> The fact that strict convexity implies convexity is obvious.
To see that strong convexity implies strict convexity, note that strong convexity of f implies

$$f(\lambda x + (1-\lambda)y) - \alpha||\lambda x + (1-\lambda)y||^2 \leq \lambda f(x) + (1-\lambda)f(y) - \lambda\alpha||x||^2 - (1-\lambda)\alpha||y||^2.$$

But

$$\lambda\alpha||x||^2 + (1-\lambda)\alpha||y||^2 - \alpha||\lambda x + (1-\lambda)y||^2 > 0, \ \forall x, y, x \neq y, \ \forall\lambda \in (0,1),$$

because $||x||^2$ is strictly convex (why?). The claim follows.

But the converse is not necessarily true.Observe that f(x) = x is convex but not strictly convex and f(x) = x^4 is strictly convex but not strongly convex

Lipschitz continuity

Bounded gradients of g ($\Leftrightarrow$ Lipschitz-continuity): the function g if convex, differentiable and has (sub)gradients uniformly bounded by B on the ball of center 0 and radius D:

$$\forall \theta \in \mathbb{R}^d, \|\theta\|_2 \leqslant D \Rightarrow \|g'(\theta)\|_2 \leqslant B$$

$$\forall \theta, \theta' \in \mathbb{R}^d, \|\theta\|_2, \|\theta'\|_2 \leqslant D \Rightarrow |g(\theta) - g(\theta')| \leqslant B\|\theta - \theta'\|_2$$

gradients change gradually

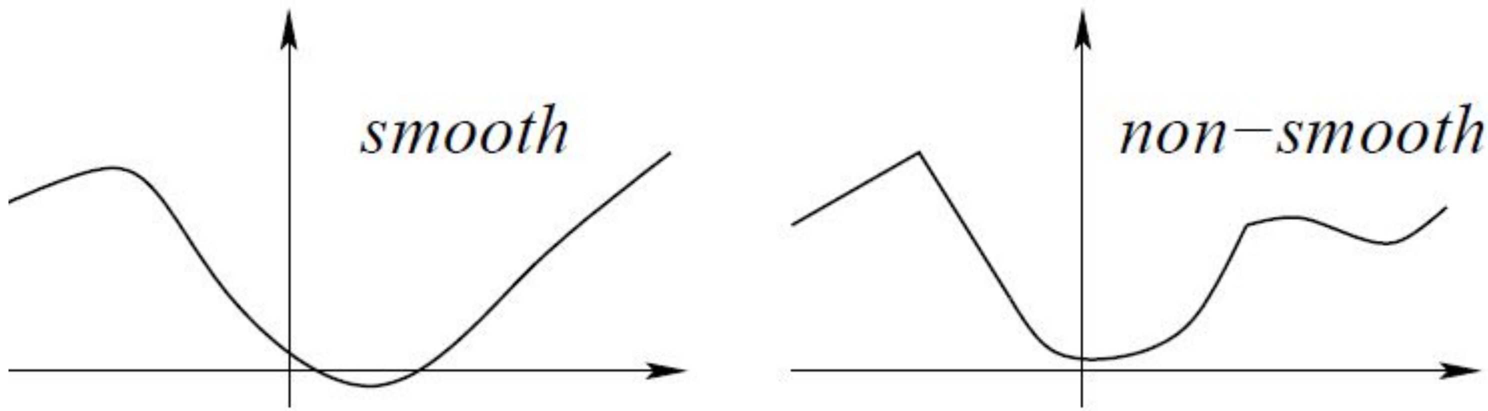
Linear function $f : \mathbb{R}^n \mapsto \mathbb{R}$ defined by $f(\mathbf{w}) = \langle \mathbf{v}, \mathbf{w} \rangle + b$, where $\mathbf{v} \in \mathbb{R}^n$ is $\|\mathbf{v}\|$ –Lipschitz. By using Cauchy-Schwartz inequality, we have

$$|f(\mathbf{w}_1) - f(\mathbf{w}_2)| = |\langle \mathbf{v}, \mathbf{w}_1 - \mathbf{w}_2 \rangle| \leq \|\mathbf{v}\| \, \|\mathbf{w}_1 - \mathbf{w}_2\| \, .$$

Smoothness

A function $g : \mathbb{R}^d \to \mathbb{R}$ is L-smooth if and only if it is differentiable and its gradient is L-Lipschitz-continuous

$$\forall \theta_1, \theta_2 \in \mathbb{R}^d,\ \|g'(\theta_1) - g'(\theta_2)\|_2 \leqslant L\|\theta_1 - \theta_2\|_2$$

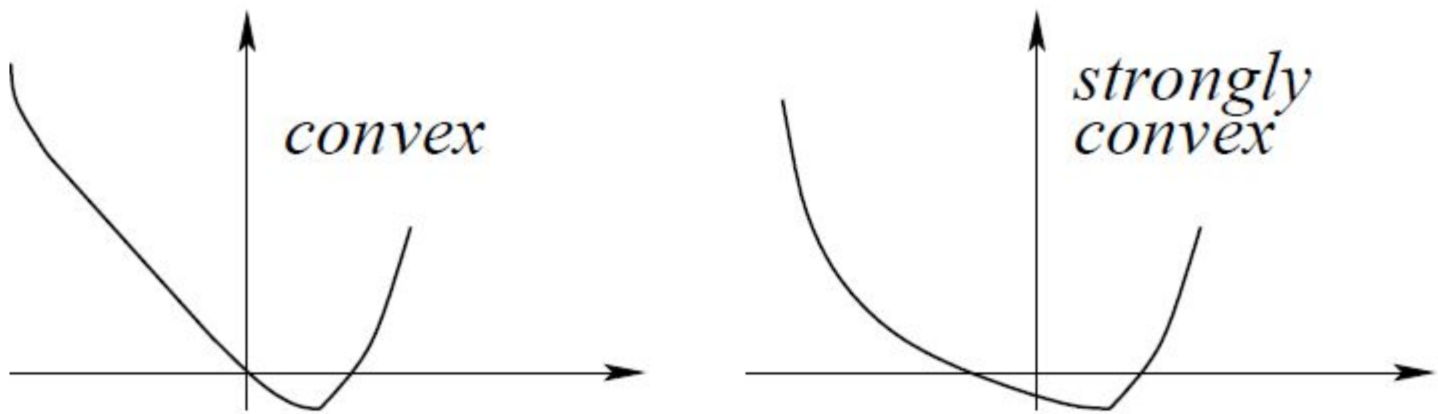


smooth function relies on the change of gradient.

Smoothness and strong convexity

A function $g : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex if and only if

$$\forall \theta_1, \theta_2 \in \mathbb{R}^d,\ g(\theta_1) \geqslant g(\theta_2) + g'(\theta_2)^\top (\theta_1 - \theta_2) + \frac{\mu}{2}\|\theta_1 - \theta_2\|_2^2$$

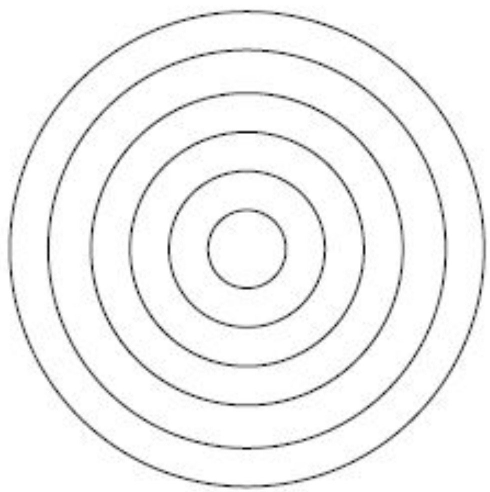


Smoothness and strong convexity

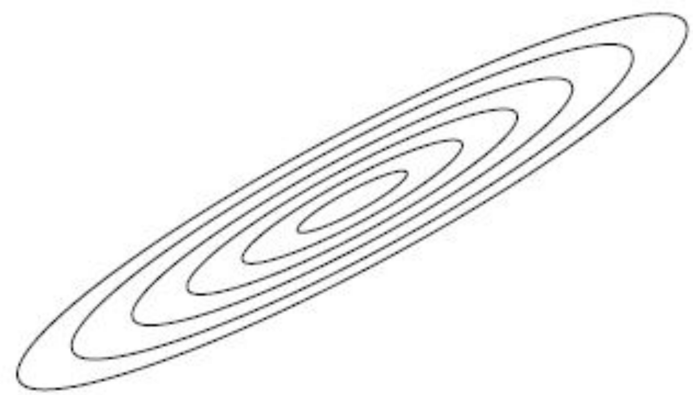
A function $g : \mathbb{R}^d \rightarrow \mathbb{R}$ is μ-strongly convex if and only if

$$\forall \theta_1, \theta_2 \in \mathbb{R}^d, \;\; g(\theta_1) \geqslant g(\theta_2) + g'(\theta_2)^\top (\theta_1 - \theta_2) + \frac{\mu}{2} \|\theta_1 - \theta_2\|_2^2$$

If g is twice differentiable: $\forall \theta \in \mathbb{R}^d, \;\; g''(\theta) \succcurlyeq \mu \cdot \mathrm{Id}$



(large μ/L)



(small μ/L)

Smoothness plus convexity

When a function is both convex and smooth, we have both upper and lower bounds on the difference between the function and its first order approximation.

$$f(\mathbf{v}) \leq f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{v} - \mathbf{w} \rangle + \frac{\beta}{2} \|\mathbf{v} - \mathbf{w}\|^2 .$$

$$f(\mathbf{v}) \geq f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{v} - \mathbf{w} \rangle .$$

Examples of smooth loss function in Machine Learning

For any $\mathbf{x} \in \mathbb{R}^n$ and $y \in \mathbb{R}$, let $f(\mathbf{w}) = (\langle \mathbf{w}, \mathbf{x} \rangle - y)^2$. Then, f is $\left(2 \|\mathbf{x}\|^2\right)$-smooth.

For any $\mathbf{x} \in \mathbb{R}^n$ and $y \in \{\pm 1\}$, let $f(x) = \log(1 + exp(-y \langle \mathbf{w}, \mathbf{x} \rangle))$. Then, f is $\left(\frac{\|\mathbf{x}\|^2}{4}\right)$-smooth.

Summary of smoothness and strong convexity

- **Bounded gradients of g (Lipschitz-continuity)**: the function g if convex, differentiable and has (sub)gradients uniformly bounded by B on the ball of center 0 and radius D:

$$\forall \theta \in \mathbb{R}^d, \|\theta\|_2 \leqslant D \Rightarrow \|g'(\theta)\|_2 \leqslant B$$

- **Smoothness of g**: the function g is convex, differentiable with L-Lipschitz-continuous gradient g' (e.g., bounded Hessians):

$$\forall \theta_1, \theta_2 \in \mathbb{R}^d, \quad \|g'(\theta_1) - g'(\theta_2)\|_2 \leqslant L\|\theta_1 - \theta_2\|_2$$

- **Strong convexity of g**: The function g is strongly convex with respect to the norm $\|\cdot\|$, with convexity constant $\mu > 0$:

$$\forall \theta_1, \theta_2 \in \mathbb{R}^d, \; g(\theta_1) \geqslant g(\theta_2) + g'(\theta_2)^\top(\theta_1 - \theta_2) + \frac{\mu}{2}\|\theta_1 - \theta_2\|_2^2$$

Convexity of sets via convex functions

- For sets of the form

$$\mathcal{C} = \{w \mid g(w) \leq \tau\},$$

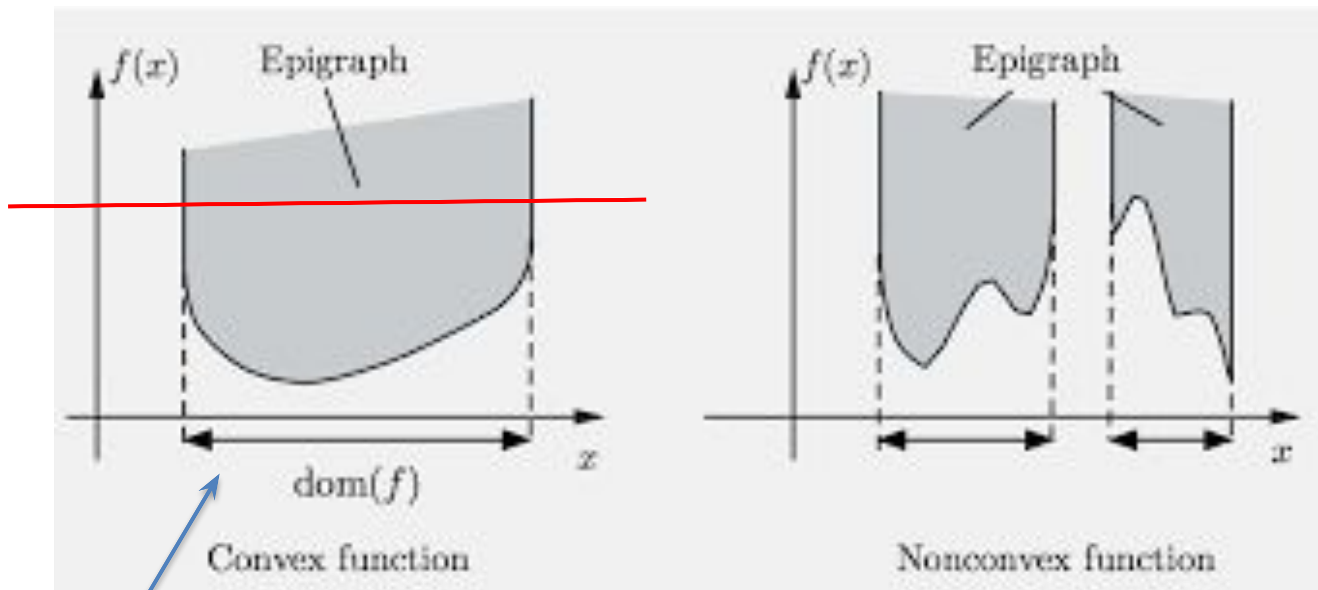
If g is a convex function, then $\mathcal{C}$ is a convex set:

$$g(\underbrace{\theta w + (1-\theta)v}_{\text{convex comb}}) \leq \underbrace{\theta g(w) + (1-\theta)g(v)}_{\text{by convexity}} \leq \underbrace{\theta\tau + (1-\theta)\tau}_{\text{definition of } g} = \tau$$

The set of S={ x | $x^2 \leq 10$ } forms a convex set by convexity of the function $g(x) = x^2$

Convexity of functions via convex sets: Epigraph

Def. A function $f : \mathbb{R}^d \to \mathbb{R}$ is **convex** if and only if its *epigraph* $\{(x, t) \subseteq \mathbb{R}^{d+1} \mid x \in \mathbb{R}^d, t \in \mathbb{R}, f(x) \leq t\}$ is a convex set.



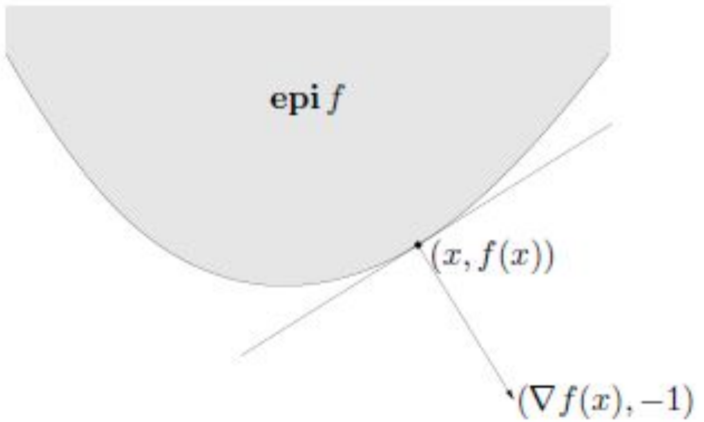
- Sublevel sets, $\{x : f(x) \leq a\}$ are convex for convex f.

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

More on Epigraph

first-order condition for convexity: $f(y) \geq f(x) + \nabla f(x)^T (y - x)$,

If $(y, t) \in \mathbf{epi}\, f$, then $t \geq f(y) \geq f(x) + \nabla f(x)^T (y - x)$.



$$(y, t) \in \mathbf{epi}\, f \implies \begin{bmatrix} \nabla f(x) \\ -1 \end{bmatrix}^T \left(\begin{bmatrix} y \\ t \end{bmatrix} - \begin{bmatrix} x \\ f(x) \end{bmatrix} \right) \leq 0.$$

hyperplane defined by $(\nabla f(x), -1)$ supports $\mathbf{epi}\, f$ at the boundary point $(x, f(x))$

Function is convex iff epigraph is convex

Let $f : S \longrightarrow \mathbb{R}$ be a function defined on the convex subset S of a real linear space L. Then, f is convex on S if and only if its epigraph is a convex subset of $S \times \mathbb{R}$; f is concave if and only if its hypograph is a convex subset of $S \times \mathbb{R}$.

Proof

$f((1-t)\mathbf{x} + t\mathbf{y}) \leqslant (1-t)f(\mathbf{x}) + tf(\mathbf{y})$ for every $\mathbf{x}, \mathbf{y} \in S$ and $t \in [0,1]$.
If $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2) \in \mathrm{epi}(f)$ we have $f(\mathbf{x}_1) \leqslant y_1$ and $f(\mathbf{x}_2) \leqslant y_2$. Therefore,

$$\begin{aligned} f((1-t)\mathbf{x}_1 + t\mathbf{x}_2) &\leqslant (1-t)f(\mathbf{x}_1) + tf(\mathbf{x}_2) \\ &\leqslant (1-t)y_1 + ty_2, \end{aligned}$$

so $((1-t)\mathbf{x}_1 + t\mathbf{x}_2, (1-t)y_1 + ty_2) = (1-t)(\mathbf{x}_1, y_1) + t(\mathbf{x}_2, y_2) \in \mathrm{epi}(f)$

Function is convex iff epigraph is convex

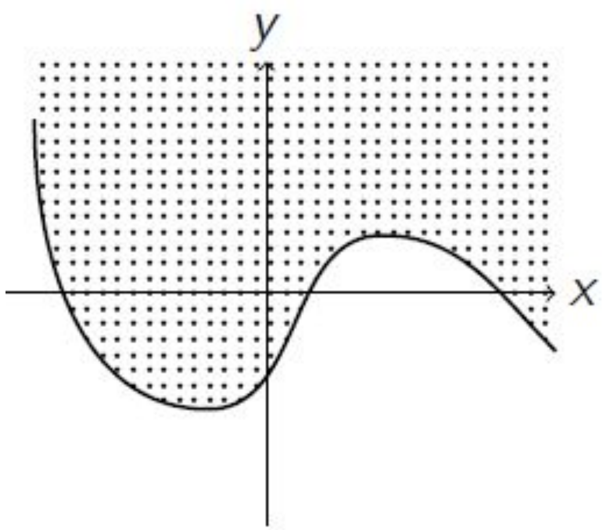
Proof (contd.)

Conversely, suppose that $\text{epi}(f)$ is convex, that is, if $(\mathbf{x}_1, y_1) \in \text{epi}(f)$ and $(\mathbf{x}_2, y_2) \in \text{epi}(f)$, then

$$(1-t)(\mathbf{x}_1, y_1) + t(\mathbf{x}_2, y_2) = ((1-t)\mathbf{x}_1 + t\mathbf{x}_2, (1-t)y_1 + ty_2) \in \text{epi}(f)$$

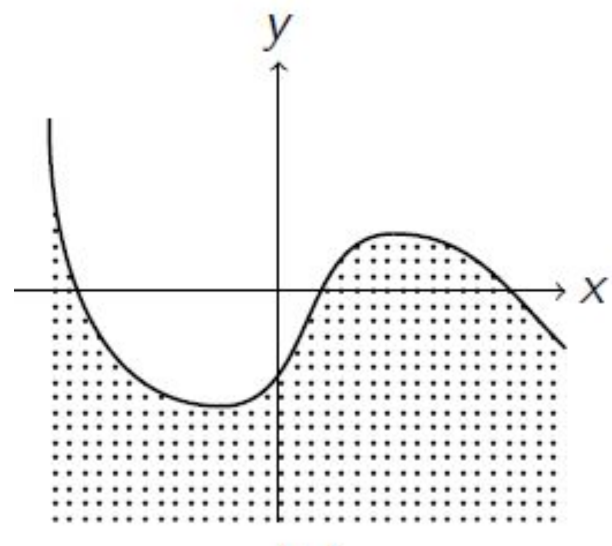
for $t \in [0,1]$. By the definition of the epigraph, this is equivalent to $f(\mathbf{x}_1) \leqslant y_1$, $f(\mathbf{x}_2) \leqslant y_2$ implies $f((1-t)\mathbf{x}_1 + t\mathbf{x}_2) \leqslant (1-t)y_1 + ty_2$. Choosing $y_1 = f(\mathbf{x}_1)$ and $y_2 = f(\mathbf{x}_2)$ yields $f((1-t)\mathbf{x}_1 + t\mathbf{x}_2) \leqslant (1-t)f(\mathbf{x}_1) + tf(\mathbf{x}_2)$, which means that f is convex.

Epigraph, Hypograph and Graph



$$\text{epi}(f) = \{(x, y) \in S \times \mathbb{R} \mid f(x) \leqslant y\}.$$

$$\text{hyp}(f) = \{(x, y) \in S \times \mathbb{R} \mid y \leqslant f(\mathbf{x})\}$$

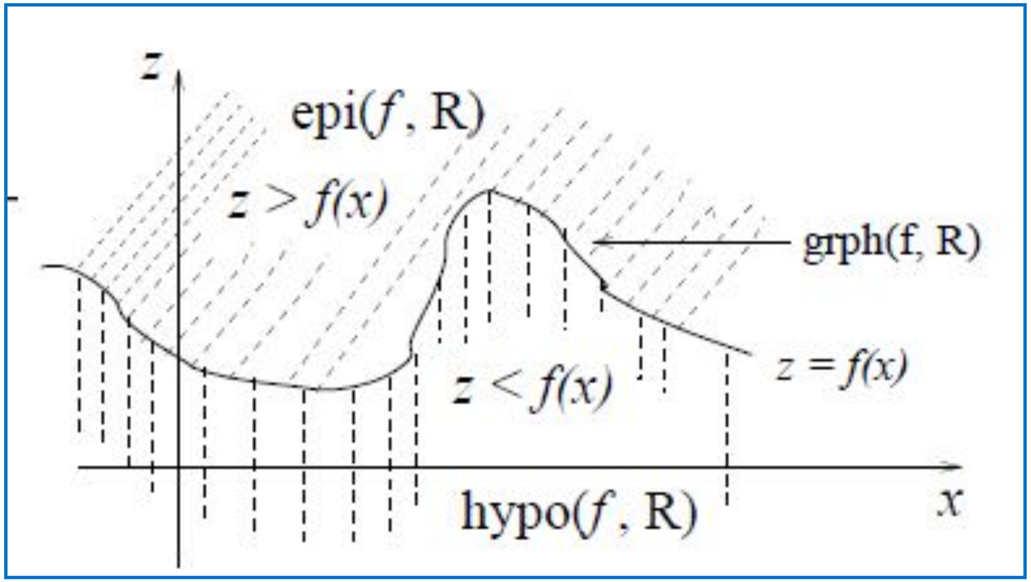


graph of the function f

$$\text{epi}(f) \cap \text{hyp}(f) = \{(x, y) \in S \times \mathbb{R} \mid y = f(x)\}$$

Epigraph, Hypograph and Graph

$$\begin{aligned}
\mathrm{grph}(f, X) &= \{(x, z)\colon\ x \in X, z = f(x)\} \\
\mathrm{epi}(f, X) &= \{(x, z)\colon\ x \in X, z \geq f(x)\} \\
\mathrm{hypo}(f, X) &= \{(x, z)\colon\ x \in X, z \leq f(x)\}
\end{aligned}$$



f is convex on X	$\leftrightarrow$	$\mathrm{epi}(f, X)$	is a convex set
f is concave on X	$\leftrightarrow$	$\mathrm{hypo}(f, X)$	is a convex set
f is affine on X	$\leftrightarrow$	$\mathrm{grph}(f, X)$	is a convex set

Convexity of twice differentiable functions

A quadratic form $Q(\mathbf{x}) = \mathbf{x}^T A \, \mathbf{x}$ and its symmetric matrix A is

- *positive definite* if $Q(\mathbf{x}) > 0$ when $\mathbf{x} \neq \mathbf{0}$
- *positive semidefinite* if $Q(\mathbf{x}) \geq 0$ when $\mathbf{x} \neq \mathbf{0}$
- *negative definite* if $Q(\mathbf{x}) < 0$ when $\mathbf{x} \neq \mathbf{0}$
- *negative semidefinite* if $Q(\mathbf{x}) \leq 0$ when $\mathbf{x} \neq \mathbf{0}$
- *indefinite* if $Q(\mathbf{x})$ takes both positive and negative values

$$\begin{aligned}\mathbf{x}^T A \, \mathbf{x} &= \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 2 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1 + 2x_2 & 2x_1 - x_2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ &= x_1^2 + 2x_2x_1 + 2x_1x_2 - x_2^2 = x_1^2 + 4x_1x_2 - x_2^2\end{aligned}$$

Definiteness and convexity

Let $Q(x_1, x_2, \ldots, x_n) = \mathbf{x}^T A\, \mathbf{x}$ be a quadratic form in n variables, with associated symmetric matrix A. Then we have:

- Q is convex $\Leftrightarrow$ A is positive semidefinite
- Q is concave $\Leftrightarrow$ A is negative semidefinite
- Q is strictly convex $\Leftrightarrow$ A is positive definite
- Q is strictly concave $\Leftrightarrow$ A is negative definite

Eigenvalues and Definiteness

Let $Q(\mathbf{x}) = \mathbf{x}^T A\,\mathbf{x}$ *be a quadratic form, let* A *be its symmetric matrix, and let* $\lambda_1, \lambda_2, \ldots, \lambda_n$ *be the eigenvalues of* A. *Then:*

- Q *is positive definite* $\Leftrightarrow \lambda_1, \lambda_2, \ldots, \lambda_n > 0$
- Q *is positive semidefinite* $\Leftrightarrow \lambda_1, \lambda_2, \ldots, \lambda_n \geq 0$
- Q *is negativ definite* $\Leftrightarrow \lambda_1, \lambda_2, \ldots, \lambda_n < 0$
- Q *is negative semidefinite* $\Leftrightarrow \lambda_1, \lambda_2, \ldots, \lambda_n \leq 0$
- A *is indefinite* $\Leftrightarrow$ *there exists* $\lambda_i > 0$ *and* $\lambda_j < 0$

$$Q(\mathbf{x}) = -x_1^2 + 6x_1x_2 - 9x_2^2 - 2x_3^2 \Rightarrow A = \begin{pmatrix} -1 & 3 & 0 \\ 3 & -9 & 0 \\ 0 & 0 & -2 \end{pmatrix}$$

$$\det(A - \lambda I) = \begin{vmatrix} -1-\lambda & 3 & 0 \\ 3 & -9-\lambda & 0 \\ 0 & 0 & -2-\lambda \end{vmatrix} = (-2-\lambda)(\lambda^2 + 10\lambda) = 0$$

λ=-2,-10,0

Principal Minor & Leading Principal Minor

Let A be an n×n matrix. A k×k submatrix of A formed by deleting n − k rows of A, and the same n − k columns of A, is called **principal submatrix** of A. The determinant of a principal submatrix of A is called a **principal minor** of A.

(**Notation**: Δ_k is the principal minor of order k.

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}, \textit{ formed by deleting column 3 and row 3;}$$

$$\begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix}, \textit{ formed by deleting column 2 and row 2;}$$

$$\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix}, \textit{ formed by deleting column 1 and row 1}$$

The kth order principal submatrix of A obtained by deleting the last n − k rows and columns of A is called the kth order **leading principal submatrix** of A, and its determinant is called the k^{th} order **leading principal minor** of A.

Notation: D_k is leading principal minor of order k

Example of Principal Minor

Let **A** =(a,b ; b,c) be a symmetric 2 x 2 matrix.
The <u>leading principal minors</u> are

$$D_1 = a \text{ and } D_2 = ac - b^2.$$

The principal minors are

$$\Delta_1 = a \text{ and } \Delta_1 = c \text{ (of order one) and}$$
$$\Delta_2 = ac - b^2 \text{ (of order two).}$$

So if $a>0$ and $ac - b^2 > 0$ then A is positive definite.

Note:

If $D_1 = a>0$ and $D_2 = ac - b^2 >0$, then $c >0$, since $ac > b^2 \geq 0$.
The characteristic equation of **A** is $\lambda^2 - (a + c)\lambda + (ac - b^2) = 0$
Solution is

$$\lambda = \frac{a+c}{2} \pm \frac{\sqrt{(a+c)^2 - 4(ac - b^2)}}{2}$$

and both solutions are positive, so A is positive definite.

Principal Minor Test

Let A be a symmetric $n \times n$ matrix. Then we have:

- *A is positive definite $\Leftrightarrow D_k > 0$ for all leading principal minors*
- *A is negative definite $\Leftrightarrow (-1)^k D_k > 0$ for all leading principal minors*
- *A is positive semidefinite $\Leftrightarrow \Delta_k \geq 0$ for all principal minors*
- *A is negative semidefinite $\Leftrightarrow (-1)^k \Delta_k \geq 0$ for all principal minors*

$$A = \begin{pmatrix} 1 & 4 & 6 \\ 4 & 2 & 1 \\ 6 & 1 & 6 \end{pmatrix} \Rightarrow \quad D_1 = 1, \quad D_2 = \begin{vmatrix} 1 & 4 \\ 4 & 2 \end{vmatrix} = -14, \quad D_3 = \begin{vmatrix} 1 & 4 & 6 \\ 4 & 2 & 1 \\ 6 & 1 & 6 \end{vmatrix} = -109$$

- *Positive definite: $D_1 > 0, D_2 > 0, D_3 > 0$*
- *Negative definite: $D_1 < 0, D_2 > 0, D_3 < 0$*
- *Positive semidefinite: $\Delta_1 \geq 0, \Delta_2 \geq 0, \Delta_3 \geq 0$ for all principal minors*
- *Negative semidefinite: $\Delta_1 \leq 0, \Delta_2 \geq 0, \Delta_3 \leq 0$ for all principal minors*

does not fit with any of these criteria.

Principal Minor Test

- A practical test for positive definiteness that does not require explicit calculation of the eigenvalues is the **principal minor test**.
- A necessary and sufficient condition that a symmetric n x n matrix be positive definite is that all n leading principal minors D_k are positive
- One particular failure of this algorithm occurs when some leading principal minor is zero, but the others fit one of the patterns above. In this case, the matrix is not definite, but may or may not be semidefinite. In this case, we must unfortunately check <u>not only the leading principal minors, but every principal minor</u>

*Source: The Principal Minor Test for Semidefinite Matrices, by John E. Prussing**

Principal Minor Test

- "A matrix will be **positive semidefinite** if all n leading principal minors are **nonnegative** " is not always true,

$$A = \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix}$$

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & a \end{bmatrix}$$

Both leading principal minors are zero and hence nonnegative, but the matrix is obviously not positive semidefinite

The leading principal minors are nonnegative ($D_1 = 1$,,$D_2 = D_3 = 0$), but the matrix is not positive semidefinite (the quadratic form is $Q(x)=(x_1+x_2+x_3)^2+(a-1)x_3^2 \geq 0$, if $a \geq 0$ when $\mathbf{x} | x_1+x_2+x_3=0$

*Source: The Principal Minor Test for Semidefinite Matrices, by John E. Prussing**

Principal Minor Test

- Thus, the condition that $D_k \geq 0$ (leading principal minor) is apparently a necessary but not a sufficient condition for positive semidefiniteness.

- The **correct** necessary and sufficient condition is that all possible principal minors are nonnegative ($\Delta_k \geq 0$.)

- For an n x n matrix, The total number of principal minors is 2^n -1

*Source: The Principal Minor Test for Semidefinite Matrices, by John E. Prussing**

Principal Minor Test

$$A = \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix}$$

If we calculate principal minors D_k formed by deleting the first rather than last n - k rows and columns, we find that Δ_1 = -1 and Δ_2=0, which clearly violates the condition.

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & a \end{bmatrix}$$

In the same way, the principal minors are
Δ_1 = a, Δ_2 = a -1, and Δ_3 = 0, satisfying the condition only if a > 1.

*Source: The Principal Minor Test for Semidefinite Matrices, by John E. Prussing**

Characterizing Convexity functions

- A C^0 function is convex if the area above the function is a convex set **(epigraph)**
- A C^0 function is convex if the function is always below its chords between points.

$$\underbrace{f(\theta w + (1-\theta)v)}_{\text{convex comb}} \le \underbrace{\theta f(w) + (1-\theta) f(v)}_{\text{“chord”}}$$

- A C^1 function is convex if the function is always above its tangent planes.
- A C^2 function is convex if it is curved upwards everywhere **(Hessian is positive semi definite)**

Examples of Convex functions

All of $\mathbb{R}^n$

Non-negative orthant, $\mathbb{R}^n_+$:

let $x \succeq 0$, $y \succeq 0$, $\alpha x + (1-\alpha)y \succeq 0$.

Norm balls $\|x\| \leq 1$, $\|y\| \leq 1$

$$\|\alpha x + (1-\alpha)y\| \leq \|\alpha x\| + \|(1-\alpha)y\| = \alpha\|x\| + (1-\alpha)\|y\| \leq 1$$

Affine subspaces: $Ax = b$, $Ay = b$,

$$A(\alpha x + (1-\alpha)y) = \alpha Ax + (1-\alpha)Ay = \alpha b + (1-\alpha)b = b.$$

positive semidefinite cone $\mathbb{S}^n_+ \subset \mathbb{R}^{n\times n}$

$A \in \mathbb{S}^n_+$ means $x^T Ax \geq 0$ for all $x \in \mathbb{R}^n$

$$A, B \in \mathbb{S}_n^+,$$

$$x^T(\alpha A + (1-\alpha)B)x$$
$$= \alpha x^T Ax + (1-\alpha)x^T Bx \geq 0.$$

Examples of Convex functions

Exponential. e^{ax} is convex on $\mathbf{R}$, for any $a \in \mathbf{R}$.

Powers. x^a is convex on $\mathbf{R}_{++}$ when $a \geq 1$ or $a \leq 0$, and concave for $0 \leq a \leq 1$.

Powers of absolute value. $|x|^p$, for $p \geq 1$, is convex on $\mathbf{R}$.

Logarithm. $\log x$ is concave on $\mathbf{R}_{++}$.

Norms. Every norm on $\mathbf{R}^n$ is convex.

Max function. $f(x) = \max\{x_1, \ldots, x_n\}$ is convex on $\mathbf{R}^n$.

Quadratic-over-linear function. The function $f(x, y) = x^2/y$, with

$$\mathbf{dom}\, f = \mathbf{R} \times \mathbf{R}_{++} = \{(x, y) \in \mathbf{R}^2 \mid y > 0\}, \quad \text{is convex}$$

Log-sum-exp. The function $f(x) = \log\left(e^{x_1} + \cdots + e^{x_n}\right)$ is convex on $\mathbf{R}^n$.

Geometric mean. The geometric mean $f(x) = \left(\prod_{i=1}^{n} x_i\right)^{1/n}$ is concave on $\mathbf{dom}\, f = \mathbf{R}^n_{++}$.

Log-determinant. The function $f(X) = \log\det X$ is concave on $\mathbf{dom}\, f = \mathbf{S}^n_{++}$.

Source: Convex Optimization, by Stephen Boyd

Examples of Convex functions

Max function. The function $f(x) = \max_i x_i$ satisfies, for $0 \le \theta \le 1$,

$$\begin{aligned} f(\theta x + (1-\theta)y) &= \max_i(\theta x_i + (1-\theta)y_i) \\ &\le \theta \max_i x_i + (1-\theta)\max_i y_i \\ &= \theta f(x) + (1-\theta)f(y). \end{aligned}$$

Quadratic-over-linear function. To show that the quadratic-over-linear function $f(x, y) = x^2/y$ is convex, we note that (for $y > 0$),

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

Geometric mean. In a similar way we can show that the geometric mean $f(x) = \left(\prod_{i=1}^n x_i\right)^{1/n}$ is concave on $\mathbf{dom}\, f = \mathbf{R}^n_{++}$. Its Hessian $\nabla^2 f(x)$ is given by

$$\frac{\partial^2 f(x)}{\partial x_k^2} = -(n-1)\frac{\left(\prod_{i=1}^n x_i\right)^{1/n}}{n^2 x_k^2}, \qquad \frac{\partial^2 f(x)}{\partial x_k \partial x_l} = \frac{\left(\prod_{i=1}^n x_i\right)^{1/n}}{n^2 x_k x_l} \quad \text{for } k \neq l,$$

$$v^T \nabla^2 f(x) v = -\frac{\prod_{i=1}^n x_i^{1/n}}{n^2}\left(n\sum_{i=1}^n v_i^2/x_i^2 - \left(\sum_{i=1}^n v_i/x_i\right)^2\right) \le 0$$

Restriction of a convex function to a line

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

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

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

Example: log determinant

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

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

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

(concave in *t*)

Convexity preserving operations

- Nonnegative weighted sum: $f(x) = \sum w_i f_i(x)$ is concave(convex) if $f_i(x)$ are concave(convex)

  Note: $\mathbf{epi}\,(wf) = \begin{bmatrix} I & 0 \\ 0 & w \end{bmatrix} \mathbf{epi}\,(f)$ is convex if $w_i \geq 0$ and f convex, because the image of the convex set $\mathbf{epi}\,(f)$ under the linear map $\mathrm{T}(\mathbf{y}) = \begin{bmatrix} I & 0 \\ 0 & w \end{bmatrix} \mathbf{y}$ is convex.

- Composition with an affine mapping: $f\colon \mathbb{R}^n \to \mathbb{R}, \mathbf{A} \in \mathbb{R}^{n \times m}, \mathbf{b} \in \mathbb{R}^m$. Let $g\colon \mathbb{R}^m \to \mathbb{R}, g(x) = f(Ax + b)$. If f is convex (concave), g is convex (concave)

Convexity preserving operation: Non-negative Sum

If f and g are convex, then $f + g$ is convex. Furthermore, if g is strictly convex, then $f + g$ is strictly convex, and if g is m-strongly convex, then $f + g$ is also m-strongly convex.

Proof. Suppose f and g are convex. Then for all $\mathrm{x}, \mathrm{y} \in \operatorname{dom}(f+g) = \operatorname{dom} f \cap \operatorname{dom} g$,

$$\begin{aligned}
(f+g)(t\mathrm{x}+(1-t)\mathrm{y}) &= f(t\mathrm{x}+(1-t)\mathrm{y}) + g(t\mathrm{x}+(1-t)\mathrm{y}) \\
&\le tf(\mathrm{x}) + (1-t)f(\mathrm{y}) + g(t\mathrm{x}+(1-t)\mathrm{y}) && \text{convexity of } f \\
&\le tf(\mathrm{x}) + (1-t)f(\mathrm{y}) + tg(\mathrm{x}) + (1-t)g(\mathrm{y}) && \text{convexity of } g \\
&= t(f(\mathrm{x}) + g(\mathrm{x})) + (1-t)(f(\mathrm{y}) + g(\mathrm{y})) \\
&= t(f+g)(\mathrm{x}) + (1-t)(f+g)(\mathrm{y})
\end{aligned}$$

If g is strictly convex, the second inequality above holds strictly for $\mathrm{x} \neq \mathrm{y}$ and $t \in (0,1)$, so $f+g$ is strictly convex.

If g is m-strongly convex, then the function $h(\mathrm{x}) \equiv g(\mathrm{x}) - \frac{m}{2}\|\mathrm{x}\|_2^2$ is convex, so $f+h$ is convex. But

$$(f+h)(\mathrm{x}) \equiv f(\mathrm{x}) + h(\mathrm{x}) \equiv f(\mathrm{x}) + g(\mathrm{x}) - \frac{m}{2}\|\mathrm{x}\|_2^2 \equiv (f+g)(\mathrm{x}) - \frac{m}{2}\|\mathrm{x}\|_2^2$$

so $f+g$ is m-strongly convex. $\square$

Convexity preserving operation: Affine composition

If f is convex, then $g(\mathbf{x}) \equiv f(\mathbf{A}\mathbf{x}+\mathbf{b})$ *is convex*

Proof. Suppose f is convex and g is defined like so. Then for all $\mathbf{x}, \mathbf{y} \in \operatorname{dom} g$,

$$\begin{aligned} g(t\mathbf{x}+(1-t)\mathbf{y}) &= f(\mathbf{A}(t\mathbf{x}+(1-t)\mathbf{y})+\mathbf{b}) \\ &= f(t\mathbf{A}\mathbf{x}+(1-t)\mathbf{A}\mathbf{y}+\mathbf{b}) \\ &= f(t\mathbf{A}\mathbf{x}+(1-t)\mathbf{A}\mathbf{y}+t\mathbf{b}+(1-t)\mathbf{b}) \\ &= f(t(\mathbf{A}\mathbf{x}+\mathbf{b})+(1-t)(\mathbf{A}\mathbf{y}+\mathbf{b})) \\ &\le tf(\mathbf{A}\mathbf{x}+\mathbf{b})+(1-t)f(\mathbf{A}\mathbf{y}+\mathbf{b}) \\ &= tg(\mathbf{x})+(1-t)g(\mathbf{y}) \end{aligned}$$

Convexity preserving operations: Pointwise Maximum

If f and g are convex, then $h(\mathrm{x}) \equiv \max\{f(\mathrm{x}), g(\mathrm{x})\}$ is convex.

Proof. Suppose f and g are convex and h is defined like so. Then for all $\mathrm{x}, \mathrm{y} \in \operatorname{dom} h$,

$$
\begin{aligned}
h(t\mathrm{x} + (1-t)\mathrm{y}) &= \max\{f(t\mathrm{x} + (1-t)\mathrm{y}), g(t\mathrm{x} + (1-t)\mathrm{y})\} \\
&\leq \max\{tf(\mathrm{x}) + (1-t)f(\mathrm{y}), tg(\mathrm{x}) + (1-t)g(\mathrm{y})\} \\
&\leq \max\{tf(\mathrm{x}), tg(\mathrm{x})\} + \max\{(1-t)f(\mathrm{y}), (1-t)g(\mathrm{y})\} \\
&= t\max\{f(\mathrm{x}), g(\mathrm{x})\} + (1-t)\max\{f(\mathrm{y}), g(\mathrm{y})\} \\
&= th(\mathrm{x}) + (1-t)h(\mathrm{y})
\end{aligned}
$$

in the first inequality we have used convexity of f and g plus the fact that
$a \leq c$; $b \leq d$ implies $\max\{a,b\} \leq \max\{c.d\}$
and in the second inequality we have used the fact that
$\max\{a+b\ ;\ c+d\} \leq \max\{a,c\} + \max\{b,d\}$

Convexity preserving operations: Pointwise Supremum

- Pointwise supremum: If for each $y \in A, f(x, y)$ is convex, then $g(x) = \sup_{y \in A} f(x, y)$ is convex
- The pointwise supremum of functions corresponds to the intersection of epigraphs

$$\mathbf{epi}\, g = \bigcap_{y \in A} \mathbf{epi}\, f(\cdot, y).$$

- E.g: Operator Norm of matrix

$$\|A\|_2 := \sup_{\|x\|_2 \neq 0} \frac{\|Ax\|_2}{\|x\|_2}. = \max_{x:\|x\|_2 \leq 1} \|Ax\|_2.$$

$\|A\|_2 = \sigma_{\max}(A)$, where $\sigma_{\max}$ is the largest singular value of A.

for convex f, $f(Ax)$ is also convex. Thus, $\|Ax\|_2$ is convex.

pointwise max of convex functions is convex.

Composition of Convex Functions

- Scalar Composition

$g : \mathbf{R}^n \to \mathbf{R}$ and $h : \mathbf{R} \to \mathbf{R}$: $\quad f(x) = h(g(x))$

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

for n=1 (**dom** f = **dom** g = $\mathbb{R}$)

- $f(x)$ is convex if $h(x)$ is convex ($h'' \geq 0$) and non-decreasing ($h' \geq 0$) and $g(x)$ is convex ($g'' \geq 0$)
- $f(x)$ is convex if $h(x)$ is convex ($h'' \geq 0$) and non-increasing ($h' \leq 0$) and $g(x)$ is concave ($g'' \leq 0$)
- $f(x)$ is concave if $h(x)$ is concave ($h'' \leq 0$) and non-decreasing ($h' \geq 0$) and $g(x)$ is concave ($g'' \leq 0$)
- $f(x)$ is concave if $h(x)$ is concave ($h'' \leq 0$) and non-increasing ($h' \leq 0$) and $g(x)$ is convex ($g'' \geq 0$)

for n>1 In place of $h(x)$ we use the extended value function $\tilde{h}(x)$, defined as $\tilde{h}(x) = \infty, x \notin \mathbf{dom}(h), \tilde{h}(x) = h(x), x \in \mathbf{dom}(h)$

Composition of Convex Functions

- Examples of Scalar Composition

$$g : \mathbf{R}^n \to \mathbf{R} \text{ and } h : \mathbf{R} \to \mathbf{R}: \quad f(x) = h(g(x))$$

- If $g(x)$ is convex then $exp(g(x))$ is convex
- If $g(x)$ is concave and positive then $\log(g(x))$ is concave
- If $g(x)$ is concave and positive then $1/g(x)$ is convex

Composition of Convex Functions

- Vector Composition

$$g : \mathbf{R}^n \to \mathbf{R}^k \text{ and } h : \mathbf{R}^k \to \mathbf{R}\text{: } f(x) = h(g(x)) = h(g_1(x), g_2(x), \ldots, g_k(x))$$

for k = 1 and f , g differentiable

$$f''(x) = g'(x)^T \nabla^2 h(g(x)) g'(x) + \nabla h(g(x))^T g''(x),$$

- $f(x)$ is convex if $h(x)$ is convex ($\nabla^2 h$ is PSD) and non-decreasing in each argument ($h'_i \geq 0$) and $g_i(x)$ are convex ($g'_i(x) \geq 0$)
- $f(x)$ is convex if $h(x)$ is convex ($\nabla^2 h$ is PSD) and non-increasing in each argument ($h'_i \leq 0$) and $g_i(x)$ are concave ($g'_i(x) \leq 0$)
- $f(x)$ is concave if $h(x)$ is concave ($\nabla^2 h$ is NSD) and non-decreasing in each argument ($h'_i \geq 0$) and $g_i(x)$ are concave ($g'_i(x) \leq 0$)

for k >1

In place of $h(x)$ we use the extended value function $\tilde{h}(x)$, defined as $\tilde{h}(x) = \infty, x \notin \mathbf{dom}(h), \tilde{h}(x) = h(x), x \in \mathbf{dom}(h)$

Composition of Convex Functions

- Examples of Vector Composition

$g : \mathbf{R}^n \to \mathbf{R}^k$ and $h : \mathbf{R}^k \to \mathbf{R}$: $f(x) = h(g(x)) = h(g_1(x), g_2(x), \ldots, g_k(x))$

- $\sum_{i=1}^{m} \log g_i(x)$ 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

Log concavity

Log-concave and log-convex functions

- A function $f(x)$ is logarithmically concave (convex) or log-concave if $f(x) > 0$, for all $x \in dom(f)$ and $\log f(x)$ is concave (convex).
- f is log-convex if and only if $1/f$ is logconcave.

Examples of Log-Convex functions

Examples of Log-concave/convex functions

- *Affine function.* $f(x) = a^T x + b$ is log-concave on $\{x \mid a^T x + b > 0\}$.
- *Powers.* $f(x) = x^a$, on $\mathbf{R}_{++}$, is log-convex for $a \leq 0$, and log-concave for $a \geq 0$.
- *Exponentials.* $f(x) = e^{ax}$ is log-convex and log-concave.
- The cumulative distribution function of a Gaussian density,

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

  is log-concave
- *Gamma function.* The Gamma function,

$$\Gamma(x) = \int_{0}^{\infty} u^{x-1} e^{-u}\, du,$$

  is log-convex for $x \geq 1$
- *Determinant.* $\det X$ is log concave on $\mathbf{S}^n_{++}$.

Closure properties of Log concave functions

- Log-concavity is closed under multiplication:
  If f, g are log-concave (convex), then $f(x)g(x)$ is also log-concave (convex)
  Since f, g are log-concave (convex),
  $\log f(x)g(x) = \log f(x) + \log g(x)$ is concave(convex)

- Log-convexity is closed under addition:
  If f, g are log-convex, then $f(x) + g(x)$ is also log-convex.
  Since f, g are log-convex, $F(x) = \log f(x)$, $G(x) = \log g(x)$ are convex, and so $\log\left(e^{F(x)} + e^{G(x)}\right) = \log(f + g)$ is convex

- Sum of log-concave functions is not, in general, log-concave

Log concavity without logarithm

- f is Log-concave if

$$f(\theta x + (1-\theta)y) \geq f(x)^{\theta} f(y)^{1-\theta},$$
for all $x, y \in dom(f), 0 \leq \theta \leq 1$

i.e., the value of a log-concave function at the average of two points is at least the geometric mean of the values at the two points

- If f is Log-convex, then it is convex, but not the converse

$\Rightarrow f(\theta x + (1-\theta)y) \leq f(x)^{\theta} f(y)^{1-\theta} \leq \theta f(x) + (1-\theta) f(y),$

$\Leftarrow f(\theta x + (1-\theta)y) = \theta f(x) + (1-\theta) f(y) > f(x)^{\theta} f(y)^{1-\theta},$

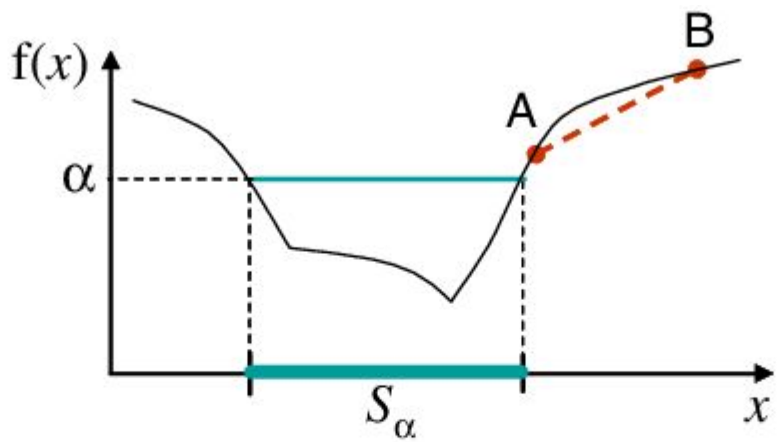
$$0 < \theta < 1, f(x) \neq f(y)$$

- If f is concave, then it is Log-concave, but not the converse: suppose, f is concave, and $0 < \theta < 1$, then $f(\theta x + (1-\theta)y) \geq \theta f(x) + (1-\theta) f(y) \geq f(x)^{\theta} f(y)^{1-\theta},$ Now suppose, $f(x) = x^2, x \in \mathbb{R}_{++}$, clearly, f is log-concave on $\mathbb{R}$ but not concave

Quasiconvex & Quasiconcave functions

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

$$S_\alpha = \{x \in \mathbf{dom}\, f \mid f(x) \leq \alpha\} \text{ are convex for all } \alpha$$



f is quasiconcave if dom f is convex and the superlevel sets $S_\alpha = \{x | f(x) \geq \alpha\}$ are convex for all α

Examples

$\sqrt{|x|}$ is quasiconvex on $\mathbf{R}$

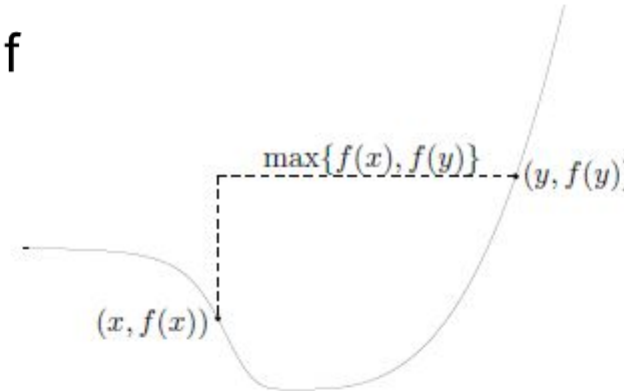
$f(x_1, x_2) = x_1 x_2$ is quasiconcave on $\mathbf{R}^2_{++}$

$\log x$ is quasilinear on $\mathbf{R}_{++}$

Quasiconvex functions

Modified Jensen's Inequaliy: f is quasiconvex if

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



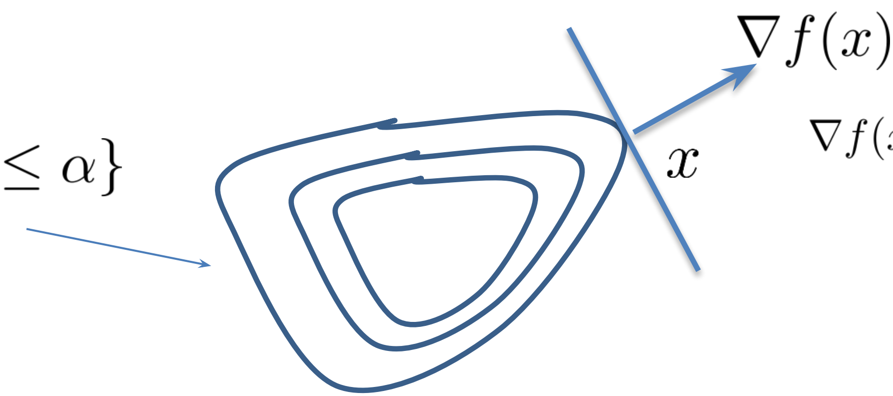
(Differentiable case): f is quasiconvex if and only if

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

equivalently $$\nabla f(x)^T(y-x) > 0 \Rightarrow f(y) > f(x)$$

Sublevel sets

$$S_\alpha = \{x | f(x) \leq \alpha\}$$



$\nabla f(x)$ is the normal to the supporting hyperplane to the sublevel sets

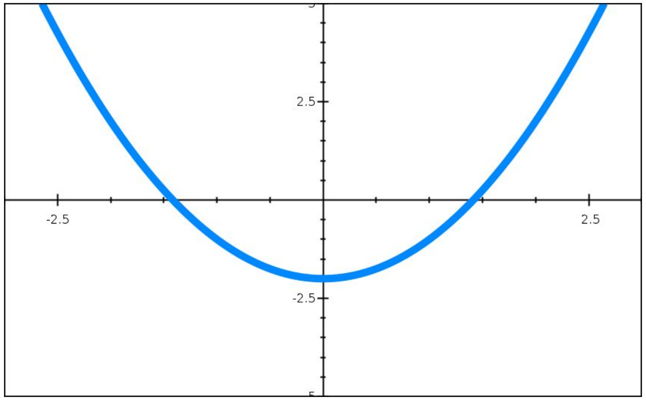
Strict & Strong Quasiconvex functions

f is strictly quasiconvex if for each x, y with $f(x) \neq f(y)$

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

$$f(x) = x^2 - 2$$

is strictly quasiconvex



$$f(\lambda x_1 + (1-\lambda)x_2) < max\{f(x_1), f(x_2)\}$$

Let $f: S \to \mathbb{R}$ be strictly quasiconvex. If x is a local minima to $min\ f(x), x \in S$. then it is also global minima *(Theorem 3.5.9, NLP book by Bazaraa)*

f is strongly quasiconvex if for each x, y with $x \neq y$

$$f(\theta x + (1-\theta)y) < \max\{f(x), f(y)\}), 0 < \theta < 1 \ (e.g., f(x) = x^2)$$

A strongly quasiconvex function is also strictly quasiconvex

(Definition 3.5.8, NLP book by Bazaraa)

Quasiconvex & Quasiconcave functions

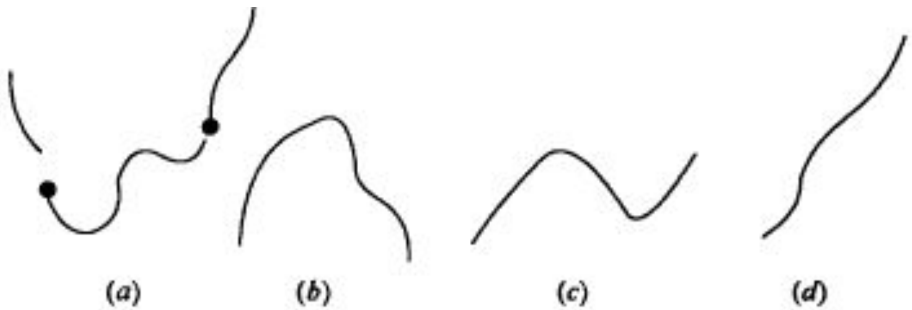


Figure 3.10 Quasiconvex and quasiconcave functions: (*a*) quasiconvex, (*b*) quasiconcave, (*c*) neither quasiconvex nor quasiconcave, (*d*) quasimonotone.

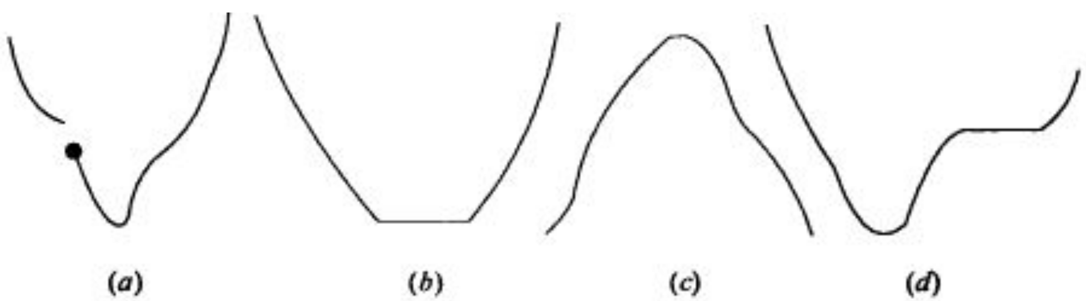


Figure 3.11 Strictly quasiconvex and strictly quasiconcave functions: (*a*) strictly quasiconvex, (*b*) strictly quasiconvex, (*c*) strictly quasiconcave, (*d*) neither strictly quasiconvex nor quasiconcave.

Source: Nonlinear Programming, Bazaraa

Pseudo-convex functions

$f : R^n \rightarrow R$ is pseudoconvex , if for each $x, y \in R^n$

$$\nabla f(x)^T (y - x) \geq 0 \Rightarrow f(y) \geq f(x)$$

Equivalently, $f(y) < f(x) \Rightarrow \nabla f(x)^T (y - x) < 0$

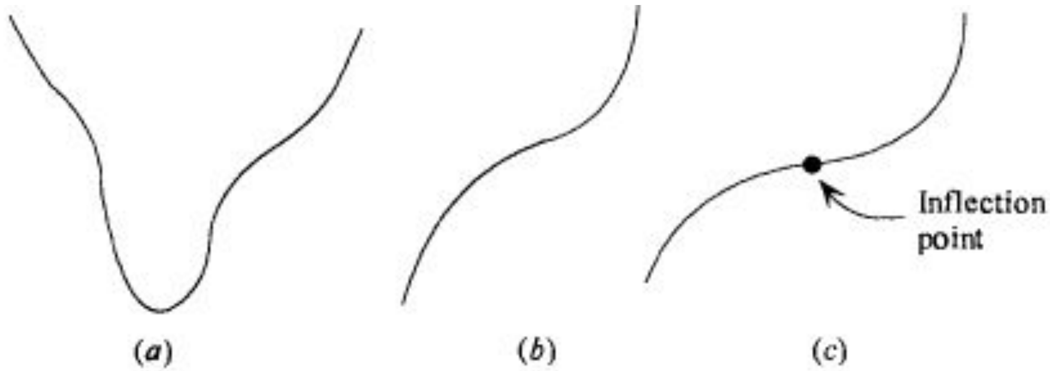


Figure 3.12 Pseudoconvex and pseudoconcave functions: (*a*) pseudoconvex, (*b*) both pseudoconvex and pseudoconcave, (*c*) neither pseudoconvex nor pseudoconcave.

Note that if $\nabla f(x) = 0$ then, $f(y) \geq f(x), \forall y$, i.e. ***x*** is global minima

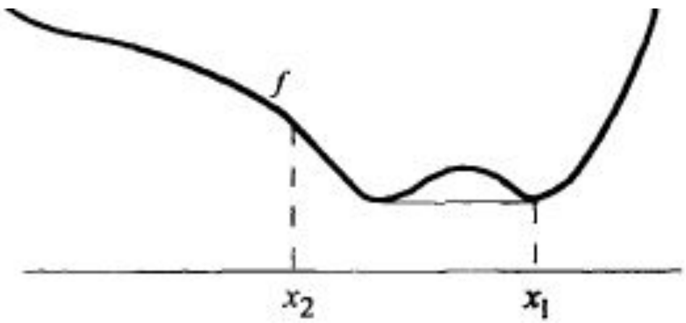
$f : R^n \rightarrow R$ is strictly pseudoconvex if for each $x \neq y$

$$\nabla f(x)^T (y - x) \geq 0 \Rightarrow f(y) \geq f(x)$$

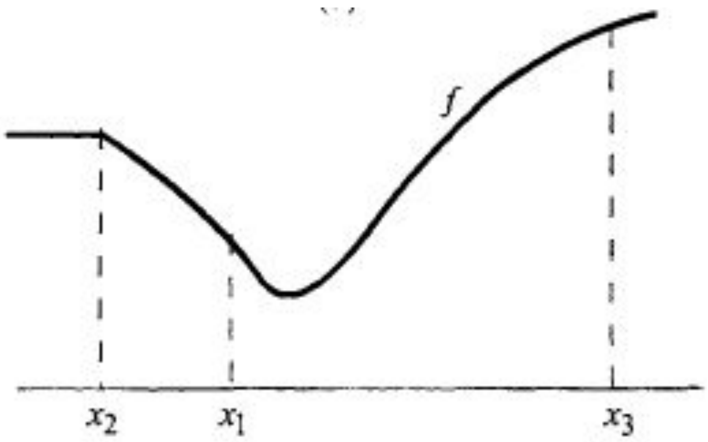
A pseudoconvex function is both quasiconvex and strictly quasiconvex

(Theorem 3.5.11, NLP book by Bazaraa)

Convexity at a point



f is pseudoconvex but not strictly pseudoconvex at x_1
f is both pseudoconvex and strictly pseudoconvex at x_2



f is quasiconvex but not strictly quasiconvex nor strongly quasiconvex at x_1
f is both quasiconvex and strictly quasiconvex at x_2 but not strongly quasiconvex
f is quasiconvex, strictly quasiconvex, and strongly quasiconvex at x_3

Linear and montonic functions

A function $f: \mathbf{X} \rightarrow \mathbf{R}$ is log-concave, and strict log-concave on **X** if $\log f$ is, respectively, concave, and strict concave on **X**.

A function $f: \mathbf{X} \rightarrow \mathbf{R}$ is linear if it is both convex and concave. The function is pseudolinear on **X** if it is both pseudoconvex and pseudoconcave.

A function $f: \mathbf{X} \rightarrow \mathbf{R}$ is monotonic if it is both quasiconcave and quasiconvex i.e. for all $x_1, x_2 \in \mathbf{X}$ and for all $\lambda \in [0, 1]$ we have

$$\min \{f(x_1), f(x_2)\} \leq f[\lambda x_1 + (1-\lambda) x_2] \leq \max \{f(x_1), f(x_2)\}$$

Relation between convex, log-convex, pseudo and quasiconvex functions

Let $\mathbf{X} \in \mathbf{R}^n$ be a convex set and let $f : \mathbf{X} \to \mathbf{R}$ be real valued function, then

a) If f is strict convex on X, then f is convex on X;
b) If f is log-convex on X, then f is convex on X;
c) If f is convex and differentiable on X (open), then f is pseudoconvex on X;
d) If f is pseudoconvex on X, then f is quasiconvex on X;

Note: the converse statements are not generally true

Proof of (b):

If f is log-convex on X, then from the relation between the arithmetic mean and the geometric mean it follows

$$f\left[\lambda x_1 + (1-\lambda)\, x_2\right] \le f(x_1)^{\lambda} f(x_2)^{1-\lambda} \le \lambda f(x_1) + (1-\lambda)\; f(x_2)$$

so, f is convex on X

Relation between convex, log-convex, pseudo and quasiconvex functions

Proof of (c)

Indeed, let $x_1, x_2 \in \mathbf{X}$, f be convex and differentiable such that $f(x_1) < f(x_2)$.

Using the property of convex function that for all $x_1, x_2 \in \mathbf{X}$, we have

$$f(x_1) - f(x_2) \geq (x_1 - x_2)^T \nabla f(x_2)$$

(i.e., function at x_1 lies above linear approximation at x_2), it follows that

$$0 > f(x_1) - f(x_2) \geq (x_1 - x_2)^T \nabla f(x_2)$$

so, f is pseudoconvex on X

Relation between convex, log-convex, pseudo and quasiconvex functions

Proof of (d)

Let $x_1, x_2 \in \mathbf{X}$, with $f(x_1) < f(x_2)$.

Let $x_\lambda = \lambda x_1 + (1-\lambda)x_2$ .We will show $f(x_\lambda) < f(x_2)$ for all $\lambda \in [0, 1]$, i.e., the function is non-decreasing as we move from x_1 to x_2.

Assume there exists a $\lambda_0 \in [0, 1]$ such that $f(x_{\lambda 0}) \geq f(x_2)$, i.e.,

$$\max_{\lambda \in [0,1]} f(x_\lambda) = f\left(x_{\lambda_0}\right), \; x_{\lambda_0} = \lambda_0 x_1 + (1-\lambda_0) x_2$$

Since function is assumed to attain maximum at $x_{\lambda 0}$, if we move away from $x_{\lambda 0}$ to x_1 or x_2 , the function will decrease and so directional derivative will be negative, i.e.,

$$(x_1 - x_{\lambda 0})^T \nabla f(x_{\lambda 0}) \leq 0 \text{ and } (x_2 - x_{\lambda 0})^T \nabla f(x_{\lambda 0}) \leq 0$$

Relation between convex, log-convex, pseudo and quasiconvex functions

Proof of (d)

Taking into account the value of $x_{\lambda 0}$ yields

$$(x_1 - x_2)^T \nabla f(x_{\lambda 0}) = 0$$

$$\text{i.e.,} \quad (x_1 - x_{\lambda 0})^T \nabla f(x_{\lambda 0}) = 0$$

Using the fact that f is pseudoconvex, it follows that

$$f(x_\lambda) \leq f(x_{\lambda 0}) \leq f(x_1)$$

From this inequality and from assumption that $f(x_{\lambda 0}) \geq f(x_2)$, it follows that $f(x_1) \geq f(x_2)$, which contradicts the assumption $f(x_1) < f(x_2)$ for x_1, $x_2 \in \mathbf{X}$, so it must be that $f(x_\lambda) < f(x_2)$, for all $\lambda \in [0, 1]$.

Relation between concave, log-concave, pseudo and quasiconcave functions

Let $\mathbf{X} \in \mathbf{R}^n$ be a convex set and let $f : \mathbf{X} \to \mathbf{R}$ be real valued function, then

a) If f is strictly concave on X, then f is convex on X;

b) If f is concave and positive on X , then f is log-concave

c) If f is log-concave and differentiable on X (open), then f is pseudoconcave on X;

d) If f is pseudoconcave on X, then f is quasiconcave on X;

Note: the converse statements are not generally true

Local and global minima of quasiconvex function

Let $\mathbf{X} \in \mathbf{R}^n$ be a convex set and let $f : \mathbf{X} \to \mathbf{R}$ be strictly quasiconvex on the convex set **X**, then any local minimum of function f is a global minimum of f on **X**.

Proof: Let $x_0 \in \mathbf{X}$, be a point of local minima.i.e., $f(x_0) < f(x)$ for all $x \in N_\varepsilon(x_0)$, for some $\varepsilon > 0$.
Let $x^* \in \mathbf{X}$, be a point of global minima, i.e. $f(x^*) < f(x_0)$
Since **X** is convex set, $\lambda x_0 + (1-\lambda)x^* \in \mathbf{X}$, for all $\lambda \in [0, 1]$.
But if $\lambda < \delta / ||x_0 - x^*||$ then $\lambda x_0 + (1-\lambda)x^* \in \mathbf{X} \cap N_\varepsilon(x^*)$ and so

$$f(\lambda x_1 + (1-\lambda)x_2) \geq f(x_0) \text{ —- (A)}$$

On the other hand, f being strictly quasiconvex on X, it follows that $f(\lambda x_1 + (1-\lambda)x_2) < \max [f(x_0), f(x^*)] = f(x_0)$ (because x^* is the global minima), and this contradicts statement (A)

Uniqueness of global minima of strictly convex function

Let $\mathbf{X} \in \mathbf{R}^n$ be a convex set and let $f : \mathbf{X} \to \mathbf{R}$ be real valued function. If f is strict convex on X, then the global minimum of the function f on X is unique;

Proof: Assume there are two different global minima at x_1 and x_2, i.e., $f(x_1)=f(x_2)$,with $x_1 \neq x_2$

Since f is strictly convex on **X**, it follows that for all $\lambda \in (0, 1)$, we have $f(\lambda x_1 + (1-\lambda)x_2) < \lambda f(x_1)+(1-\lambda) f(x_2)=f(x_1)=f(x_2)$

But this contradicts x_1 and x_2 being the minima

Uniqueness of global minima of strictly pseudoconvex function

Let $\mathbf{X} \in \mathbf{R}^n$ be a convex set and let $f : \mathbf{X} \to \mathbf{R}$ be real valued function. If f is strict pseudoconvex on **X**, then the global minimum of the function f on X is unique

Proof: Assume there are two different global minima at x_1 and x_2, i.e., $f(x_1)=f(x_2)$,with $x_1 \neq x_2$
If the function is strictly pseudoconvex, then $f(x_1) = f(x_2)$ implies $(x_1 - x_2)^T \nabla f(x_2) < 0$ which means if we move from x_2 along the direction $x_1 - x_2$ the function decreases. Since $x_1, x_2 \in \mathbf{X}$ is convex, we can select $\lambda \in [0, 1]$, such that $x_\lambda = \lambda x_1 + (1-\lambda)x_2$ .and $f(x_\lambda) < f(x_2) = f(x_1)$,

But this contradicts x_1 and x_2 being minima