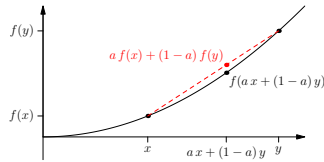


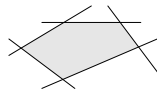
Convexity



Convex sets, convex functions, Jensen's inequality

Convex Regions

- Convex regions are familiar



- For any two points x and y in a region \mathcal{R} then for any $a \in [0, 1]$ if

$$z = ax + (1 - a)y \in \mathcal{R}$$

- then \mathcal{R} is a convex region



Positive Semi-Definite Matrices

- Recall that a matrix \mathbf{M} is positive semi-definite if for any vector v

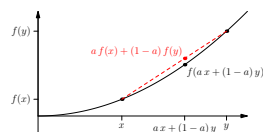
$$v^T \mathbf{M} v \geq 0$$

(i.e. any quadratic form of the matrix is non-negative)

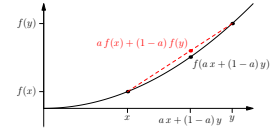
- (We showed this also implies that all the eigenvalues are non-negative)
- We denote the fact that \mathbf{M} is positive semi-definite by $\mathbf{M} \succeq 0$, and $\mathbf{M} \succ 0$ if it is positive definite
- The set of positive semi-definite (PSD) matrices (or kernels) form a convex set

Outline

- Convex sets
- Convex functions
- Jensen's inequality



- Convex sets
- Convex functions
- Jensen's inequality



Convex Sets

- For any set, \mathcal{S} , where addition and scalar multiplication is defined (e.g. a vector space) then:

If for any two elements $x, y \in \mathcal{S}$ and any $a \in [0, 1]$

$$z = ax + (1 - a)y \in \mathcal{S}$$

then \mathcal{S} is said to be a convex set

Proof

- Consider any two arbitrarily chosen PSD matrices \mathbf{M}_1 and \mathbf{M}_2 and any $a \in [0, 1]$ then let

$$\mathbf{M}_3 = a\mathbf{M}_1 + (1 - a)\mathbf{M}_2$$

- Then for any vector v

$$\begin{aligned} v^T \mathbf{M}_3 v &= v^T (a\mathbf{M}_1 + (1 - a)\mathbf{M}_2) v \\ &= av^T \mathbf{M}_1 v + (1 - a)v^T \mathbf{M}_2 v \\ &= am_1 + (1 - a)m_2 \end{aligned}$$

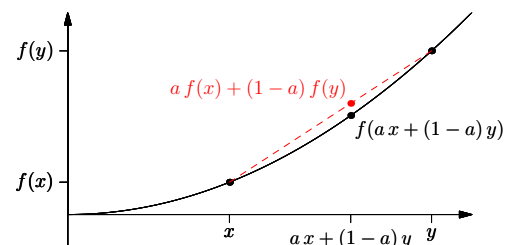
where $m_1 = v^T \mathbf{M}_1 v$ and $m_2 = v^T \mathbf{M}_2 v$

- But $m_1, m_2 \geq 0$ since $\mathbf{M}_1, \mathbf{M}_2 \succeq 0$. Thus $am_1 + (1 - a)m_2 \geq 0$ and so $\mathbf{M}_3 \succeq 0$ \square

Convex Functions

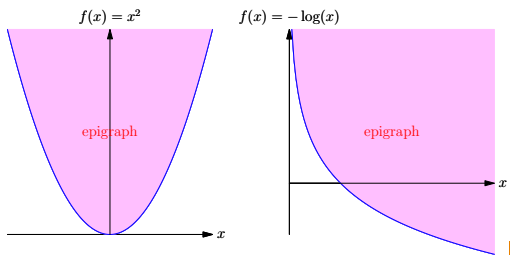
- Any function $f(x)$ is said to be a **convex function** if for any two points x and y and any $a \in [0, 1]$

$$f(ax + (1 - a)y) \leq af(x) + (1 - a)f(y)$$



Epigraph

- The **epigraph** of a function is the area that lies above the function
- The epigraph of a convex function is a convex region



Convex-Down or Concave Functions

- Any function, $f(x)$, that satisfies the inverse inequality

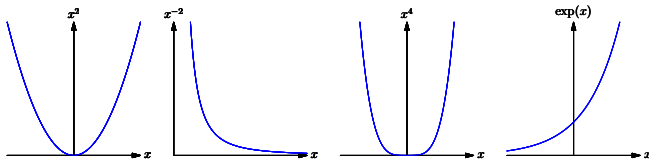
$$f(ax + (1-a)y) \geq af(x) + (1-a)f(y)$$

for any points x and y and any $a \in [0,1]$ is said to be a **convex-down** or **concave** function

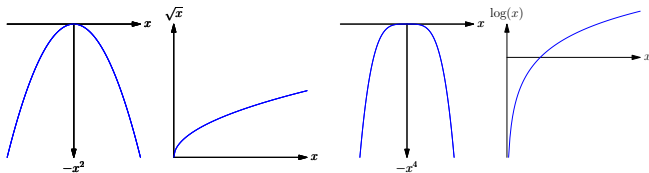
- Everything true for a convex(-up) function carries over to a convex-down function with a small modification
- If $f(x)$ is a convex-up function then $g(x) = -f(x)$ is a convex-down function
- The area that lies below a convex-down function is a convex region

Examples

Convex-Up Functions



Convex-Down Functions



Linear Functions

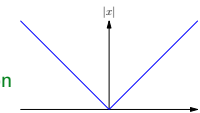
- Linear functions are given by

$$f(x) = mx + c$$

- They satisfy the **equality**

$$f(ax + (1-a)y) = af(x) + (1-a)f(y)$$

- As such they are both convex(-up) and convex-down function



- $|x|$ is a convex-up function

Strictly Convex Function

- Functions that satisfy the strict inequality (for $0 < a < 1$ and $x \neq y$)

$$f(ax + (1-a)y) < af(x) + (1-a)f(y)$$

are said to be **strictly convex functions**

- A strictly convex-down function satisfies the reverse strict inequality
- Strictly convex-(up or down) functions don't contain any linear regions

Convexity in High Dimensions

- If $f: \mathbb{R}^n \rightarrow \mathbb{R}$ (i.e. $f(x)$ maps high dimensional point $x \in \mathbb{R}^n$ to a real value) satisfies

$$f(ax + (1-a)y) \leq af(x) + (1-a)f(y)$$

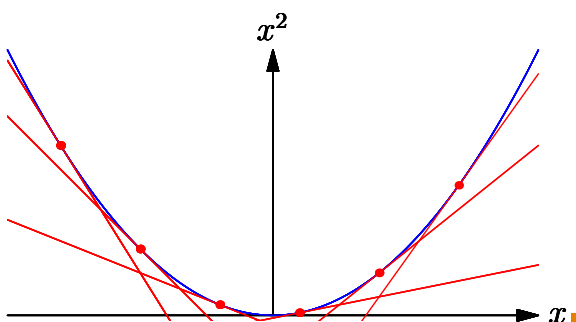
for any $x, y \in \mathbb{R}^n$ and any $a \in [0,1]$ then $f(x)$ is a convex function

- $\|x\|_2^2 = \sum_i x_i^2$ is a (strictly) convex function
- $\|x\|_1 = \sum_i |x_i|$ is a convex function

Properties of Convex Functions

- Convex functions lie on or above any tangent line

$$f(x) \geq f(x^*) + (x - x^*)f'(x^*)$$



Second Derivatives

- As $f(x)$ lies on or above its tangent line then for any $\epsilon > 0$

$$f'(x + \epsilon) \geq f'(x)$$

therefore $f''(x) = \lim_{\epsilon \rightarrow 0} (f'(x + \epsilon) - f'(x))/\epsilon \geq 0$ at all points x

- In high dimensions a convex function lies above its tangent plane

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

- The matrix of second derivatives (the Hessian) must be positive semi-definite at all points x

$$H = \begin{pmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} \succeq 0$$

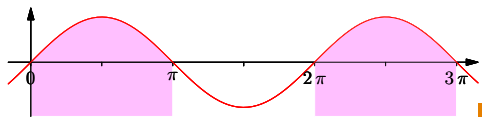
Proving Convexity

- $f(x) = x^2$ is strictly convex as $f''(x) = 2 > 0$
- $f(x) = e^{cx}$ is strictly convex as $f''(x) = c^2 e^{cx} > 0$
- $f(x) = \log(x)$ is strictly convex-down as $f''(x) = -\frac{1}{x^2} < 0$
- $f(x, y) = \frac{x^2}{y}$ is convex for $y > 0$ as

$$\mathbf{H} = \begin{pmatrix} \frac{\partial^2 f(x, y)}{\partial x^2} & \frac{\partial^2 f(x, y)}{\partial x \partial y} \\ \frac{\partial^2 f(x, y)}{\partial y \partial x} & \frac{\partial^2 f(x, y)}{\partial y^2} \end{pmatrix} = \begin{pmatrix} \frac{2}{y} & -\frac{2x}{y^2} \\ -\frac{2x}{y^2} & \frac{2x^2}{y^3} \end{pmatrix} = \frac{2}{y^3} \begin{pmatrix} y^2 & -xy \\ -xy & x^2 \end{pmatrix}$$
- Now $T = \text{tr} \mathbf{H} = \frac{2}{y^3}(x^2 + y^2)$, $D = \det(\mathbf{H}) = 0$
- $\lambda_{1,2} = T/2 \pm \sqrt{T^2/4 - D} = \{0, T\} = \{0, \frac{2(x^2 + y^2)}{y^3}\} \geq 0 \Rightarrow \mathbf{H} \succeq 0$

Convex Functions Defined on Convex Sets

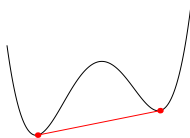
- All the properties we have discussed hold for functions defined on a convex set
- $\sin(x)$ is not generally neither convex up or down
- $\sin(x)$ for $x \in [0, \pi]$ is convex-down (its second derivative $-\sin(x)$ is less than or equal to 0 in this range)



- For a convex function defined on a non-convex set, \mathcal{S} , there exists points $x, y \in \mathcal{S}$ such that for some $a \in [0, 1]$ there will be points $z = ax + (1 - a)y \notin \mathcal{S}$ (the function isn't defined on such points)

Unique Minimum

- Strictly convex function have a unique minimum
- The existence of a local minimum would break convexity
 - ★ The line connecting a local minimum to a global minimum would be strictly decreasing
 - ★ Thus there are points next to the local minimum with lower values
 - ★ This is a contradiction
- This remains true if we consider convex functions that are constrained to live in a convex set



Linear Regression

- For linear regression the loss function

$$L(w) = \|Xw - y\|^2 = w^T X^T X w - 2w^T X^T y + y^T y$$

is convex

- Since the Hessian $\mathbf{H} = 2X^T X \succeq 0$ (positive semi-definite) (For any vector v then $v^T \mathbf{H} v = 2v^T X^T X v = 2\|Xv\|^2 \geq 0$)
- If $\mathbf{H} \succ 0$ there will be a unique minima while if \mathbf{H} has some zero eigenvalues there will be a family of solutions

Sums of Convex Functions

- For any set of convex functions $f_1(x), f_2(x), \dots$ and any set of non-negative scalars a_1, a_2, \dots then

$$g(x) = \sum_i a_i f_i(x)$$

is convex

- Proof

$$g''(x) = \sum_i a_i f_i''(x)$$

but $f_i''(x) \geq 0$ so $g''(x)$ is a sum on non-negative terms

- This generalises to higher dimensions as the sum of PSD matrices times positive factors is a PSD matrix

Constraints

- Often we impose constraints on the set of points, e.g.

$$x_i > 0 \quad a^T x = b \quad x^T M x \leq 1$$

- Linear constraints (e.g. $x_i > 0$ or $a^T x = b$ or $a^T x \leq b$) always define a convex region
- Multiple linear constraints always define a convex region
- Non-linear constraints may or may not define a convex region ($\{x \in \mathbb{R}^n | x^T M x \leq 1, M \succeq 0\}$ does while $\{x \in \mathbb{R}^n | x^T M x \geq 1, M \succeq 0\}$ doesn't)

Convex Set of Minima

- If $f(x)$ is **convex** but not **strictly convex** then there might exist a convex set $\mathcal{M} \subset \mathcal{X}$ of minima such that for all $x, y \in \mathcal{M}$ and any $z \in \mathcal{X}$ we have $f(x) = f(y) \leq f(z)$
- This set of minima is convex, that is, if $x, y \in \mathcal{M}$ then for any $a \in [0, 1]$ the point $z = ax + (1 - a)y \in \mathcal{M}$
- The sum of a convex function, $f(x)$, and a strictly convex function $g(x)$ will always be strictly convex since

$$f''(x) + g''(x) > 0$$

as $f''(x) \geq 0$ and $g''(x) > 0$

Regularised Linear Regression

- In ridge regression we minimise a loss

$$L(w) = \|Xw - y\|^2 + \eta \|w\|^2 = w^T (X^T X + \eta I) w - 2w^T X^T y + y^T y$$

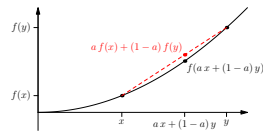
- Because $\|w\|^2$ is strictly convex the loss function is strictly convex and so will have a unique solution
- Using an L_1 regulariser (Lasso)

$$L(w) = \|Xw - y\|^2 + \eta \|w\|_1$$

again $\|w\|_1$ is convex and so $L(w)$ will be convex

- Using an L_1 and an L_2 regulariser (elastic net) also gives a unique solution

1. Convex sets
2. Convex functions
3. Jensen's inequality



Proof

- We said before that a convex function must lie on or above its tangent plane at any point x^*

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

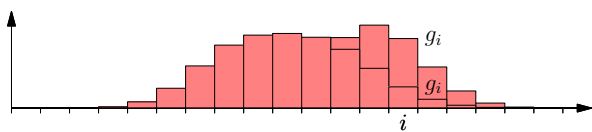
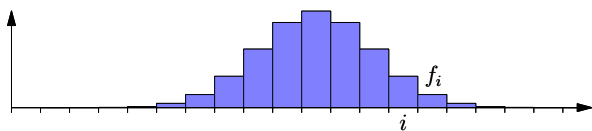
- Taking $x^* = \mathbb{E}[X]$

$$f(X) \geq f(\mathbb{E}[X]) + (X - \mathbb{E}[X])^T \nabla f(\mathbb{E}[X])$$

- Taking expectations of both sides

$$\begin{aligned} \mathbb{E}[f(X)] &\geq f(\mathbb{E}[X]) + (\mathbb{E}[X] - \mathbb{E}[X])^T \nabla f(\mathbb{E}[X]) \\ &= f(\mathbb{E}[X]) \quad \square \end{aligned}$$

Kullback-Leibler Divergence



$$\text{KL}(f\|g) = -\sum_{i=1}^n f_i \log\left(\frac{g_i}{f_i}\right) = 0.235$$

Lessons

- Although we haven't talked much about machine learning, convexity is heavily used in many machine learning applications
- A lot of ML algorithms involve convex functions e.g. SVMs
- As such they will have a unique minimum (or a convex set of minima)
- Convexity is an elegant idea which is relatively easy to prove theorems about
- One of the most useful tools is Jensen's inequality

- In proving many properties of learning machines inequalities are really useful
- One of the most useful inequalities involve expectations of convex functions, this is known as **Jensen's Inequality**

- If $f(x)$ is a convex(-up) function then

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$$

- If $f(x)$ is a convex(-down) function then

$$\mathbb{E}[f(X)] \leq f(\mathbb{E}[X])$$

Simple Proofs with Jensen's Inequality

- Since $f(x) = x^2$ is convex by Jensen's inequality

$$\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2 \quad \text{or} \quad \mathbb{E}[X^2] - \mathbb{E}[X]^2 \geq 0$$

(i.e. variance are non-negative)

- The KL-divergence $\text{KL}(f\|g)$ between two categorical probability distributions (f_1, f_2, \dots) and (g_1, g_2, \dots) is defined as

$$\text{KL}(f\|g) = -\sum_i f_i \log\left(\frac{g_i}{f_i}\right)$$

(note $f_i, g_i \geq 0$ and $\sum_i f_i = \sum_i g_i = 1$)

Proof of Gibbs' Inequality

- To show that $\text{KL}(f\|g) \geq 0$ (Gibbs' inequality) we note that since the logarithm is a convex-down function

$$\begin{aligned} \text{KL}(f\|g) &= -\sum_i f_i \log\left(\frac{g_i}{f_i}\right) = \mathbb{E}_f\left[-\log\left(\frac{g_i}{f_i}\right)\right] \\ &\geq -\log\left(\mathbb{E}_f\left[\frac{g_i}{f_i}\right]\right) \\ &= -\log\left(\sum_i f_i \frac{g_i}{f_i}\right) = -\log\left(\sum_i g_i\right) = -\log(1) = 0 \end{aligned}$$

- We will meet KL-divergences later on