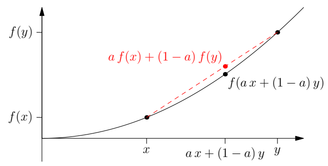
Advanced Machine Learning

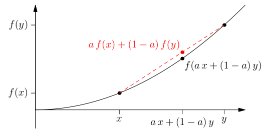
Convexity



Convex sets, convex functions, Jensen's inequality

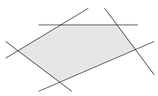
Outline

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



Convex Regions

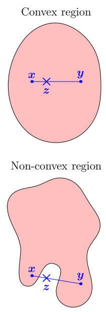
- Convex regions are familiar



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

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

- then $\mathcal{R}$ is a convex region



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 $\boldsymbol{x},\boldsymbol{y} \in \mathcal{S}$ and any $a \in [0,1]$

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

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

Positive Semi-Definite Matrices

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

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

  (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

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 $\boldsymbol{v}$

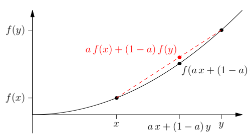
$$\begin{aligned}\boldsymbol{v}^{\mathsf{T}}\mathbf{M}_3\boldsymbol{v} &= \boldsymbol{v}^{\mathsf{T}}(a\mathbf{M}_1 + (1-a)\mathbf{M}_2)\boldsymbol{v}\\ &= a\boldsymbol{v}^{\mathsf{T}}\mathbf{M}_1\boldsymbol{v} + (1-a)\boldsymbol{v}^{\mathsf{T}}\mathbf{M}_2\boldsymbol{v}\\ &= am_1 + (1-a)m_2\end{aligned}$$

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

- 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 $\mathbf{M}_3 \succeq 0$ $\square$

Outline

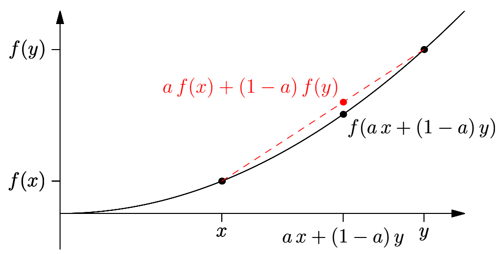
1. Convex sets
2. **Convex functions**
3. Jensen's inequality



Convex Functions

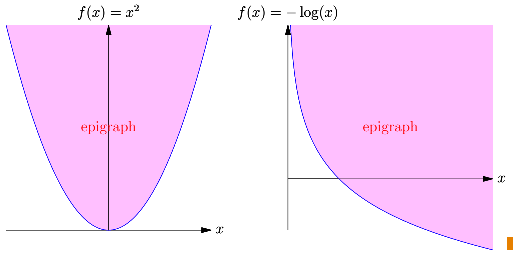
- Any function $f(x)$ is said to be a **convex function** if for any two points x and y and any $a \in [0,1]$

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



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

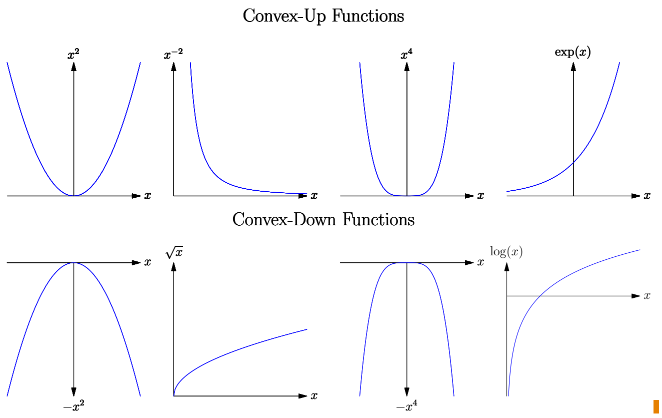
- Any function, $f(x)$, that satisfies the inverse inequality

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

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



Linear Functions

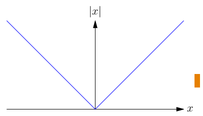
- Linear functions are given by

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

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

$$f(a\boldsymbol{x} + (1-a)\boldsymbol{y}) \leq af(\boldsymbol{x}) + (1-a)f(\boldsymbol{y})$$

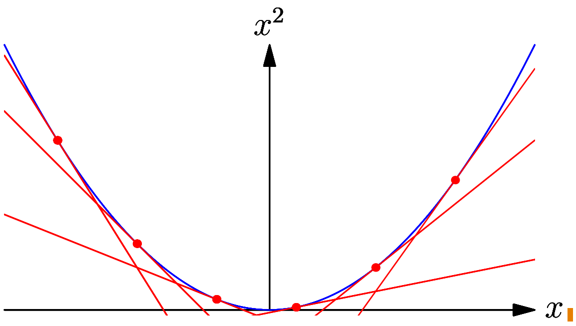
for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ and any $a \in [0,1]$ then $f(\boldsymbol{x})$ is a convex function

- $\|\boldsymbol{x}\|_2^2 = \sum_i x_i^2$ is a (strictly) convex function
- $\|\boldsymbol{x}\|_1 = \sum_i |x_i|$ 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 $f''(x) = \lim_{\epsilon \to 0}(f'(x+\epsilon) - f'(x))/\epsilon \geq 0$ at all points x

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

$$f(\boldsymbol{x}) \geq f(\boldsymbol{x}^*) + (\boldsymbol{x} - \boldsymbol{x}^*)^{\mathsf{T}} \boldsymbol{\nabla} f(\boldsymbol{x}^*)$$

- The matrix of second derivatives (the Hessian) must be positive semi-definite at all points $\boldsymbol{x}$

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

Proving Convexity

- $f(x) = x^2$ is strictly convex as $f''(x) = 2 > 0$
- $f(x) = \mathrm{e}^{cx}$ is strictly convex as $f''(x) = c^2 \mathrm{e}^{cx} > 0$
- $f(x) = \log(x)$ 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 $T = \mathrm{tr}\,\mathbf{H} = \frac{2}{y^3}(x^2 + y^2)$, $D = \det(\mathbf{H}) = 0$
- $\lambda_{1,2} = T/2 \pm \sqrt{T^2/4 - D} = \{0, T\} = \{0, \frac{2(x^2+y^2)}{y^3}\} \geq 0 \Rightarrow \mathbf{H} \succeq 0$

Sums of Convex Functions

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

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

is convex

- Proof

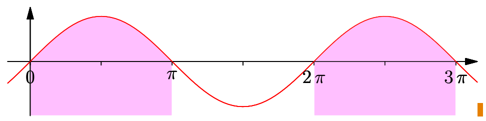
$$g''(x) = \sum_i a_i f_i''(x)$$

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 matix

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

Constraints

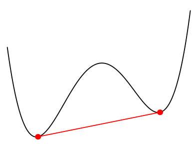
- Often we impose constraints on the set of points, e.g.

$$x_i > 0 \qquad \boldsymbol{a}^\mathsf{T}\boldsymbol{x} = b \qquad \boldsymbol{x}^\mathsf{T}\mathbf{M}\boldsymbol{x} \leq 1$$

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

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

Convex Set of Minima

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

Linear Regression

- For linear regression the loss function

$$L(\boldsymbol{w}) = \|\mathbf{X}\boldsymbol{w} - \boldsymbol{y}\|^2 = \boldsymbol{w}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\boldsymbol{w} - 2\boldsymbol{w}^\mathsf{T}\mathbf{X}^\mathsf{T}\boldsymbol{y} + \boldsymbol{y}^\mathsf{T}\boldsymbol{y}$$

is convex

- Since the Hessian $\mathbf{H} = 2\boldsymbol{X}^\mathsf{T}\boldsymbol{X} \succeq 0$ (positive semi-definite) (For any vector $\boldsymbol{v}$ then $\boldsymbol{v}^\mathsf{T}\mathbf{H}\boldsymbol{v} = 2\boldsymbol{v}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\boldsymbol{v} = 2\|\mathbf{X}\boldsymbol{v}\|^2 \geq 0$)
- 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

Regularised Linear Regression

- In ridge regression we minimise a loss

$$L(\boldsymbol{w}) = \|\mathbf{X}\boldsymbol{w} - \boldsymbol{y}\|^2 + \eta\|\boldsymbol{w}\|^2 = \boldsymbol{w}^\mathsf{T}\left(\mathbf{X}^\mathsf{T}\mathbf{X} + \eta\mathbf{I}\right)\boldsymbol{w} - 2\boldsymbol{w}^\mathsf{T}\mathbf{X}^\mathsf{T}\boldsymbol{y} + \boldsymbol{y}^\mathsf{T}\boldsymbol{y}$$

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

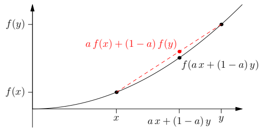
$$L(\boldsymbol{w}) = \|\mathbf{X}\boldsymbol{w} - \boldsymbol{y}\|^2 + \eta\|\boldsymbol{w}\|_1$$

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

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

Outline

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



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 $f(\boldsymbol{x})$ is a convex(-up) function then

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

- If $f(\boldsymbol{x})$ is a convex(-down) function then

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

Proof

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

$$f(\boldsymbol{x}) \geq f(\boldsymbol{x}^*) + (\boldsymbol{x} - \boldsymbol{x}^*)^{\mathsf{T}} \boldsymbol{\nabla} f(\boldsymbol{x}^*)$$

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

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

- Taking expectations of both sides

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

Simple Proofs with Jensen's Inequality

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

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

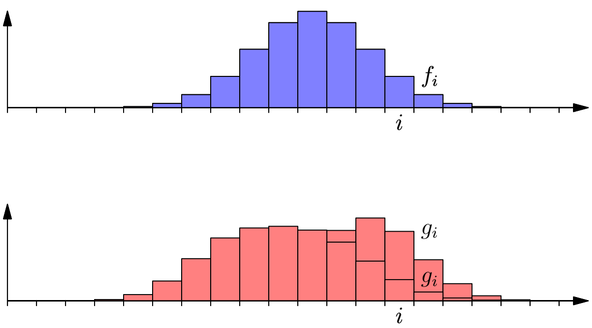
(i.e. variance are non-negative)

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

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

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

Kullback-Leibler Divergence



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

Proof of Gibbs' Inequality

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

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

- We will meet KL-divergences later on

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