Introduction to Machine Learning (CS 771A, IIT Kanpur)

Course Notes and Exercises

Suggested Citation: P. Kar. IIT Kanpur CS 771A, Course Notes and Exercises: Introduction to Machine Learning, 2019.

Purushottam Kar
IIT Kanpur
purushot@cse.iitk.ac.in

IIT Kanpur

Contents

Introduction to Machine Learning (CS 771A, IIT Kanpur)

Purushottam Kar[1]*

[1]*IIT Kanpur; purushot@cse.iitk.ac.in*

ABSTRACT

Machine Learning is the art and science of designing algorithms that can learn patterns and concepts from data to modify their own behavior without being explicitly programmed to do so. This monograph is intended to accompany a course on an introduction to the design of machine learning algorithms with a modern outlook. Some of the topics covered herein are *Preliminaries* (multivariate calculus, linear algebra, probability theory), *Supervised Learning* (local/proximity-based methods, learning by function approximation, learning by probabilistic modeling), *Unsupervised Learning* (discriminative models, generative models), practical aspects of machine learning, and additional topics.

Although the monograph will strive to be self contained and revisit basic tools in areas such as calculus, probability, and linear algebra, the reader is advised to not completely rely on these refresher discussions but rather refer to a standard textbook devoted to these topics.

*The contents of this monograph were developed as a part of successive offerings of various machine learning related courses at IIT Kanpur.

1

Introduction

2

Learning with Prototypes

3

Nearest Neighbors

4

Decision Trees

5

Support Vector Machines

Acknowledgements

The author is thankful to the students of successive offerings of the course for their inputs and pointing out various errata in the lecture material. This monograph was typeset using the beautiful style of the Foundations and Trends® series published by now publishers.

Appendices

A

Vector Space Refresher

B

Calculus Refresher

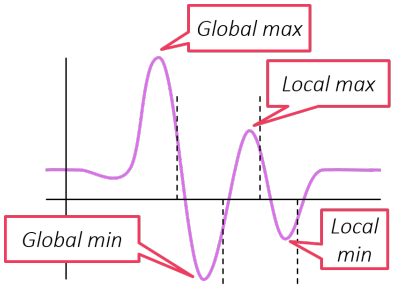
In this chapter we will take a look at basic tools from calculus that would be required to design and execute machine learning algorithms. Before we proceed, we caution the reader that the treatment in this chapter will *not* be mathematically rigorous and frequently, we will appeal to concepts and results based on informal arguments and demonstration, rather than proper proofs. This was done in order to provide the reader with a working knowledge of the topic without getting into excessive formalism. We direct the reader to texts in mathematics, of which several excellent ones are available, for a more rigorous treatment of this subject.

B.1 Extrema

The vast majority of machine learning algorithms learn models by trying to obtain the best possible performance on training data. What changes from algorithm to algorithm is how "performance" is defined and what constitutes "best". Frequently, performance can be defined in terms of an objective function f that takes in a model (say, a linear model $\mathbf{w}$) and outputs a real number $f(\mathbf{w}) \in \mathbb{R}$ called the objective value. Depending on the algorithm designer a large objective value may be better or a small score may be better (e.g. if f encodes margin then we want a large objective value, on the other hand if f encodes the classification error then we want a small objective value).

objective function

objective value



Given a function $f : \mathbb{R}^d \to \mathbb{R}$, a point $\mathbf{x}^* \in \mathbb{R}^d$ is said to be the global maximum of this function if $f(\mathbf{x}^*) \geq f(\mathbf{x})$ for all other $\mathbf{x} \in \mathbb{R}^d$. We similarly define a global minimum of this function as a point $\tilde{\mathbf{x}}$ such that $f(\tilde{\mathbf{x}}) \leq f(\mathbf{x})$ for all other $\mathbf{x} \in \mathbb{R}^d$. Note that a function may have multiple global maxima and global minima. For example the function $f(x) = \sin(x)$ has global maxima at all values of x that are of the form $2k\pi + \frac{\pi}{2}$

global maximum

global minimum

and global minima at all values of x that are of the form $2k\pi - \frac{\pi}{2}$.

However, apart from global extrema which achieve the largest or the smallest value of a function among all possible input points, we can also have local extrema, i.e. local minimum and local maximum. These are points which achieve the best value of the function (min for local minima and max for local maxima) only in a certain (possibly small) region surrounding the point.

local extrema

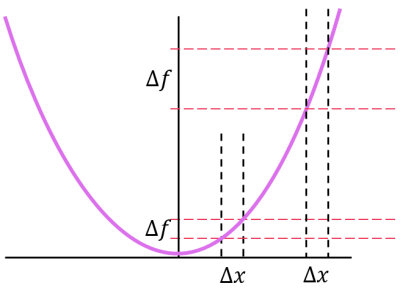
A practical example to understand the distinction between local and global extrema can be that of population: the city of Kanpur has a large population (that of 3.2 million) which is the highest only if we restrict ourselves to cities within the state of Uttar Pradesh and is thus a local maximum. If we go outside the state, we find cities like Mumbai with a population of 12.4 million. However, even Mumbai is a local maxima since the global maxima is achieved at Chongqing, China which has a population of 30.1 million (source: Wikipedia).

It is be clear from the above definitions that all global extrema are necessarily local extrema. For example, Chongqing clearly has the largest population within China itself and thus a local maximum. However, not all local extrema need be global extrema.

B.2 Derivatives

Derivatives are an integral part of calculus (pun intended) and are the most direct way of finding how function values change (increase/decrease) if we move from one point to another. Given a univariate function i.e. a function $f : \mathbb{R} \to \mathbb{R}$ that takes a single real number as input and outputs a real number (we will take care of multivariate functions later), the derivative of f at a point x^0 tells us two things. Firstly, if the sign of the derivative is positive i.e. $f'(x^0) > 0$, then the function value will increase if we move a little bit to the right on the number line (i.e. go from x^0 to $x^0 + \Delta x$ for some $\Delta x > 0$) and it will decrease if we move a little bit to the left on the number line. Similarly if $f'(x^0) < 0$, then moving right decreases the function value whereas moving left increases the function value.

univariate function



Secondly, the magnitude of the derivative i.e. $|f'(x^0)|$ tells us by how much would the function value increase or decrease if we move a little bit left or right from the point x^0. For example, consider the function $f(x) = x^2$. Its derivative is $f'(x) = 2x$. This tells us that if $x^0 < 0$ (where the derivative is negative), the function value would decrease if we moved right and increase if we moved left. Similarly, if $x^0 > 0$, the derivative is positive and thus, the function value would increase if we moved to the right and decrease if we moved to the left. However, since the magnitude of the derivative is $2|x|$ which increases as we go away from the origin, it can be seen that the increase in function value, for the same change in the value of x^0 s much steeper if x^0 is far from the origin.

It is important to note that the above observations (e.g. function value goes

up if $f(x^0) > 0$ and we move to the right) hold true only if the movement Δx is "small". For example, $f(x) = x^2$ has a negative derivative at $x^0 = -2$ and so the function value should decrease if we moved right little bit. However, if we move right too much (say we move to $x^0 = 3$) then the above promise does not hold since $f(3) = 9 > 4 = f(-2)$. In fact a corollary of the Taylor's theorem states

Taylor's theorem (first order)

$$f(x^0 + \Delta x) \approx f(x^0) + f'(x) \cdot \Delta x, \text{if } \Delta x \text{ is "small".}$$

How small is small enough for the above result to hold may depend on both the function f as well as the point x^0 where we are applying the result.

B.3 Second Derivative

Just as the derivative of a function tells us how does the function value changes (i.e. goes up/down) and by how much, the second derivative tells us how does the derivative change (i.e. go up/down) and by how much. Intuitively, the second derivative can be thought of as similar to acceleration if we consider the derivative as similar to velocity and the function value as being similar to displacement. If at a point x^0 we have $f''(x^0) > 0$, then this means that the derivative will go up if we move to the right and decrease if we move to the left (similarly if $f''(x^0) < 0$ at a point).

The Taylor's theorem does extend to second order derivatives as well

$$f'(x^0 + \Delta x) \approx f'(x^0) + f''(x^0) \cdot \Delta x, \text{if } \Delta x \text{ is "small".}$$

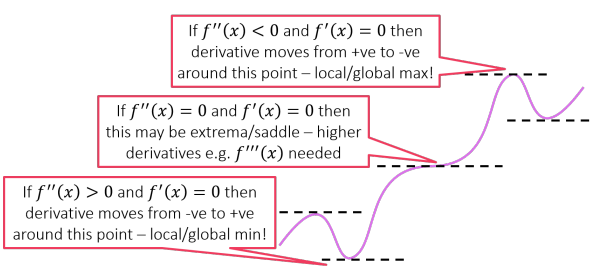
Integrating both sides and applying the fundamental theorem of algebra

$$f(x^0 + \Delta x) \approx f(x^0) + f'(x^0) \cdot \Delta x + \frac{1}{2} f''(x^0) \cdot (\Delta x)^2, \text{if } \Delta x \text{ is "small".}$$

Taylor's theorem (second order)

Although the above derivation is not strictly rigorous, the result is nevertheless true. Thus, knowing the second derivative can help us get a better approximation of the change in function value if we move a bit. The second derivative is most commonly used in machine learning in designing very efficient optimization algorithms (known as *Newton methods* which we will study later). In fact there exist 3rd and higher order derivatives as well (the third derivative telling us how does the second derivative change from point to point etc) but since they are not used all that much, we will not study them here.

B.4 Stationary Points



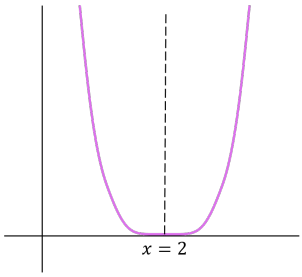
The stationary points of a function $f : \mathbb{R} \to \mathbb{R}$ are defined as the points where the derivative of the function vanishes i.e. $f'(x) = 0$. The stationary points of a function correspond to either the

local/global maxima or minima or else saddle points. Given a stationary point, the second derivative test is used to distinguish extrema from saddle points. second derivative test

If the second derivative of the function is positive at a stationary point x^0 i.e. $f'(x^0) = 0$ and $f''(x^0) > 0$ then x^0 is definitely a local minimum. This result follows directly from the second order Taylor's theorem we studied above. Since $f'(x^0) = 0$, we have

$$f(x^0 + \Delta x) \approx f(x^0) + \frac{1}{2} f''(x^0) \cdot (\Delta x)^2 \geq f(x^0)$$

This means that irrespective of whether $\Delta x < 0$ or $\Delta x > 0$ (i.e. irrespective of whether we move left or right), the function value always increases. Recall that this is the very definition of a local minimum. Similarly, we can intuitively see that if $f'(x^0) = 0$ and $f''(x^0) < 0$ then x^0 is definitely a local maximum.



If we have $f'(x^0) = 0$ and $f''(x^0) = 0$ at a point then the second derivative test is actually silent and fails to tell us anything informative. The reader is warned that the first and second derivatives both vanishing *does not* mean that the point is a saddle point. For example, consider the case of the function $f(x) = (x-2)^4$. Clearly $x^0 = 2$ is a local (and global) minimum. However, it is also true that $f'(2) = 0 = f''(2)$. In such inconclusive cases, higher order derivatives e.g. $f^{(3)}(x) = f'''(x), f^{(4)}(x)$ have to be used to figure out what is the status of our stationary point.

B.5 Useful Rules for Calculating Derivatives

Several rules exist that can help us calculate the derivative of complex-looking functions with relative ease. These are given below followed by some examples applying them to problems.

1. (Constant Rule) If $h(x) = c$ where c is not a function of x then $h'(x) = 0$
2. (Sum Rule) If $h(x) = f(x) + g(x)$ then $h'(x) = f'(x) + g'(x)$
3. (Scaling Rule) If $h(x) = c \cdot f(x)$ and if c is not a function of x then $h'(x) = c \cdot f'(x)$
4. (Product Rule) If $h(x) = f(x) \cdot g(x)$ then $h'(x) = f'(x) \cdot g(x) + g'(x) \cdot f(x)$
5. (Quotient Rule) If $h(x) = \frac{f(x)}{g(x)}$ then $h'(x) = \frac{f'(x) \cdot g(x) - g'(x) \cdot f(x)}{g^2(x)}$
6. (Chain Rule) If $h(x) = f(g(x)) \triangleq (f \circ g)(x)$, then $h'(x) = f'(g(x)) \cdot g'(x)$

Apart from this, some handy rules exist for polynomial functions e.g. if $f(x) = x^c$ where c is not a function of x, then $f'(x) = c \cdot x^{c-1}$, the logarithmic function i.e. if $f(x) = \ln(x)$ then $f'(x) = \frac{1}{x}$, the exponential function i.e. if $f(x) = \exp(x)$ then $f'(x) = \exp(x)$ and trigonometric functions i.e. if $f(x) = \sin(x)$ then $f'(x) = \cos(x)$ and if $f(x) = \cos(x)$ then $f'(x) = -\sin(x)$. The most common use of the chain rule is finding $f'(x)$ when f is a function of some variable, say t but t itself is a function of x i.e. $t = g(x)$.

Example B.1. Let $\ell(x) = (a \cdot x - b)^2$ where $a, b \in \mathbb{R}$ are constants that do not depend on x. Then we can write $\ell(t) = t^2$ where $t(x) = a \cdot x - b$. Thus, applying the chain rule tells us that $\ell'(x) = \ell'(t) \cdot t'(x)$. By applying the rules above we have $\ell'(t) = 2 \cdot t$ (polynomial rule) and $t'(x) = a$ (constant rule and scaling rule). This gives us $\ell'(x) = 2a \cdot (a \cdot x - b)$.

Example B.2. Let $\sigma(x) = \frac{1}{1+\exp(-B \cdot x)}$ be the sigmoid function where $B \in \mathbb{R}$ is a constant that does not depend on x. Then we can write $\sigma(t) = (t)^{-1}$ where $t(s) = 1 + \exp(s)$ where $s(x) = -B \cdot x$. Thus, applying the chain rule tells us that $\sigma'(x) = \sigma'(t) \cdot t'(s) \cdot s'(x)$. By applying the rules above we have $\sigma'(t) = -\frac{1}{t^2}$ (polynomial rule), $t'(s) = \exp(s)$ (constant rule and exponential rule), $s'(x) = -B$ (scaling rule). This gives us $\sigma'(x) = B\frac{\exp(-B \cdot x)}{(1+\exp(-B \cdot x))^2} = B \cdot \sigma(x)(1 - \sigma(x))$

B.6 Multivariate Functions

In the previous sections we looked at functions of one variable i.e. univariate functions $f : \mathbb{R} \to \mathbb{R}$. We will now extend our intuitions about derivatives to multivariate functions i.e. functions of multiple variables i.e. of the form $f : \mathbb{R}^d \to \mathbb{R}$ which take a d-dimensional vector as input and output a real number.

multivariate functions

B.6.1 First Derivatives

As before, the first derivative tells us how much the function value changes and in what direction, if we move a bit from our current location. Since in d dimensions, there are d directions along which we can move, "moving" means going from $\mathbf{x}^0 \in \mathbb{R}^d$ to a point $\mathbf{x}^0 + \Delta\mathbf{x}$ where $\Delta\mathbf{x} \in \mathbb{R}^d$ (but $\Delta\mathbf{x}$ is "small" i.e. $\|\Delta\mathbf{x}\|_2$ is small). To capture how the function value may change as a result of such movement, the gradient of the function captures how much the function changes if we move just long one of the axes.

More specifically, the gradient of a multivariate function f at a point $\mathbf{x}^0 \in \mathbb{R}^d$ is a vector $\nabla f(\mathbf{x}^0) = \left(\frac{\partial f}{\partial \mathbf{x}_1}, \frac{\partial f}{\partial \mathbf{x}_2}, \ldots, \frac{\partial f}{\partial \mathbf{x}_d}\right)$ where for any $j \in [d]$, $\frac{\partial f}{\partial \mathbf{x}_j}$ indicates whether the function value increases or decreases and by how much, if we keep all coordinates of $\mathbf{x}^0$ fixed except the j^{th} coordinate whih we increase by a small amount i.e. if $\Delta\mathbf{x} = (0, 0, \ldots, 0, \underset{j}{\delta}, 0, \ldots, 0)$, then our friend the Taylor's theorem tells us that

gradient

$$f(\mathbf{x}^0 + \Delta\mathbf{x}) \approx f(\mathbf{x}^0) + \delta \cdot \frac{\partial f}{\partial \mathbf{x}_j}$$

We can use the gradient to find out how much the function value would change if we moved a little bit in a general direction by summing up the individual contributions from all the axes. Suppose we move along $\Delta\mathbf{x}$ where now all coordinates of $\Delta\mathbf{x}$ may be non-zero (but small), then the following holds

$$\begin{aligned} f(\mathbf{x}^0 + \Delta\mathbf{x}) &\approx f(\mathbf{x}^0) + \nabla f(\mathbf{x}^0)^\top \Delta\mathbf{x} \\ &= f(\mathbf{x}^0) + \sum_{j=1}^{d} \Delta\mathbf{x}_j \cdot \frac{\partial f}{\partial \mathbf{x}_j} \end{aligned}$$

Multivariate Taylor's theorem (first order)

The gradient also has the very useful property of being the direction of steepest ascent. This means that among all the directions in which we could move, if we move along the direction of the gradient, then the function value would experience the maximum amount of increase. However, for machine learning applications, a related property holds more importance: among all the directions in which we could move, if we move along the direction opposite to that of the gradient i.e. we move along $-\nabla f(\mathbf{x}^0)$, then the function value would experience the maximum amount of *decrease* – this means that the direction opposite to the gradient offers the steepest descent.

steepest ascent

steepest descent

B.6.2 Second Derivatives

Second derivatives play a similar role of documenting how the first derivative changes as we move a little bit from point to point. However, since we have d partial derivatives here and d possible axes directions along which to move, the second derivative for multivariate functions is actually a $d \times d$ matrix, called the Hessian and denoted as $\nabla^2 f(\mathbf{x}^0)$.

Hessian

$$\nabla^2 f(\mathbf{x}^0) = \begin{bmatrix} \frac{\partial^2 f}{\partial \mathbf{x}_1^2} & \frac{\partial^2 f}{\partial \mathbf{x}_1 \partial \mathbf{x}_2} & \cdots & \frac{\partial^2 f}{\partial \mathbf{x}_1 \partial \mathbf{x}_d} \\ \frac{\partial^2 f}{\partial \mathbf{x}_2 \partial \mathbf{x}_1} & \frac{\partial^2 f}{\partial \mathbf{x}_2^2} & \cdots & \frac{\partial^2 f}{\partial \mathbf{x}_2 \partial \mathbf{x}_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial \mathbf{x}_d \partial \mathbf{x}_1} & \frac{\partial^2 f}{\partial \mathbf{x}_d \partial \mathbf{x}_2} & \cdots & \frac{\partial^2 f}{\partial \mathbf{x}_d^2} \end{bmatrix}$$

Clairaut's theorem tells us that if the function f is "nice" (basically the second order partial derivatives are all continuous), then $\frac{\partial^2 f}{\partial \mathbf{x}_i \partial \mathbf{x}_j} = \frac{\partial^2 f}{\partial \mathbf{x}_j \partial \mathbf{x}_i}$ i.e. the Hessian matrix is symmetric. The $(i,j)^{\text{th}}$ entry of this Hessian matrix – styled as $\frac{\partial^2 f}{\partial \mathbf{x}_i \partial \mathbf{x}_j}$ – records how much the i^{th} partial derivative changes if we move a little bit along the j^{th} axis i.e. if $\Delta\mathbf{x} = (0, 0, \ldots, 0, \underset{j}{\delta}, 0, \ldots, 0)$, then

$$\frac{\partial f}{\partial \mathbf{x}_i}(x^0 + \Delta x) \approx \frac{\partial f}{\partial \mathbf{x}_i}(x^0) + \frac{\partial^2 f}{\partial \mathbf{x}_i \partial \mathbf{x}_j}(x^0) \cdot \Delta\mathbf{x}, \text{if } \Delta\mathbf{x} \text{ is "small".}$$

Just as in the univariate case, the Hessian can be incorporated into the Taylor's theorem to obtain a finer approximation of the change in function value. Denote $H = \nabla^2 f(\mathbf{x}^0)$ for sake of notational simplicity

Multivariate Taylor's theorem (second order)

$$\begin{aligned} f(\mathbf{x}^0 + \Delta\mathbf{x}) &\approx f(\mathbf{x}^0) + \nabla f(\mathbf{x}^0)^\top \Delta\mathbf{x} + (\Delta\mathbf{x})^\top H(\Delta\mathbf{x}) \\ &= f(\mathbf{x}^0) + \sum_{j=1}^{d} \Delta\mathbf{x}_j \cdot \frac{\partial f}{\partial \mathbf{x}_j} + \sum_{i=1}^{d}\sum_{j=1}^{d} \Delta\mathbf{x}_i \Delta\mathbf{x}_j \frac{\partial^2 f}{\partial \mathbf{x}_i \partial \mathbf{x}_j}(x^0) \end{aligned}$$

B.6.3 Stationary Points

Just as in the univariate case, here also we define stationary points as those where the gradient of the function vanishes i.e. $\nabla f(\mathbf{x}^0) = \mathbf{0}$. As before, stationary points can either be local minima/maxima or else saddle points and the second derivative test is used to decide which is the case. However, the *multivariate second derivative test* looks a bit different.

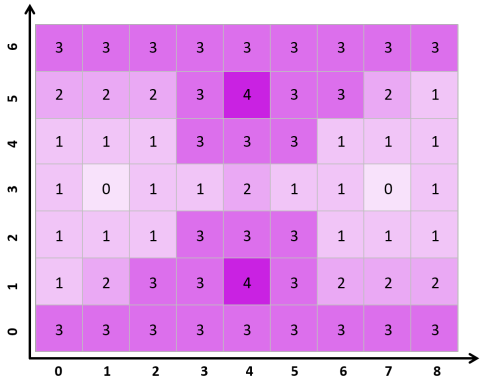
If the Hessian of the function is positive semi definite (PSD) at a stationary point $\mathbf{x}^0$ i.e. $\nabla f(\mathbf{x}^0) = \mathbf{0}$ and $H = \nabla^2 f(\mathbf{x}^0) \succeq 0$ then $\mathbf{x}^0$ is definitely a local minimum. Recall that a square symmetric matrix $A \in \mathbb{R}^{d \times d}$ is called positive semi definite if for all vectors $\mathbf{v} \in \mathbb{R}^d$, we have $\mathbf{v}^\top A \mathbf{v} \geq 0$. As before, this result follows directly from the multivariate second order Taylor's theorem we studied above. Since $\nabla f(\mathbf{x}^0) = \mathbf{0}$, we have

$$f(\mathbf{x}^0 + \Delta\mathbf{x}) \approx f(\mathbf{x}^0) + \frac{1}{2}(\Delta\mathbf{x})^\top H(\Delta\mathbf{x}) \geq f(\mathbf{x}^0)$$

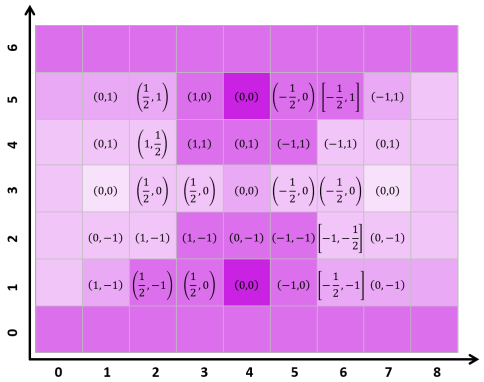
This means that no matter in which direction we move from $\mathbf{x}^0$, the function value always increases. This is the very definition of a local minimum. Similarly, we can intuitively see that if the Hessian of the function is negative semi definite (NSD) at a stationary point $\mathbf{x}^0$ i.e. $\nabla f(\mathbf{x}^0) = \mathbf{0}$ and $\nabla^2 f(\mathbf{x}^0) \preceq 0$ then $\mathbf{x}^0$ is a local maximum. Recall that a square symmetric matrix $A \in \mathbb{R}^{d \times d}$ is called negative semi definite if for all vectors $\mathbf{v} \in \mathbb{R}^d$, we have $\mathbf{v}^\top A \mathbf{v} \leq 0$.

B.7 Visualizing Multivariate Derivatives

We now take a toy example to help the reader visualize how multivariate derivatives operate. We will take $d = 2$ to allow us to explicitly show gradients and function values on a 2D grid. The function we will study will not be continuous but discrete but will nevertheless allow us to revise the essential aspects of the topics we studied above.



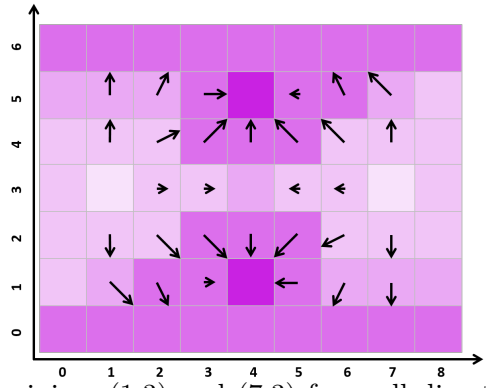
Consider the function $f : [0, 8] \times [0, 6] \rightarrow \mathbb{R}$ on the left. The function is discrete – darker shades indicate a higher function value (which is also written inside the boxes) and lighter shades indicate a smaller function value. Since discrete functions are non-differentiable, we will use approximations to calculate the gradient of this function at all the points. Note that the input to this function are two integers (x, y) where $0 \leq x \leq 8$ and $0 \leq y \leq 6$.



Given this, we may estimate the gradient of the function at a point (x_0, y_0) using the formula $\nabla f(x_0, y_0) = \left(\frac{\Delta f}{\Delta x}, \frac{\Delta f}{\Delta y}\right)$ where

$$\frac{\Delta f}{\Delta x} = \frac{f(x_0 + 1, y_0) - f(x_0 - 1, y_0)}{2}$$
$$\frac{\Delta f}{\Delta y} = \frac{f(x_0, y_0 + 1) - f(x_0, y_0 - 1)}{2}$$

The values of the gradients calculated using the above formula are shown on the left. Notice that we have five locations where the gradient vanishes (4,5), (1,3), (4,3), (7,3) and (4,1): these are stationary points.



It may be more instructive to see the gradients represented as arrows which the figure on the left does. Notice that gradients converge toward the local maxima (4,5) and (4,1) from all directions (this is expected since the point has a greater function value than all its neighbors). Similarly, gradients diverge away from the local minima (1,3) and (7,3) from all directions (this is expected as well since the point has a smaller function value than all its neighbors). However, the point (4,3) being a saddle point, has gradients converging to it in the x direction but diverging away from it in the y direction.

In order to verify which of our stationary points are local maxima/minima and which are saddle points, we need to estimate the Hessian of this function. To do so, we use the following formulae

$$\nabla^2 f(x_0, y_0) = \begin{bmatrix} \frac{\Delta^2 f}{\Delta x^2} & \frac{\Delta^2 f}{\Delta x \Delta y} \\ \frac{\Delta^2 f}{\Delta x \Delta y} & \frac{\Delta^2 f}{\Delta y^2}, \end{bmatrix}$$

where we calculate each of the mixed partial derivative terms as follows

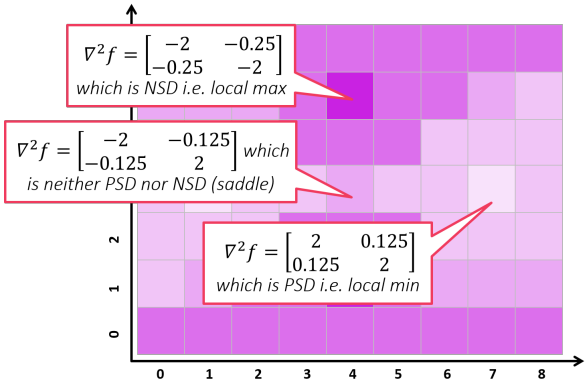
$$\frac{\Delta^2 f}{\Delta x^2} = \frac{f(x_0+1, y_0) + f(x_0-1, y_0) - 2f(x_0, y_0)}{1^2}$$

$$\frac{\Delta^2 f}{\Delta y^2} = \frac{f(x_0, y_0+1) + f(x_0, y_0-1) - 2f(x_0, y_0)}{1^2}$$

$$\frac{\Delta^2 f}{\Delta x \Delta y} = \frac{f_{xy} + f_{yx}}{2}$$

$$f_{xy} = \frac{\frac{\Delta f}{\Delta x}(x_0, y_0+1) - \frac{\Delta f}{\Delta x}(x_0, y_0-1)}{2}$$

$$f_{yx} = \frac{\frac{\Delta f}{\Delta y}(x_0+1, y_0) - \frac{\Delta f}{\Delta y}(x_0-1, y_0)}{2}$$



Deriving the above formulae is relatively simple but we do not do so here. Also, the expression for $\frac{\Delta^2 f}{\Delta x \Delta y}$ was made such so as to obtain a symmetric Hessian since Clairaut's theorem does not apply to our toy example. However, having done so, we can verify that the Hessian is indeed PSD at the local minima, NSD at the local maxima and neither NSD nor PSD at the saddle point. This verifies our earlier second derivative test rules.

B.8 Useful Rules for Calculating Gradients

The rules that we studied in the context of univariate functions (Constant Rule, Sum Rule, Scaling Rule, Product Rule, Quotient Rule, Chain Rule) continue

to apply in the multivariate setting as well. However, we present here a few more handy rules for calculating gradients. In the following, $\mathbf{x} \in \mathbb{R}^d$

1. (Dot Product Rule) If $f(\mathbf{x}) = \mathbf{a}^\top \mathbf{x}$ where $\mathbf{a} \in \mathbb{R}^d$ is a vector that does not depend on $\mathbf{x}$, then $\nabla f(\mathbf{x}) = \mathbf{a}$. This is a generalization of the scaling rule in the univariate case.

2. (Quadratic Rule) If $f(\mathbf{x}) = \mathbf{x}^\top A\mathbf{x}$ where $A \in \mathbb{R}^{d\times d}$ is a symmetric matrix that is not a function of $\mathbf{x}$, then $\nabla f(\mathbf{x}) = 2A\mathbf{x}$. If A is not symmetric, then $\nabla f(\mathbf{x}) = A\mathbf{x} + A^\top \mathbf{x}$.

3. (Chain Rule) If $g : \mathbb{R}^d \to \mathbb{R}$ and $f : \mathbb{R} \to \mathbb{R}$, then if we define $h(\mathbf{x}) = f(g(\mathbf{x}))$, then $\nabla h(\mathbf{x}) = f'(g(\mathbf{x})) \cdot \nabla g(\mathbf{x})$.

A very handy rule to remember while taking derivatives with respect to vectors (i.e. gradients) or even matrices is the dimensionality rule which states that the dimensionality of the derivative must be the same as that of the variable with respect to which the derivative is being taken. Thus, if $\mathbf{x} \in \mathbb{R}^d$ and $f : \mathbb{R}^d \to \mathbb{R}$ then $\frac{df}{d\mathbf{x}}$ must also be a vector of d-dimensions. Similarly, if $X \in \mathbb{R}^{d\times d}$ and $f : \mathbb{R}^{d\times d} \to \mathbb{R}$, then $\frac{df}{dX}$ must also be a $d \times d$ matrix. We now illustrate the use of these rules using some examples

dimensionality rule

Example B.3. Let $f(x) = \|\mathbf{x}\|_2$. We can rewrite this as $f = \sqrt{t}$, where $t(\mathbf{x}) = \|\mathbf{x}\|_2^2 = \mathbf{x}^\top \mathbf{x} = \mathbf{x}^\top I_d \mathbf{x}$ where I_d is the $d \times d$ identity matrix. Thus, using the chain rule we have $\nabla f(\mathbf{x}) = f'(t) \cdot \nabla t(\mathbf{x})$. Using the polynomial rule we have $f'(t) = \frac{1}{2\sqrt{t}}$, whereas using the quadratic rule, we get $\nabla t(\mathbf{x}) = 2\mathbf{x}$. Thus we have $\nabla f(\mathbf{x}) = \frac{\mathbf{x}}{\|\mathbf{x}\|_2}$. Note that in this case, the gradient is always a unit vector.

Example B.4. Let $\sigma(\mathbf{x}) = \frac{1}{1+\exp(-\mathbf{a}^\top \mathbf{x})}$ where $\mathbf{a} \in \mathbb{R}^d$ is a constant vector that does not depend on $\mathbf{x}$. Then we can write $\sigma(t) = (t)^{-1}$ where $t(s) = 1 + \exp(s)$ where $s(x) = -\mathbf{a}^\top \mathbf{x}$. Thus, applying the chain rule tells us that $\nabla\sigma(\mathbf{x}) = \sigma'(t) \cdot t'(s) \cdot \nabla s(\mathbf{x})$. By applying the rules above we have $\sigma'(t) = -\frac{1}{t^2}$ (polynomial rule), $t'(s) = \exp(s)$ (constant rule and exponential rule), $\nabla s(\mathbf{x}) = -\mathbf{a}$ (dot product rule). This gives us $\nabla\sigma(\mathbf{x}) = \frac{\exp(-\mathbf{a}^\top \mathbf{x})}{(1+\exp(-\mathbf{a}^\top \mathbf{x}))^2} \cdot \mathbf{a} = \sigma(\mathbf{x})(1 - \sigma(\mathbf{x})) \cdot \mathbf{a}$.

B.9 Useful Rules for Calculating Hessians

The Hessian of function $f : \mathbb{R}^d \to \mathbb{R}$ can be shown to be defined as the derivative $\nabla^2 f(\mathbf{x}) = \frac{\partial^2 f}{\partial \mathbf{x} \partial \mathbf{x}^\top} = \frac{\partial g}{\partial \mathbf{x}^\top}$ where $g(\mathbf{x}) = \frac{\partial f}{\partial \mathbf{x}} = \nabla f(\mathbf{x})$. Note that the Hessian of a multivariate function is always a square symmetric matrix and indeed, once we take the derivative of $g(\mathbf{x})$ (which is itself a vector) with respect to $\mathbf{x}^\top$, we will get a matrix.

The usual rules of sum, product, scaling, chain continue to apply to calculations of Hessians as well. However, we present below some of the handy rules that make life easier when calculating Hessians.

1. (Linear Rule) If $f(\mathbf{x}) = \mathbf{a}^\top \mathbf{x}$ where $\mathbf{a}$ is a constant vector, then $g(\mathbf{x}) = \nabla f(\mathbf{x}) = \mathbf{a}$ which is a constant and thus $\nabla^2 f(\mathbf{x}) = \frac{\partial g}{\partial \mathbf{x}^\top} = \mathbf{0}\mathbf{0}^\top \in \mathbb{R}^{d\times d}$

using the constant rule. Thus, linear (or even affine functions of the form $f(\mathbf{x}) = \mathbf{a}^\top \mathbf{x} + b$ where b is a scalar constant) have zero Hessians.

2. (Quadratic Rule) If $f(\mathbf{x}) = \mathbf{x}^\top A\mathbf{x}$ where $A \in \mathbb{R}^{d\times d}$ is a symmetric matrix that is not a function of $\mathbf{x}$, then $g(\mathbf{x})\nabla f(\mathbf{x}) = 2A\mathbf{x}$ and $\nabla^2 f(\mathbf{x}) = \frac{\partial g}{\partial \mathbf{x}^\top} = 2A \in \mathbb{R}^{d\times d}$. If A is not symmetric, then $\nabla^2 f(\mathbf{x}) = A + A^\top$.

B.10 Exercises

Exercise B.1. Let $f(x) = x^4 - 4x^2 + 4$. Find all stationary points of this function. Which of them are local maxima and minima?

Exercise B.2. Let $g : \mathbb{R}^2 \to \mathbb{R}$ be defined as $g(x, y) = f(x) + f(y) + 8$. Find all stationary points of this function. Which of them are local maxima and minima? Which one of these are saddle points?

Exercise B.3. Let $\mathbf{x}, \mathbf{a} \in \mathbb{R}^d, b \in \mathbb{R}$ where $\mathbf{a}$ is a constant vector that does not depend on $\mathbf{x}$ and b is a constant real number that does not depend on $\mathbf{x}$. Let $f(\mathbf{x}) = (\mathbf{a}^\top \mathbf{x} - b)^2$. Calculate $\nabla f(\mathbf{x})$.

Exercise B.4. Now suppose we have n such constant vectors $\mathbf{a}^1, \ldots, \mathbf{a}^n$ and n such real constants $b_1, \ldots, b_n$. Let $f(\mathbf{x}) = \sum_{i=1}^n (\mathbf{x}^\top \mathbf{a}^i - b_i)^2$. Calculate $\nabla f(\mathbf{x})$.

Exercise B.5. Given a natural number $n \in \mathbb{N}$ e.g. $2, 8, 97$ and a real number $x^0 \in \mathbb{R}$, design a function $f : \mathbb{R} \to \mathbb{R}$ so that $f^{(k)}(x^0) = 0$ for all $k = 1, 2, \ldots, n$. Here $f^{(k)}(x^0)$ denotes the k^{th} order derivative of f at x^0 e.g. $f^{(1)}(x^0) = f'(x^0), f^{(3)}(x^0) = f'''(x^0)$ etc.

Exercise B.6. Let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$ be constant vectors and let $f(\mathbf{x}) = \mathbf{a}^\top \mathbf{x}\mathbf{x}^\top \mathbf{b}$. Calculate $\nabla f(\mathbf{x})$.
Hint: write $f(\mathbf{x}) = g(\mathbf{x}) \cdot h(\mathbf{x})$ *where* $g(\mathbf{x}) = \mathbf{a}^\top \mathbf{x}$ *and* $h(\mathbf{x}) = \mathbf{b}^\top \mathbf{x}$ *and apply the product rule.*

Exercise B.7. Let $\mathbf{b} \in \mathbb{R}^d$ a constant vector and $A \in \mathbb{R}^{d\times d}$ be a constant symmetric matrix. Let $f(\mathbf{x}) = \mathbf{b}^\top A\mathbf{x}$. Calculate $\nabla f(\mathbf{x})$.
Hint: write $f(\mathbf{x}) = \mathbf{c}^\top \mathbf{x}$ *where* $\mathbf{c} = A^\top \mathbf{b}$.

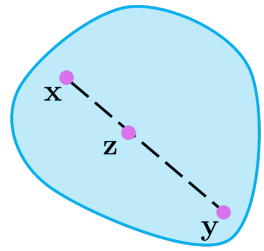
Exercise B.8. Let $A, B, C \in \mathbb{R}^{d\times d}$ be three symmetric and constant matrices and $\mathbf{p}, \mathbf{q} \in \mathbb{R}^d$ be two constant vectors. Let $f(\mathbf{x}) = (A\mathbf{x} + \mathbf{p})^\top C(B\mathbf{x} + \mathbf{q})$. Calculate $\nabla f(\mathbf{x})$.

C

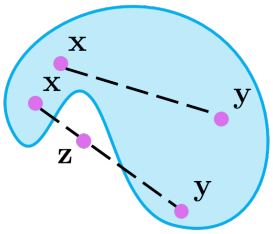
Convex Analysis Refresher

Convex sets and functions remain the favorites of practitioners working on machine learning algorithms since these objects have several beautiful properties that make it simple to design efficient algorithms. Of course, the recent years have seen several strides in *non-convex optimization* as well due to areas such as deep learning, robust learning, sparse learning gaining prominence.

C.1 Convex Set



CONVEX SET



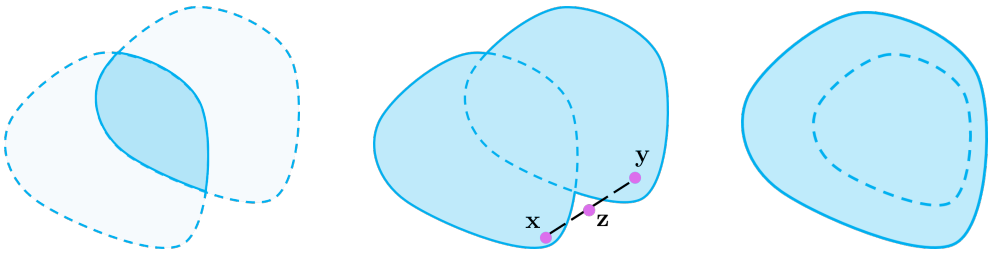
NON-CONVEX SET

Given a set of points (or a region) $\mathcal{C} \subset \mathbb{R}^d$, we call this set or region a convex set if the set contains all line segments that join two points inside that set. More formally, for a set $\mathcal{C}$ to be convex, no matter which two points we take in the set $\mathbf{x}, \mathbf{y} \in \mathcal{C}$, for every $\lambda \in [0, 1]$, we must have $\mathbf{z} \in \mathcal{C}$ where $\mathbf{z} = \lambda \cdot \mathbf{x} + (1 - \lambda) \cdot \mathbf{y}$. It is noteworthy that the vectors $\mathbf{z}$ defined this way completely capture all points on the *line segment* joining $\mathbf{x}$ and $\mathbf{y}$. Indeed, with $\lambda = 0$, we have $\mathbf{z} = \mathbf{z}$, $\lambda = 1$ gives us $\mathbf{z} = \mathbf{y}$ and $\lambda = 0.5$ gives us the midpoint of the line segment. It is noteworthy however, that we must have $\lambda \in [0, 1]$. If λ starts taking negative values or values greater than 1, then we would start getting points outside the line segment.

convex set

For well behaved sets, in order to confirm convexity, it is sufficient to verify that $\mathbf{z} = \frac{\mathbf{x}+\mathbf{y}}{2} \in \mathcal{C}$ i.e. we need not take the trouble of verifying for all $\lambda \in [0, 1]$ and simply verifying mid-point convexity is enough to verify convexity. It is also important to note that non-convex sets, such as the one depicted in the figure, may contain *some* of the line segments that join points within them – this does not make the set convex! Only if a set contains *all* its line segments is it called convex.

mid-point convexity

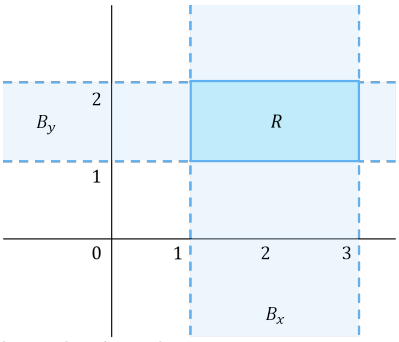


The reader would have noticed that convex sets *bulge* outwards in all directions. The presence of any inward bulges typically makes a set non-convex.

Theorem C.1. Given two convex sets $\mathcal{C}_1, \mathcal{C}_2 \subset \mathbb{R}^d$, the intersection of these two sets i.e. $\mathcal{C}_1 \cap \mathcal{C}_2$ is always convex. However, the union of these two sets i.e. $\mathcal{C}_1 \cup \mathcal{C}_2$ need not be convex.

Proof. We first deal with the case of intersection. The intersection of two sets (not necessarily convex) is defined to be the set of all points that are contained in both the sets i.e. $\mathcal{C}_1 \cap \mathcal{C}_2 \triangleq \left\{\mathbf{x} \in \mathbb{R}^d : \mathbf{x} \in \mathcal{C}_1 \text{ and } \mathbf{x} \in \mathcal{C}_2\right\}$. Consider two points $\mathbf{x}, \mathbf{y} \in \mathcal{C}_1 \cap \mathcal{C}_2$. Since $\mathbf{x}, \mathbf{y} \in \mathcal{C}_1$, we know that $\mathbf{z} = \frac{\mathbf{x}+\mathbf{y}}{2} \in \mathcal{C}_1$ since $\mathcal{C}_1$ is convex. However, by the same argument, we get that $\mathbf{z} = \frac{\mathbf{x}+\mathbf{y}}{2} \in \mathcal{C}_2$ as well. Since $\mathbf{z} \in \mathcal{C}_1$ and $\mathbf{z} \in \mathcal{C}_2$, we conclude that $\mathbf{z} \in \mathcal{C}_1 \cap \mathcal{C}_2$. This proves that the intersection of any two convex sets must necessarily be convex. The first figure above illustrates the intersection region of two convex sets. intersection of two sets

The union of two sets (not necessarily convex) is defined to be the set of all points that are contained in either of the sets (including points that are present in both sets). More specifically, we define $\mathcal{C}_1 \cap \mathcal{C}_2 \triangleq \left\{\mathbf{x} \in \mathbb{R}^d : \mathbf{x} \in \mathcal{C}_1 \text{ or } \mathbf{x} \in \mathcal{C}_2\right\}$. The second figure above shows that the union of two convex sets may be non-convex. However, the union of two convex sets may be convex in some very special cases, for example, if one set is contained in the other i.e. $\mathcal{C}_1 \subseteq \mathcal{C}_2$ which is illustrated in the third figure. □ union of two sets



Example C.1. Are rectangles convex? Let $R \triangleq \{(x, y) \in \mathbb{R}^2 : x \in [1, 3] \text{ and } y \in [1, 2]\}$ be a rectangle with side lengths 1 and 2. We could show R to be convex directly as we do in the example below. However, there exists a neater way. Consider the bands $B_x \triangleq \{(x, y) \in \mathbb{R}^2 : x \in [1, 3]\}$ and $B_y \triangleq \{(x, y) \in \mathbb{R}^2 : y \in [1, 2]\}$. It is easy to see that $R = B_x \cap B_y$. Thus, if we show that the bands are convex, we could then use Theorem C.1 to show that R is convex too! Showing that B_x is convex is pretty easy: if two points have their x coordinate in the range $[1, 3]$, then the average of those two points clearly satisfies this as well. This establishes that B_x is convex. Similarly B_y is convex which tells us that R is convex.

Example C.2. Consider the set of all points which are at a Euclidean distance at most 1 from the origin i.e. the unit ball $\mathcal{B}_2(1) \triangleq \left\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 \leq 1\right\}$. To

show that this set is convex, we take $\mathbf{x}, \mathbf{y} \in \mathcal{B}_2(1)$ and consider $\mathbf{z} = \frac{\mathbf{x}+\mathbf{y}}{2}$. Now, instead of showing $\|\mathbf{z}\|_2 \leq 1$ (which will establish convexity), we will instead show $\|\mathbf{z}\|_2^2 \leq 1$ which is equivalent but easier to analyze. We have $\|\mathbf{z}\|_2^2 = \left\|\frac{\mathbf{x}+\mathbf{y}}{2}\right\|_2^2 = \frac{\|\mathbf{x}\|_2^2+\|\mathbf{y}\|_2^2+2\mathbf{x}^\top\mathbf{y}}{4}$. Now, the Cauchy-Schwartz inequality tells us that for any two vectors $\mathbf{a}, \mathbf{b}$ we have $\left|\mathbf{a}^\top\mathbf{b}\right| \leq \|\mathbf{a}\|_2 \|\mathbf{b}\|_2$. Thus, we have $\|\mathbf{z}\|_2^2 \leq \frac{\|\mathbf{x}\|_2^2+\|\mathbf{y}\|_2^2+2\|\mathbf{x}\|_2\|\mathbf{y}\|_2}{4}$. Since $\mathbf{x}, \mathbf{y} \in \mathcal{B}_2(1)$, we have $\|\mathbf{x}\|_2, \|\mathbf{y}\|_2 \leq 1$ which gives us $\|\mathbf{z}\|_2^2 \leq 1$ which establishes the unit ball $\mathcal{B}_2(1)$ is a convex set.

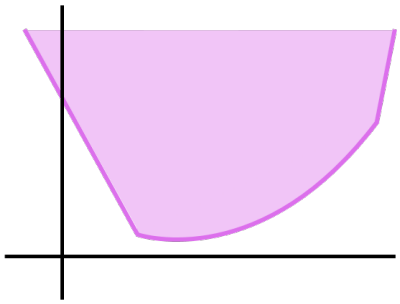
Cauchy-Schwartz inequality

C.2 Convex Functions

We now move on to convex functions. These functions play an important role in several optimization based machine learning algorithms such as SVMs and logistic regression. There exist several definitions of convex functions, some that apply only to differentiable functions, and others that apply even to non-differentiable functions. We look at these below.

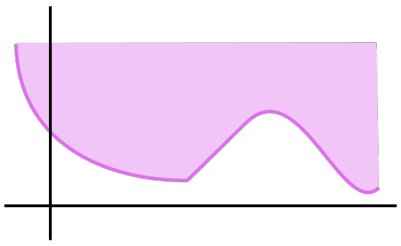
C.2.1 Epigraph Convexity

This is the most fundamental definition of convexity and applies to all functions, whether they are differentiable or not. This definition is also quite neat in that it simply uses the definition of convex sets to define convex functions.



Definition C.1 (Epigraph). Given a function $f : \mathbb{R}^d \to \mathbb{R}$, its epigraph is defined as the set of points that lie on or above the graph of the function i.e. $\text{epi}(f) \triangleq \left\{(\mathbf{x}, y) \in \mathbb{R}^{d+1} : y \geq f(\mathbf{x})\right\}$. Note that the epigraph is a $d+1$-dimensional set and not a d dimensional set.

Epigraph

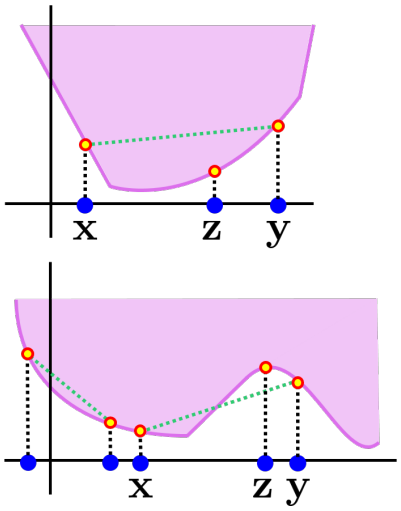


Definition C.2 (Epigraph Convexity). A function $f : \mathbb{R}^d \to \mathbb{R}$ is defined to be convex if its epigraph $\text{epi}(f) \in \mathbb{R}^{d+1}$ is a convex set. On the left, we have a non-convex function whose epigraph is a non-convex set (notice the inward bulge) whereas in the figure above, we have a convex function whose epigraph is a convex set.

Epigraph Convexity

C.2.2 Chord Convexity

The above definition, although fundamental, is not used quite often since there exist simpler definitions. One of these definitions exploits the fact that convexity of the epigraph set need only be verified at the lower boundary of the set i.e. at the surface of the function graph. Applying the mid-point definition of convex sets then gives us this new definition of convex functions.



Definition C.3 (Chord). Given a function $f : \mathbb{R}^d \to \mathbb{R}$ and any two points $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, the line segment joining the two points $(\mathbf{x}, f(\mathbf{x}))$ and $(\mathbf{y}, f(\mathbf{y}))$ is called a chord of this function. Chord

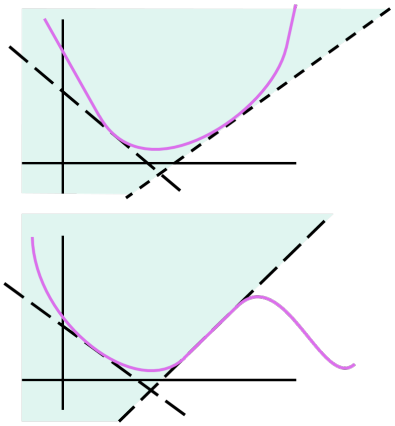
Definition C.4 (Chord Convexity). A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if and only if the function graph lies below all its chords i.e. for any two points $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ we must have $f\left(\frac{\mathbf{x}+\mathbf{y}}{2}\right) \leq \frac{f(\mathbf{x})+f(\mathbf{y})}{2}$. Chord Convexity

The figures depict a convex function that lies above all its chords and a non-convex function which does not do so.

It is also important to note that non-convex functions may lie below *some* chords (as the figure on the bottom shows) – this does not make the function convex! Only if a functions lies below *all* its chords is it called convex.

C.2.3 Tangent Convexity

This definition holds true only for differentiable functions but is usually easier to apply when checking whether a function is convex or not.



Definition C.5 (Tangent). Given a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, the tangent of the function at a point $\mathbf{x}^0 \in \mathbb{R}^d$ is the hyperplane $\nabla f(\mathbf{x}^0)^\top(\mathbf{x} - \mathbf{x}^0) + f(\mathbf{x}^0) = 0$ i.e. of the form $\mathbf{w}^\top\mathbf{x} + b$ where $\mathbf{w} = \nabla f(\mathbf{x}^0)$ and $b = f(\mathbf{x}^0) - \nabla f(\mathbf{x}^0)^\top\mathbf{x}^0$. Tangent

Definition C.6 (Tangent Convexity). A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if and only if the function graph lies above all its tangents i.e. for all $\mathbf{x}^0, \mathbf{x} \in \mathbb{R}^d$, we have $f(\mathbf{x}) \geq \nabla f(\mathbf{x}^0)^\top(\mathbf{x} - \mathbf{x}^0) + f(\mathbf{x}^0)$. Tangent Convexity

Note that the point $(\mathbf{x}^0, f(\mathbf{x}^0))$ always lies on the tangent hyperplane at $\mathbf{x}^0$. The figures above depict a convex function that lies above all its tangents and a non-convex function which fails to lie above at least one of its tangents. It is important to note that non-convex functions may lie above *some* of their tangenst (as the figure on the bottom shows) – this does not make the function convex! Only if a functions lies above *all* its tangents is it called convex.

It is also useful to clarify that the epigraph and chord definitions of convexity continue to apply here as well. Its is just that the tangent definition is easier to use in several cases. A rough analogy is that of deciding the income of individuals – although we can find out the total income of any citizen of India, it may be tedious to do so. However, it is much easier to find the income of a person if that person files income tax returns (truthfully, of course).

C.2.4 Hessian Convexity

Hessian Convexity

For doubly differentiable functions, we have an even simpler definition of convexity. A doubly differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if and only if its Hessian is positive semi-definite at all points i.e. $\nabla^2 f(\mathbf{x}^0) \succeq 0$ for all $\mathbf{x}^0 \in \mathbb{R}^d$. Recall that this implies that for all $\mathbf{v} \in \mathbb{R}^d$, we have $\mathbf{v}^\top \nabla^2 f(\mathbf{x}^0)\mathbf{v} \geq 0$.

C.2.5 Concave Functions

Concave functions are defined as those whose negative is a convex function i.e. $f : \mathbb{R}^d \to \mathbb{R}$ is defined to be concave if the function $-f$ is convex. Convex functions typically look like upturned cups (think of the function $f(x) = x^2$ which is convex and looks like a right-side-up cup). Concave functions on the other hand look like inverted cups, for example $f(x) = -x^2$. To check whether a function is concave or not, we need to simply check (using the epigraph, chord, tangent, or Hessian methods) whether the negative of that function is convex or not.

Example C.3. Let us look at the example of the Euclidean norm $f(\mathbf{x}) = \|\mathbf{x}\|_2 = \sqrt{\mathbf{x}^\top \mathbf{x}}$. This function is non-differentiable at the origin i.e. at $\mathbf{x} = \mathbf{0}$ so we have to use the chord definition of convexity. Given two points $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we have $f\left(\frac{\mathbf{x}+vy}{2}\right) = \left\|\frac{\mathbf{x}+\mathbf{y}}{2}\right\| = \frac{1}{2}\|\mathbf{x}+\mathbf{y}\|$ by using the homogeneity property of the Euclidean distance (if we halve a vector, its length gets halved too). However, the triangle inequality tells us that for any two vectors $\mathbf{p}, \mathbf{q}$, we have $\|\mathbf{p}+\mathbf{q}\|_2 \leq \|\mathbf{p}\|_2 + \|\mathbf{q}\|_2$. This gives us $f\left(\frac{\mathbf{x}+\mathbf{y}}{2}\right) \leq \frac{\|\mathbf{x}\|_2+\|\mathbf{y}\|_2}{2} = \frac{f(\mathbf{x})+f(\mathbf{y})}{2}$ which proves the convexity of the norm.

triangle inequality

C.3 Operations with Convex Functions

We can take convex functions and manipulate them to obtain new convex functions. Here we explore some such operations that are useful in machine learning applications.

1. Affine functions are always convex[1].
2. Scaling a convex function by a positive scale factor always yields a convex function[2].
3. The sum of two convex functions is always convex (see Theorem C.2).
4. If $g : \mathbb{R}^d \to \mathbb{R}$ is a (multivariate) convex function and $f : \mathbb{R} \to \mathbb{R}$ is a (univariate) convex and non-decreasing function i.e. $a \leq b \Leftrightarrow f(a) \leq f(b)$, then the function $h \triangleq f \circ g$ i.e. $h(\mathbf{x}) = f(g(\mathbf{x}))$ is also convex (see Theorem C.3).
5. If $f : \mathbb{R} \to \mathbb{R}$ is a (univariate) convex function then for any $\mathbf{a} \in \mathbb{R}^d, b \in \mathbb{R}$, the (multivariate) function $g : \mathbb{R}^d \to \mathbb{R}$ defined as $g(\mathbf{x}) = f(\mathbf{a}^\top \mathbf{x} + b)$ is always convex (see Theorem C.4).

[1] See Exercise C.6.
[2] See Exercise C.7.

6. If $f, g : \mathbb{R}^d \to \mathbb{R}$ are two (multivariate) convex functions then the function $h \triangleq \max\{f, g\}$ is also convex (see Theorem C.5).

Theorem C.2 (Sum of Convex Functions)**.** Given two convex functions $f, g : \mathbb{R}^d \to \mathbb{R}$, the function $h \triangleq f + g$ is always convex.

Proof. We will use the chord definition of convexity here since there is no surety that f and g are differentiable. Consider two points $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. We have

$$\begin{aligned} h\left(\frac{\mathbf{x}+\mathbf{y}}{2}\right) &= f\left(\frac{\mathbf{x}+\mathbf{y}}{2}\right) + g\left(\frac{\mathbf{x}+\mathbf{y}}{2}\right) \\ &\leq \frac{f(\mathbf{x})+f(\mathbf{y})}{2} + \frac{g(\mathbf{x})+g(\mathbf{y})}{2} \\ &= \frac{(f(\mathbf{x})+g(\mathbf{x}))+(f(\mathbf{y})+g(\mathbf{y}))}{2} \\ &= \frac{h(\mathbf{x})+h(\mathbf{y})}{2} \end{aligned}$$

where in the second step, we used the fact that f and g are both convex. This proves that h is convex by the chord definition of convexity. □

Theorem C.3 (Composition of Convex Functions)**.** Suppose $g : \mathbb{R}^d \to \mathbb{R}$ is a (multivariate) convex function and $f : \mathbb{R} \to \mathbb{R}$ is a (univariate) convex and non-decreasing function i.e. $a \leq b \Leftrightarrow f(a) \leq f(b)$, then the function $h \triangleq f \circ g$ i.e. $h(\mathbf{x}) = f(g(\mathbf{x}))$ is always convex.

Proof. We will use the chord definition of convexity here since there is no surety that f and g are differentiable. Consider two points $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. We have

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

Now, since g is convex, we have

$$g\left(\frac{\mathbf{x}+\mathbf{y}}{2}\right) \leq \frac{g(\mathbf{x})+g(\mathbf{y})}{2}$$

Let us denote the left hand side of the above inequality by p and the right hand side by q for sake of notational simplicity. Thus, the above inquality tells us that $p \leq q$. However, since f is non-decreasing, we get $f(p) \leq f(q)$ i.e.

$$f\left(g\left(\frac{\mathbf{x}+\mathbf{y}}{2}\right)\right) \leq f\left(\frac{g(\mathbf{x})+g(\mathbf{y})}{2}\right)$$

Let us denote $u \triangleq g(\mathbf{x})$ and $v \triangleq g(\mathbf{y})$ for sake of notational simplicity. Since f is convex, we have

$$f\left(\frac{u+v}{2}\right) \leq \frac{f(u)+f(v)}{2}$$

This is the same as saying

$$f\left(\frac{g(\mathbf{x})+g(\mathbf{y})}{2}\right) \leq \frac{f(g(\mathbf{x}))+f(g(\mathbf{y}))}{2} = \frac{h(\mathbf{x})+h(\mathbf{y})}{2}$$

Thus, n the chain of inequalities established above, we have shown that

$$h\left(\frac{\mathbf{x}+\mathbf{y}}{2}\right) \leq \frac{h(\mathbf{x})+h(\mathbf{y})}{2}$$

which proves that h is a convex function. □

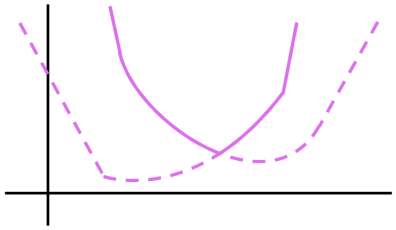
Theorem C.4 (Convex Wrappers over Affine Functions)**.** If $f : \mathbb{R} \to \mathbb{R}$ is a (univariate) convex function then for any $\mathbf{a} \in \mathbb{R}^d, b \in \mathbb{R}$, the (multivariate) function $g : \mathbb{R}^d \to \mathbb{R}$ defined as $g(\mathbf{x}) = f(\mathbf{a}^\top \mathbf{x} + b)$ is always convex.

Proof. We will yet again use the chord definition of convexity. Consider two points $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. We have

$$\begin{aligned} g\left(\frac{\mathbf{x}+\mathbf{y}}{2}\right) &= f\left(\mathbf{a}^\top\left(\frac{\mathbf{x}+\mathbf{y}}{2}\right) + b\right) \\ &= f\left(\frac{(\mathbf{a}^\top\mathbf{x}+b)+(\mathbf{a}^\top\mathbf{y}+b)}{2}\right) \\ &\leq \frac{f(\mathbf{a}^\top\mathbf{x}+b)+f(\mathbf{a}^\top\mathbf{y}+b)}{2} \\ &= \frac{g(\mathbf{x})+g(\mathbf{y})}{2} \end{aligned}$$

where in the second step we used the linearity of the dot product i.e. $\mathbf{c}^\top(\mathbf{a}+\mathbf{b}) = \mathbf{c}^\top\mathbf{a}+\mathbf{c}^\top\mathbf{b}$ and in the third step we used convexity of f. This shows that the function g is convex. □

Theorem C.5 (Maximum of Convex Functions)**.** If $f, g : \mathbb{R}^d \to \mathbb{R}$ are two (multivariate) convex functions then the function $h \triangleq \max\{f, g\}$ is also convex.



Proof. To prove this result, we will need the following simple monotonicity property of the max function: Let $a, b, c, d \in \mathbb{R}$ be four real numbers such that $a \leq c$ and $b \leq d$. Then we must have $\max\{a, b\} \leq \max\{c, d\}$. This can be shown in a stepwise manner ($\max\{a, b\} \leq \max\{c, b\} \leq \max\{c, d\}$) Now, consider two points $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. We have

$$\begin{aligned} &h\left(\frac{\mathbf{x}+\mathbf{y}}{2}\right) = \max\left(f\left(\frac{\mathbf{x}+\mathbf{y}}{2}\right), g\left(\frac{\mathbf{x}+\mathbf{y}}{2}\right)\right) \\ &\leq \max\left(\frac{f(\mathbf{x})+f(\mathbf{y})}{2}, \frac{g(\mathbf{x})+g(\mathbf{y})}{2}\right) \\ &\leq \max\left(\frac{\max\{f(\mathbf{x}), g(\mathbf{x})\}+\max\{f(\mathbf{y}), g(\mathbf{y})\}}{2}, \frac{g(\mathbf{x})+g(\mathbf{y})}{2}\right) \\ &\leq \max\left(\frac{\max\{f(\mathbf{x}), g(\mathbf{x})\}+\max\{f(\mathbf{y}), g(\mathbf{y})\}}{2}, \frac{\max\{f(\mathbf{x}), g(\mathbf{x})\}+\max\{f(\mathbf{y}), g(\mathbf{y})\}}{2}\right) \\ &= \frac{\max\{f(\mathbf{x}), g(\mathbf{x})\}+\max\{f(\mathbf{y}), g(\mathbf{y})\}}{2} = \frac{h(\mathbf{x})+h(\mathbf{y})}{2} \end{aligned}$$

where in the second step, we used the fact that f, g are convex functions and the monotonicity property of the max function. The third and the fourth steps also use the monotonicity property. The fifth step uses the fact that $\max\{a, a\} = a$. This proves that h is a convex function.

Example C.4. The functions $f(x) = \ln(x)$ and $g(x) = \sqrt{x}$ are concave. Since both of these are doubly differentiable functions, we may use the Hessian

definition to decide their concavity. Recall that a function is concave if and only if its negation is convex. Let $p(x) = -\ln(x)$. Then $p''(x) = \frac{1}{x^2} \geq 0$ for all $x > 0$. This confirms that $p(x)$ is convex and that $\ln(x)$ is concave. Similarly, define $q(x) = -\sqrt{x}$. Then $q''(x) = \frac{1}{4x\sqrt{x}} \geq 0$ for all $x \geq 0$ which confirms that $q(x)$ is convex and that $\sqrt{x}$ is concave.

Example C.5. Let us show that squared Euclidean norm i.e. the function $h(\mathbf{x}) = \|\mathbf{x}\|_2^2$ is convex. We have already shown above that the function $g(\mathbf{x}) = \|\mathbf{x}\|_2$ is convex. We can write $h(\mathbf{x}) = f(g(\mathbf{x}))$ where $f(t) = t^2$. Now, $f''(t) = 1$ i.e. f is convex. Also, $\|\mathbf{x}\|_2 \geq 0$ for all $\mathbf{x} \in \mathbb{R}^d$ and the function f is indeed an increasing function on the positive half of the real line. Thus, Theorem C.3 tells us that $h(\mathbf{x})$ is convex as well.

Example C.6. Let us show that the hinge loss is a convex function $\ell_{\text{hinge}}(t) = \max\{1 - t, 0\}$. Note that the hinge loss function is treated as a univariate function here i.e. $\ell_{\text{hinge}} : \mathbb{R} \to \mathbb{R}$. Exercise C.6 shows us that affine functions are convex. Thus $f(x) = 1 - x$ and $g(x) = 0$ are both convex functions. Thus, by applying Theorem C.5, we conclude that the hinge loss function is convex.

Example C.7. We will now show that the objective function used in the C-SVM formulation is a convex function of the model vector $\mathbf{w}$. For sake of simplicity, we will show this result without the bias parameter b although we stress that the result holds even if the bias parameter is present (recall that the bias can always be hidden inside the model by adding a fake dimension into the data). Let $\{(\mathbf{x}^i, y^i)\}_{i=1}^n$ be n data points with $\mathbf{x}^i \in \mathbb{R}^d$ and $y^i \in \{-1, 1\}$. Denote $\mathbf{z}^i \triangleq y^i \cdot \mathbf{x}^i$ for sake of notational simplicity. The the C-SVM objective function is reproduced below:

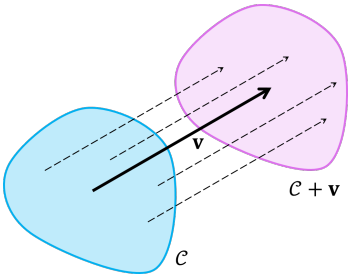
$$f_{\text{C-SVM}}(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^{n} \ell_{\text{hinge}}(y^i \cdot \mathbf{w}^\top \mathbf{x}^i) = \frac{1}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^{n} \ell_{\text{hinge}}(\mathbf{w}^\top \mathbf{z}^i)$$

Note that the feature vectors $\mathbf{x}^i, i = 1, \ldots, n$ and the labels $y^i, i = i, \ldots, n$ (and hence the vectors $\mathbf{z}^i$) are treated as constants since we cannot change our training data. The only variable here is the model vector $\mathbf{w}$ which we learn using the training data. We have already shown that $\|\mathbf{w}\|_2^2$ is a convex function of $\mathbf{w}$, Exercise C.7 shows that $\frac{1}{2}\|\mathbf{w}\|_2^2$ is convex too. We showed above that ℓ_{hinge} is a convex function and thus, Theorem C.4 shows that $h_i(\mathbf{w}) = \ell_{\text{hinge}}(\mathbf{w}^\top \mathbf{z}^i)$ is a convex function of $\mathbf{w}$ for every i. Theorem C.2 shows that the sum of convex functions is convex which shows that $f_{\text{C-SVM}}(\mathbf{w})$ is a convex function.

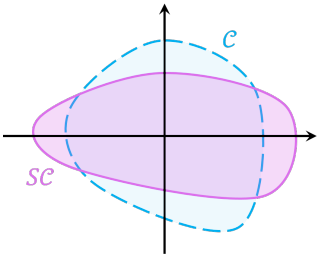
C.4 Exercises

Exercise C.1. Let A be a positive semi-definite matrix and let us define a Mahalanobis distance using A as $d_A(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^\top A(\mathbf{x} - \mathbf{y})}$. Consider the unit ball according to this distance i.e. the set of all points that are less than or equal to unit Mahalanobis distance from the origin i.e. $\mathcal{B}_A(1) \triangleq \left\{\mathbf{x} \in \mathbb{R}^d : d_A(\mathbf{x}, \mathbf{0}) \leq 1\right\}$. Show that $\mathcal{B}_A(1)$ is a convex set.

Exercise C.2. Consider the hyperplane given by the equation $\mathbf{w}^\top \mathbf{x} + b = 0$ i.e. $H \triangleq \left\{\mathbf{x} \in \mathbb{R}^d : \mathbf{w}^\top \mathbf{x} + b = 0\right\}$ where $\mathbf{w}$ is the normal vector to the hyperplane and b is the bias term. Show that H is a convex set.



Exercise C.3. If I take a convex set and shift it, does it remain convex? Let $\mathcal{C} \subset \mathbb{R}$ be a convex set and let $\mathbf{v} \in \mathbb{R}^d$ be any vector (whether "small" or "large"). Define $\mathcal{C} + \mathbf{v} \triangleq \{\mathbf{x} : \mathbf{x} = \mathbf{z} + \mathbf{v} \text{ for some } \mathbf{z} \in \mathcal{C}\}$. Show that the set $\mathcal{C} + \mathbf{v}$ will always be convex, no matter what $\mathbf{v}$ or convex set $\mathcal{C}$ we choose.



Exercise C.4. If I take a convex set and scale it, does it remain convex? Let $\mathcal{C} \subset \mathbb{R}$ be a convex set and let me scale dimension j using a scaling factor $s_j > 0$ i.e. for a vector $\mathbf{x}$, the scaled vector is $\tilde{\mathbf{x}}$ where $\tilde{\mathbf{x}}_j = s_j \cdot \mathbf{x}_j$. We can represent this operation using a diagonal matrix $S \in \mathbb{R}^{d \times d}$ where $S_{ii} = s_i$ and $S_{ij} = 0$ if $i \neq j$ i.e. $\tilde{\mathbf{x}} = S\mathbf{x}$. In the figure to the left, the x axis has been scaled up (expanded) 33% i.e. $s_1 = 1.333$ and the y axis has been scaled down (shrunk) by 33% i.e. $s_2 = 0.667$. Thus, in this example $S = \begin{bmatrix} 1.333 & 0 \\ 0 & 0.667 \end{bmatrix}$. Define $S\mathcal{C} \triangleq \{\mathbf{x} : \mathbf{x} = S\mathbf{z} \text{ for some } \mathbf{z} \in \mathcal{C}\}$. Show that the set $S\mathcal{C}$ will always be convex, no matter what positive scaling factors or convex set $\mathcal{C}$ we choose. Does this result hold even if (some of) the scaling factors are negative? What if some of the scaling factors are zero?

Exercise C.5. Above, we saw two operations (shifting a.k.a *translation* and scaling) that keep convex sets convex. However, can these operations turn a non-convex set into a convex set i.e. can there exist a non-convex set $\mathcal{C}$ such that $\mathcal{C} + \mathbf{v}$ is convex or else $S\mathcal{C}$ is convex when all scaling factors are non-zero? What if some (or all) scaling factors are zero?

Exercise C.6. Show that affine functions are always convex i.e. for any $\mathbf{a} \in \mathbb{R}^d, b \in \mathbb{R}$, the function $f(\mathbf{x}) = \mathbf{a}^\top \mathbf{x} + b$ is a convex function. Next, show that affine functions are always concave as well. In fact, affine functions are the only functions that are both convex as well as concave.

Exercise C.7. Show that affine functions when scaled by a positive constant, remain convex i.e. for convex function $f : \mathbb{R}^d \to \mathbb{R}$ and any $c > 0$, the function $g = c \cdot f$ is also convex. Next, show that if $c < 0$ then g is concave. What happens if $c = 0$ – does g become convex or concave?

Exercise C.8. The logistic loss function is very popular in machine learning and is defined as $\ell_{\text{logistic}}(t) = \ln(1 + \exp(-t))$. Show that given data points $\{(\mathbf{x}^i, y^i)\}_{i=1}^n$ (which are to be treated as constants) the function $f(\mathbf{w}) \triangleq \sum_{i=1}^n \ell_{\text{logistic}}(y^i \cdot \mathbf{w}^\top \mathbf{x}^i)$ is a convex function of $\mathbf{w}$.

References