# 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

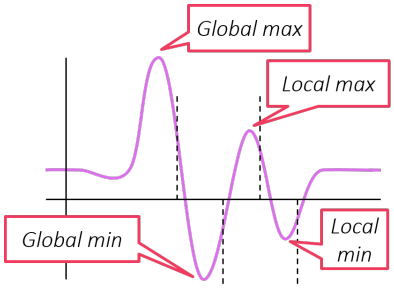
In this chapter we will take a look at basic tools from calculus that would be required to design and execute machine learning algorithms. Before we proceed, we caution the reader that the treatment in this chapter will *not* be mathematically rigorous and frequently, we will appeal to concepts and results 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.

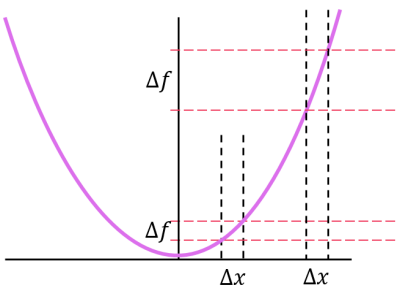
local extrema

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 (we will take care of multivariate functions later) $f : \mathbb{R} \to \mathbb{R}$, 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.




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 $\Delta x$

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

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

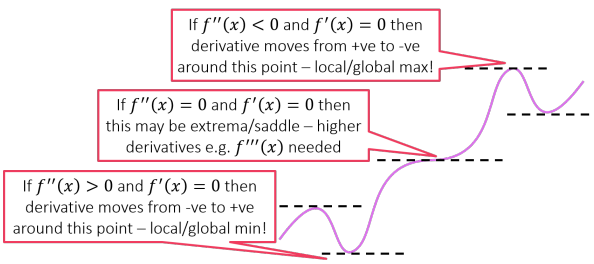
Integrating both sides and applying the fundamental theorem of algebra

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

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



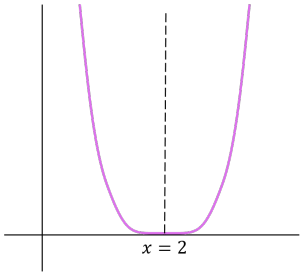

The stationary points of a function $f : \mathbb{R} \to \mathbb{R}$ are defined as the points where the derivative of the function vanishes i.e. $f'(x) = 0$. The stationary points of a function correspond to either the local/global maxima or minima or else saddle points. Given a stationary point, the 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 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.

### 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 $j^{\text{th}}$ 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 $i^{\text{th}}$ partial derivative changes if we move a little bit along the $j^{\text{th}}$ 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.

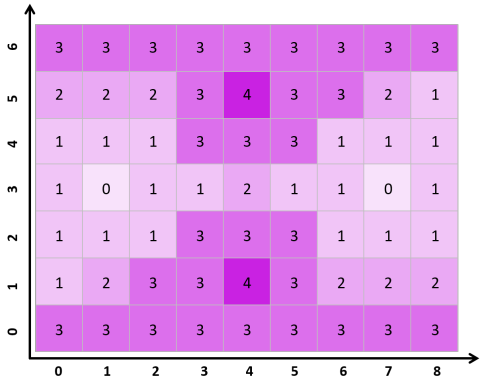
If the Hessian of the function is positive semi definite (PSD) at a stationary point $\mathbf{x}^0$ i.e. $\nabla f(\mathbf{x}^0) = \mathbf{0}$ and $H = \nabla^2 f(\mathbf{x}^0) \succeq 0$ then $\mathbf{x}^0$ is definitely a local minimum. Recall that a square symmetric matrix $A \in \mathbb{R}^{d \times d}$ 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

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

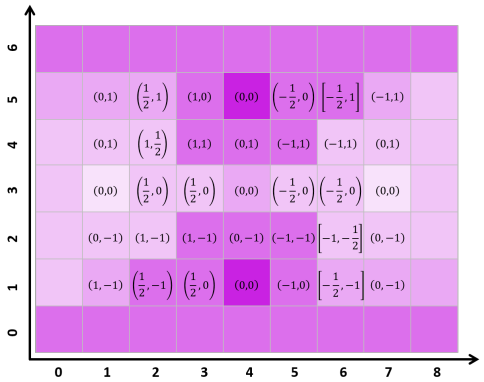
This means that no matter in which direction we move from $\mathbf{x}^0$, the function value always increases. This is the very definition of a local minimum. Similarly, we can intuitively see that if the Hessian of the function is negative semi definite (NSD) at a stationary point $\mathbf{x}^0$ i.e. $\nabla f(\mathbf{x}^0) = \mathbf{0}$ 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.



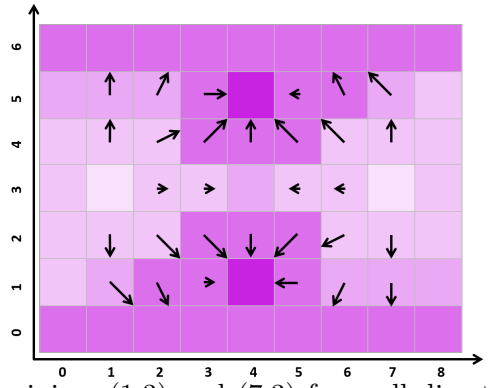

Consider the function $f : [0, 8] \times [0, 6] \rightarrow \mathbb{R}$ on the left. The function is discrete – darker shades indicate a higher function value (which is also written inside the boxes) and lighter shades indicate a smaller function value. Since discrete functions are non-differentiable, 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

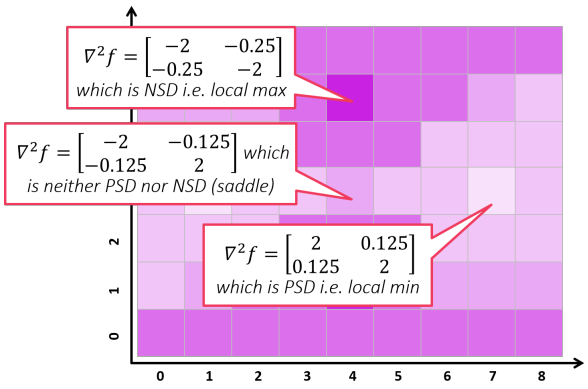
$$\frac{\Delta^2 f}{\Delta x^2} = \frac{f(x_0+1, y_0) + f(x_0-1, y_0) - 2f(x_0, y_0)}{1^2}$$

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

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

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

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




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 $k^{\text{th}}$ 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

# References