Math and Probability Review

Justin Silverman

Table of Contents

This material is intended as a review / quick primer only

The material in this lecture is intended to be a primer / review of basic concepts in calculus, linear algebra, and probability. This information is not intended to be comprehensive. If you find this material unfamiliar, or if during this course you come across material not covered in this lecture I expect you to bring yourself up to speed on these topics. The following resources may provide additional insights and background or should provide nice references throughout this course.

- Mathematics for Machine Learning (link)
- Linear Algebra Review and Reference (link)
- Linear algebra explained in 4 pages (link)
- Probability CheatSheet (link)

The above resources were chosen as they are "high-yield", short and sweet. Still, those wishing/needing a more formal introduction are encouraged to consult basic texts on these topics.

Material Source

This lecture contains material from the following sources:

- Nihit Desai, Sameep Bagadia, Review of Linear Algebra for CS224W at Stanford
- Garrett Thomas, Mathematics for Machine Learning (link)
- rhome Medium article Convex optimization, unconstrained

Section 1

Review of Linear Algebra

Matrices and Vectors

- Matrix: A rectangular array of numbers, e.g., $A \in \mathbb{R}^{m \times n}$:

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}$$

- Vector: A matrix consisting of only one column (default) or one row, e.g., $x \in \mathbb{R}^n$

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

Matrix Addition

Vectors and Matricies add Component wise – so they must be of the same dimension.

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} \\ a_{21} + b_{21} & a_{22} + b_{22} \end{bmatrix}$$

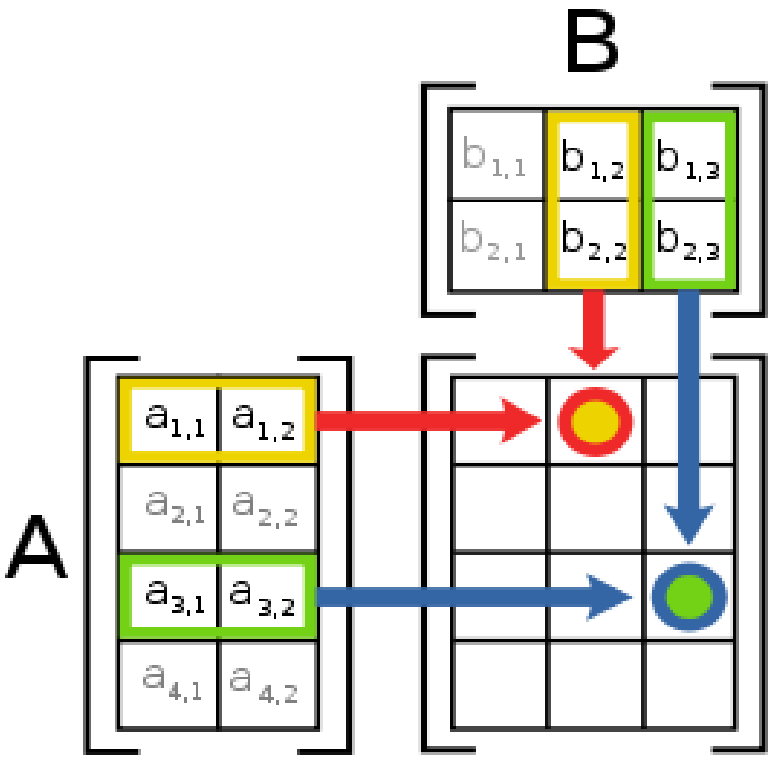
Matrix Multiplication

- If $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{n \times p}$, then their product $AB \in \mathbb{R}^{m \times p}$
  - Number of columns of A **must** equal number of rows of B
- We can compute the product $C = AB$ using this formula:

$$C_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}$$

Matrix Multiplication

$$(AB)(x) = A(B(x))$$



Properties of Matrix Multiplication

- Associative: $(AB)C = A(BC)$
- Distributive: $A(B + C) = AB + AC$
- Non-commutative: $AB \neq BA$
  - They don't even have to be the same size!

Linear Transformation and Matrices

A linear transformation T is a function from $\mathbb{R}^n$ to $\mathbb{R}^m$ that satisfies two properties:

1. For all $x, y \in \mathbb{R}^n$,

$$T(x + y) = T(x) + T(y)$$

2. For all $x \in \mathbb{R}^n$ and all $a \in \mathbb{R}$ (scalar)

$$T(ax) = aT(x)$$

Every linear transformation can be represented by a matrix.
Every matrix is a linear transformation.

Matrix Transpose

- Transpose: $A \in \mathbb{R}^{m \times n}$, then $A^T \in \mathbb{R}^{n \times m}$: $(A^T)_{ij} = A_{ji}$.
  For example, if

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$$

  then

$$A^T = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$$

- Properties:
  - $(A^T)^T = A$
  - $(AB)^T = B^T A^T$
  - $(A + B)^T = A^T + B^T$

Matrix Inverse

- If $A \in \mathbb{R}^{n \times n}$ and invertible, then the inverse of A, denoted A^{-1} is the matrix that: $AA^{-1} = A^{-1}A = I$
- Properties:
  - $(A^{-1})^{-1} = A$
  - $(AB)^{-1} = B^{-1}A^{-1}$
  - $(A^{-1})^T = (A^T)^{-1}$

Identity Matrix

- Identity matrix: $I = I_n \in \mathbb{R}^{n \times n}$:

$$I_{ij} = \begin{cases} 1 & \text{i=j,} \\ 0 & \text{otherwise.} \end{cases}$$

- $\forall A \in \mathbb{R}^{m \times n}$: $AI_n = I_m A = A$

$$I_1 = \begin{bmatrix} 1 \end{bmatrix}, \; I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \; I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \; \cdots, \; I_n = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

Diagonal Matrix

- Diagonal matrix: $D = diag(d_1, d_2, \ldots, d_n)$:

$$D_{ij} = \begin{cases} d_i & \text{j=i,} \\ 0 & \text{otherwise.} \end{cases}$$

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & -3 \end{bmatrix}$$

Other Special Matrices

- Symmetric matrices: $A \in \mathbb{R}^{n \times n}$ is symmetric if $A = A^T$.
- Orthogonal matrices: $U \in \mathbb{R}^{n \times n}$ is orthogonal if $UU^T = I = U^T U$
  - Every column is orthogonal to every other column (dot product = 0)
  - The inverse of an orthogonal matrix is its transpose

Linear combinations and Span

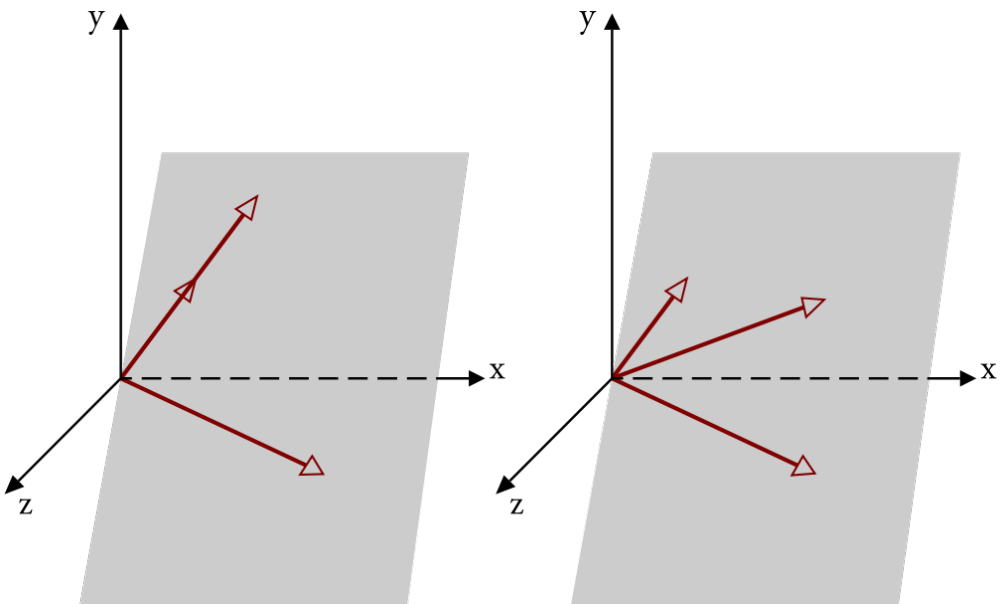
- Given a set of vectors $S = \{x_1, \ldots, x_n\}$ where $x_i \in \mathbb{R}^n$, a linear combination of this set of vectors is an expression of the form: $\sum_{i=1}^{n} \alpha_i x_i$ where $\alpha_i \in \mathbb{R}$
- $Span(S)$ is the set of all linear combinations of the elements of S

Linear Independence and Rank

- A set of vectors $S = \{x_1, \ldots, x_n\}$ is linearly independent if the following holds: $\sum_{i=1}^{n} \alpha_i x_i = 0$ only if $\alpha_1 = \alpha_2 \ldots = \alpha_n = 0$
- Rank: $A \in \mathbb{R}^{m \times n}$, then $rank(A)$ is the maximum number of linearly independent columns (or equivalently, rows)
- Properties:
  - $rank(A) \leq \min\{m, n\}$
  - $rank(A) = rank(A^T)$
  - $rank(AB) \leq \min\{rank(A), rank(B)\}$
  - $rank(A + B) \leq rank(A) + rank(B)$

Example of Linear Dependence

These three vectors are linearly dependent because they all lie in the same plane. The matrix with these three vectors as rows has rank 2.



Eigenvalues and Eigenvectors

- Given $A \in \mathbb{R}^{n \times n}$, $\lambda \in \mathbb{C}$ is an eigenvalue of A with the corresponding eigenvector $x \in \mathbb{C}^n$ ($x \neq 0$) if:

$$Ax = \lambda x$$

For example, if

$$A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

then the vector $\begin{bmatrix} 3 \\ -3 \end{bmatrix}$ is an eigenvector with eigenvalue 1, because

$$A\mathbf{x} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 3 \\ -3 \end{bmatrix} = \begin{bmatrix} 2 \cdot 3 + 1 \cdot (-3) \\ 1 \cdot 3 + 2 \cdot (-3) \end{bmatrix} = \begin{bmatrix} 3 \\ -3 \end{bmatrix} = 1 \cdot \begin{bmatrix} 3 \\ -3 \end{bmatrix}.$$

Intuitive View of Eigenvectors and Eigenvalues

Solving for Eigenvalues/Eigenvectors

- Characteristic Polynomial: If $Ax = \lambda x$ then

$$(A - \lambda I)x = 0$$

so $(A - \lambda I)$ is singular (not full rank), so

$$\det(A - \lambda I) = 0.$$

Thus the eigenvalues are exactly the n possibly complex roots of the degree n polynomial equation $det(A - \lambda I) = 0$. This is known as the characteristic polynomial.

- Once we solve for all λ's, we can plug in to find each corresponding eigenvector.

Eigenvalue/Eigenvector Properties

- Usually eigenvectors are normalized to unit length.
- If A is symmetric, then all the eigenvalues are real
- $tr(A) = \sum_{i=1}^{n} \lambda_i$
- $det(A) = \prod_{i=1}^{n} \lambda_i$

Matrix Eigendecomposition

$A \in \mathbb{R}^{n \times n}$, $\lambda_1, \ldots, \lambda_n$ the eigenvalues, and $x_1, \ldots, x_n$ the eigenvectors. $P = [x_1|x_2|\ldots|x_n]$, $D = diag(\lambda_1, \ldots, \lambda_n)$, then:

$$
\begin{aligned}
AP &= A\begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 & \cdots & \mathbf{X}_k \end{bmatrix} \\
&= \begin{bmatrix} A\mathbf{X}_1 & A\mathbf{X}_2 & \cdots & A\mathbf{X}_k \end{bmatrix} \\
&= \begin{bmatrix} \lambda_1 \mathbf{X}_1 & \lambda_2 \mathbf{X}_2 & \cdots & \lambda_k \mathbf{X}_k \end{bmatrix} \\
&= \begin{bmatrix} \lambda_1 x_{11} & \lambda_2 x_{21} & \cdots & \lambda_k x_{k1} \\ \lambda_1 x_{12} & \lambda_2 x_{22} & \cdots & \lambda_k x_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_1 x_{1k} & \lambda_2 x_{2k} & \cdots & \lambda_k x_{kk} \end{bmatrix} \\
&= \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{k1} \\ x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1k} & x_{2k} & \cdots & x_{kk} \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k \end{bmatrix} \\
&= PD,
\end{aligned}
$$

Eigendecomposition and Matrix Approximation (but simpler)

For many square matrices (diagonalizable aka non-defective square matrices) X we can write the Eigen-decomposition can be written as

$$X = VDV^T$$

for orthonormal matrix V and diagonal matrix D.

The Singular Value Decomposition

The singular value decomposition (SVD) is a generalization of the eigen decomposition to non-square matrices.

For a square or non-square matrix Y we write the SVD as $Y = UDV^T$ where U and V have certain properties but must not be equivalent and don't have to be square.

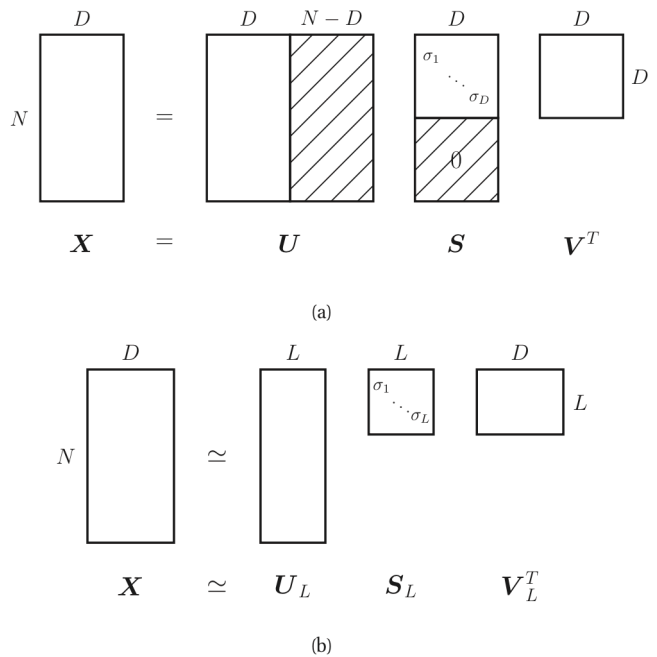


Figure 12.8 (a) SVD decomposition of non-square matrices $\mathbf{X} = \mathbf{USV}^T$. The shaded parts of $\mathbf{S}$, and all the off-diagonal terms, are zero. The shaded entries in $\mathbf{U}$ and $\mathbf{S}$ are not computed in the economy-sized version, since they are not needed. (b) Truncated SVD approximation of rank L.

Matrix Inverse from Eigen Decomposition

If $X = VDV^T$ then

$$X^{-1} = VD^{-1}V^T.$$

Since D is diagonal its inverse is given by just taking the inverse of each of its diagonal elements.

Positive Definite Matricies

A symmetric matrix is called **positive definite** if all its eigenvalues are strictly greater than zero.

A symmetric matrix is called **positive semi-definite** if all of its eigenvalues are greater than zero (can include zero).

Why you should care: Covariance matricies are symmetric positive semi-definite matrices.

Matrix Square Roots

The square root of a square matrix X is defined as a matrix V such that $X = VV^T$. In this case we write that $X^{1/2} = V$.

The Eigen-decomposition provides one means of finding such a matrix V for a square matrix X.

$$X^{1/2} = VD^{1/2}$$

where it is clear that $[D]_{ii}^{1/2} = \sqrt{D_{ii}}$.

The Cholesky Decomposition

For Symmetric Positive Definite Matrices have a special decomposition that is often faster than the eigen decomposition to compute. It is called the Cholesky decomposition. For SPD matrix Σ it has a Cholesky decompostion of

$$\Sigma = LL^T$$

where L is a lower triangular matrix.

The Cholesky decomposition is very useful for calculating matrix square roots and inverses of symmetric positive definite matricies.

Vector p-norms

Think of a norm as the length of the vector. We have the standard Euclidean (aka ℓ_2 norm)

$$||x||_2 = \sqrt{\sum_i x_i^2}.$$

We also have the city-block norm (aka ℓ_1 norm)

$$||x||_1 = \sum_i |x_i|.$$

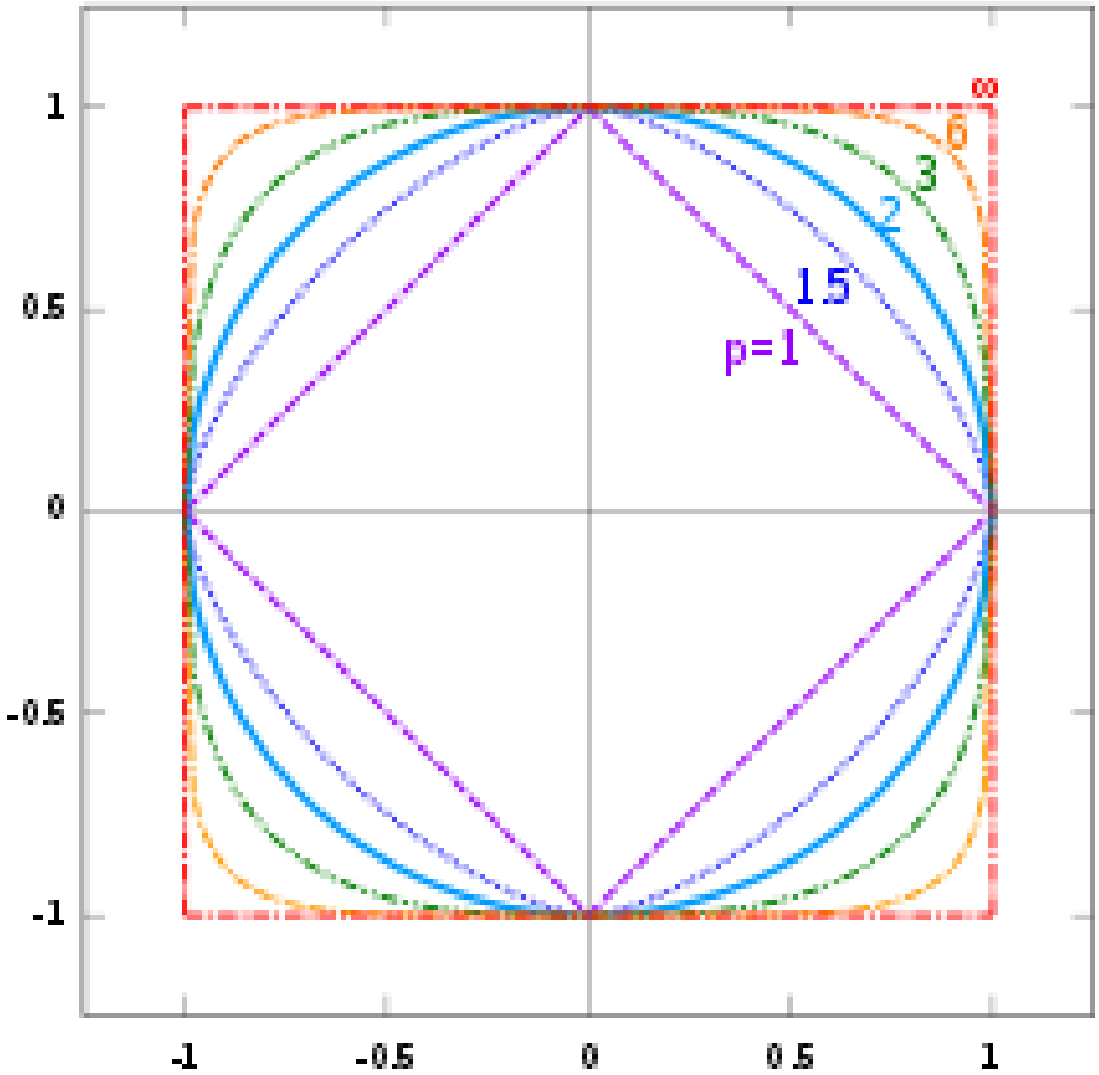
p-norms

For $p \geq 1$ the p-norm (also called the ℓ_p-norm) of a vector x is given by

$$||x||_p = \left(\sum_i |x_i|^p \right)^{1/p}$$

Visualizing p-norms

There are an infinite number of p-norms and we can visualize them by thinking about the areas of equal "length" in a 2D plane [1].



Why is linear algebra so central?

Linear algebra provides a means of operating on linear systems of equations. What is below should make you think of multivariate linear regression.

$$y_1 = \beta_1 x_{11} - \beta_2 x_{12}$$
$$y_2 = \beta_1 x_{21} - \beta_2 x_{22}$$

is equivalent to

$$y = X\beta$$

We can solve for β as $\beta = X^{-1}y$.

Section 2

Review of Derivatives

The Gradient

The single most important concept from calculus in the context of machine learning is the *gradient*. Gradients generalize derivatives to scalar functions of several variables. The gradient of $f : \mathcal{R}^d \to \mathcal{R}$, denoted ∇f, is given by

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix} \quad \text{i.e.} \quad [\nabla f]_i = \frac{\partial f}{\partial x_i}$$

Gradients have the following very important property: $\nabla f(x)$ points in the direction of *steepest ascent* from x. Similarly, $-\nabla f(x)$ points in the direction of *steepest descent* from x. We will use this fact frequently when iteratively minimizing a function via *gradient descent*.

The Jaccobian

The *Jaccobian* of $f : \mathcal{R}^n \to \mathcal{R}^m$ is a matrix of of first-order partial derivaties:

$$\mathbf{J}_f = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

Note the special case where $m = 1$ in which $\nabla f = \mathbf{J}_f^T$. This comes in handy when working with changes of variables.

The Hessian

The *Hessian* matrix of $f : \mathcal{R}^n \to \mathcal{R}$ is a matrix of second order partial derivatives:

$$\nabla_f^2 = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_1} \end{bmatrix}$$

Recall that if the partial derivatives are continuous, the order of differentiation can be interchanged (Clairaut's theorem), so the Hessian matrix will be symmetric. This will typically be the case for differentiable functions that we work with.

The Hessian is used in some optimization algorithms such as Newton's method. It is expensive to calculate but can drastically reduce the number of iterations needed to converge to a local minimum by providing information about the curvature of f. It is also used as the basis for a Laplace Approximation to a probability density.

Section 3

Review of Optimization

Preliminaries

- Why Optimization?

$$\hat{\theta} = \underset{\theta}{\mathit{argmax}}\ p(Y, X|\theta)$$

Preliminaries

- Why Optimization?

$$\hat{\theta} = \underset{\theta}{\mathit{argmax}}\ p(Y, X|\theta)$$

- Local vs. Global Minima vs. Maxima

Preliminaries

- Why Optimization?

$$\hat{\theta} = \underset{\theta}{argmax}\ p(Y, X|\theta)$$

- Local vs. Global Minima vs. Maxima

- (For nicely behaved functions) at an optima (Local or Global, Minima or Maxima) the gradient is zero.

Review of Optimization

Again, for nicely behaved functions the gradient is zero at a local extrema (minima or maxima). However the gradient can also be zero for saddle points so be careful. You also need to know curvature to say if its an extrema or a saddle point.

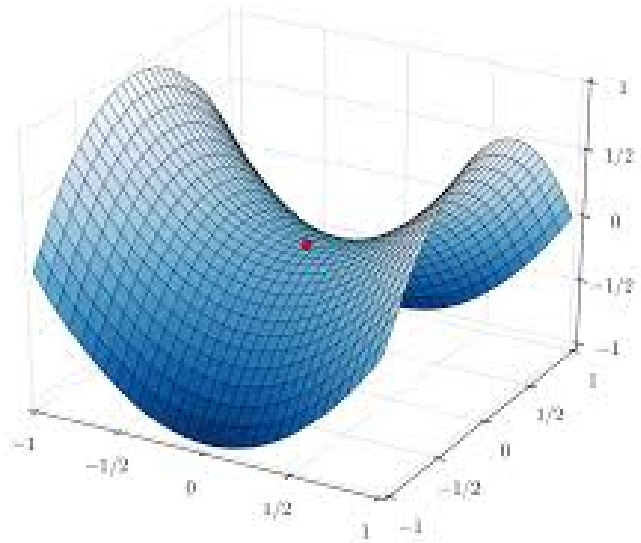


Figure 1: Image from Wikipedia

Convexity

Convexity is a term that can apply to both sets and functions. We are not going to deal formally with convexity but instead deal with the intuition behind why convexity matters.

Informally: A *set is convex if* any line through the set only enters and exits the set once. In other words, think about convex sets as being things that are not shaped like the letter "C".

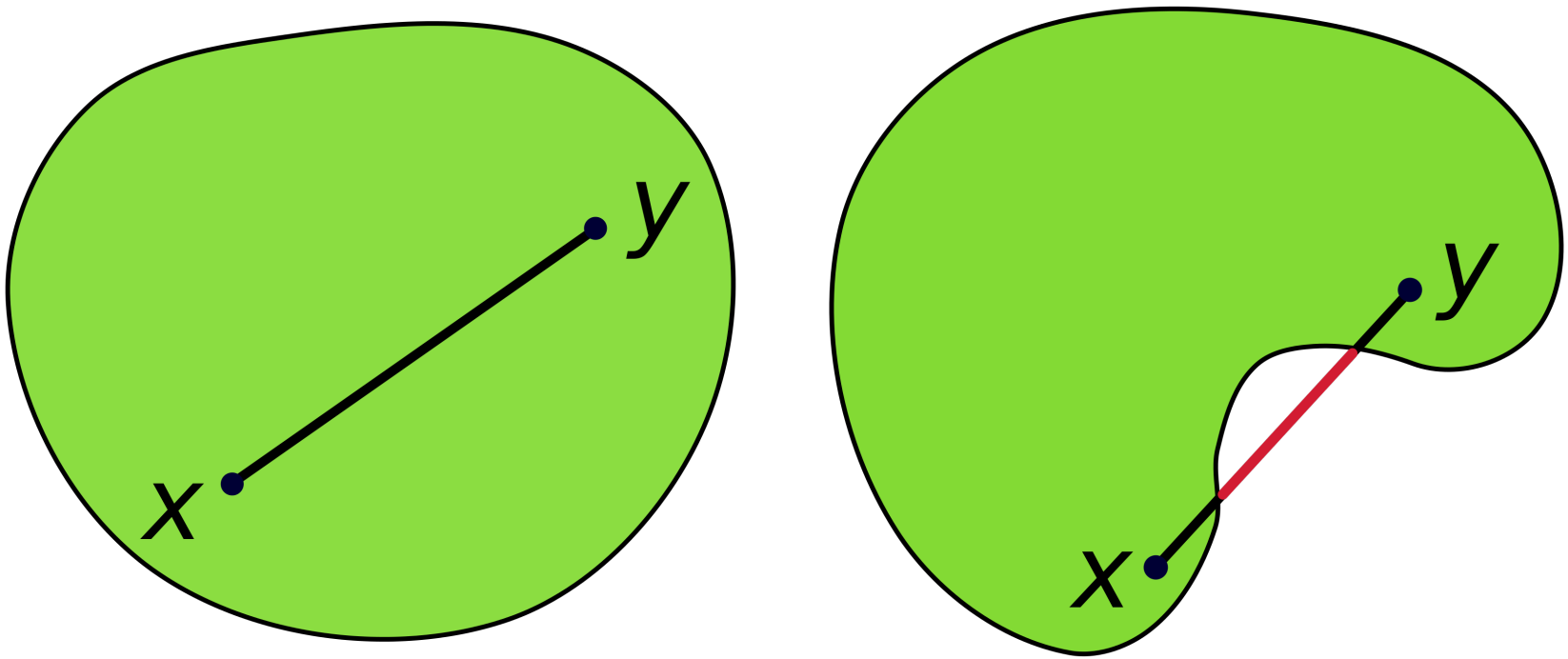
Informally: A *function is convex if* its shaped like a hill or a valley (but not a hill with multiple peaks or a valley with multiple low-points).[2]

Why we care: Think about how easy it is to get "stuck" during optimization when you have a non-convex function or a non-convex set.

[2] Norms are convex.

Convex Set Examples

Informally: A *set is convex if* any line through the set only enters and exists the set once. In other words, think about convex sets as being things that are not shaped like the letter "C".



(a) A convex set

(b) A non-convex set

Figure 2: What convex sets look like

Convex Function Example

Informally: A *function is convex if* its shaped like a hill or a valley (but not a hill with multiple peaks or a valley with multiple low-points).

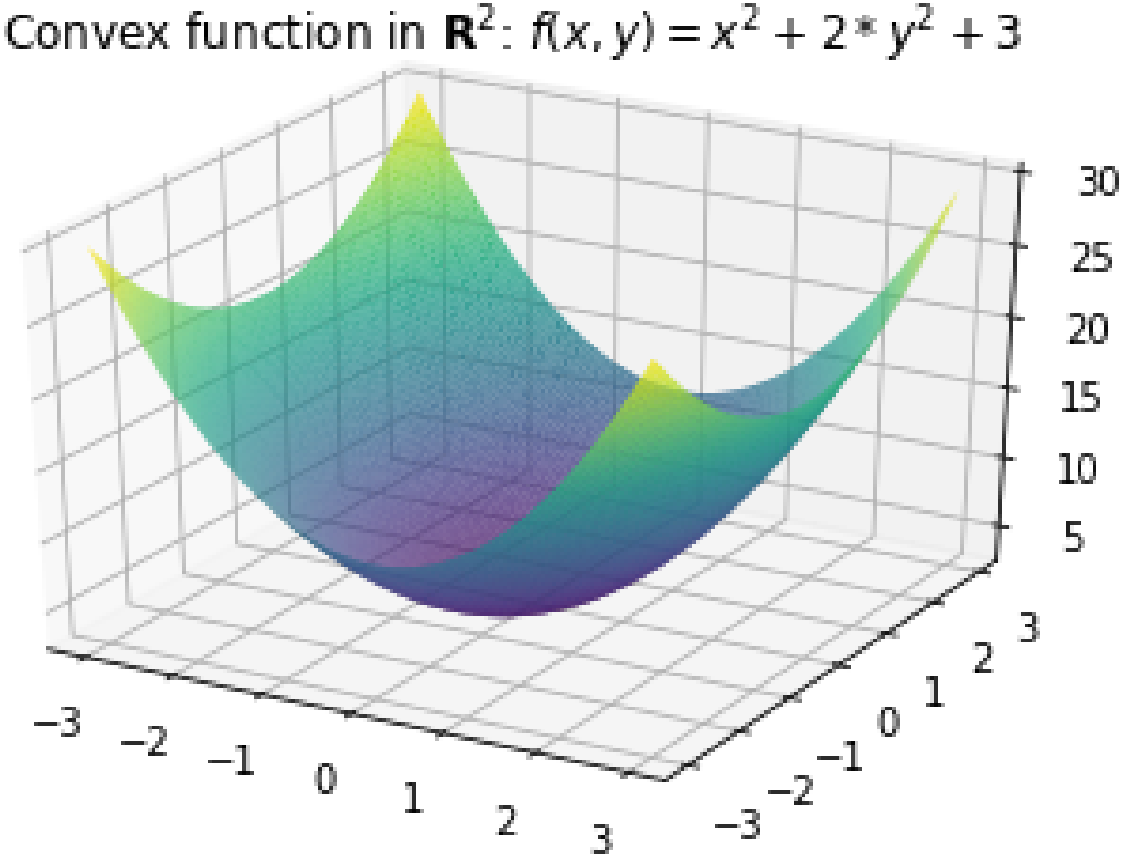


Figure 3: An example of a convex function[3]

[3] image from rhome review of Convex optimization on medium

Common Convex Functions

- Exponential: $f(x) = e^{ax} \geq 0, \quad \forall a \in R$
- Powers: $f(x) = x^a \geq 0, \quad \forall a \geq 1 \text{ or } a < 0, x > 0$
- Negative Logarithm: $f(x) = (-1)*\log(x)$
- Negative entropy: $f(x) = x*\log(x)$
- Max, norms, negative geometric mean... and more.

Popular convex functions

- Weighted sum: $g(x) = \sum_i w_i * f_i(x), \quad for\ w_i \geq 0$
- Affine composition: $g(x) = f(A * x + b)$
- Composition:
  - If g is convex, then exp(g(x)) is convex
  - If g is concave and positive, then log(g(x)) is concave
  - If g is concave and positive, then 1/g(x) is convex

Popular operations and their effect on convexity

Figure 4: Common Convex Functions[4]

[4] image from rhome review of Convex optimization on medium

Gradient based optimization

All gradient based optimization algorithms have a similar form. Assuming we are trying to optimize a function $f : \mathcal{R}^p \to \mathcal{R}$

- Find an initial point, $x_0 \in \mathcal{R}^p$.[5]
- Repeat the following update until convergence:
  - Determine a search direction a_k[6]
  - Determine a step size s_k
  - Update x: $x_{k+1} = x_k + a_k s_k$
  - (evaluate convergence criteria)

Typically these algorithms are guaranteed to converge to the global optima in finite time if the function and the set over which you are optimizing is convex. However some gradient based algorityms are faster than others.

[5] The closer you can get to the optima the better but often its just a random guess.

[6] Typically based on gradient and hessian

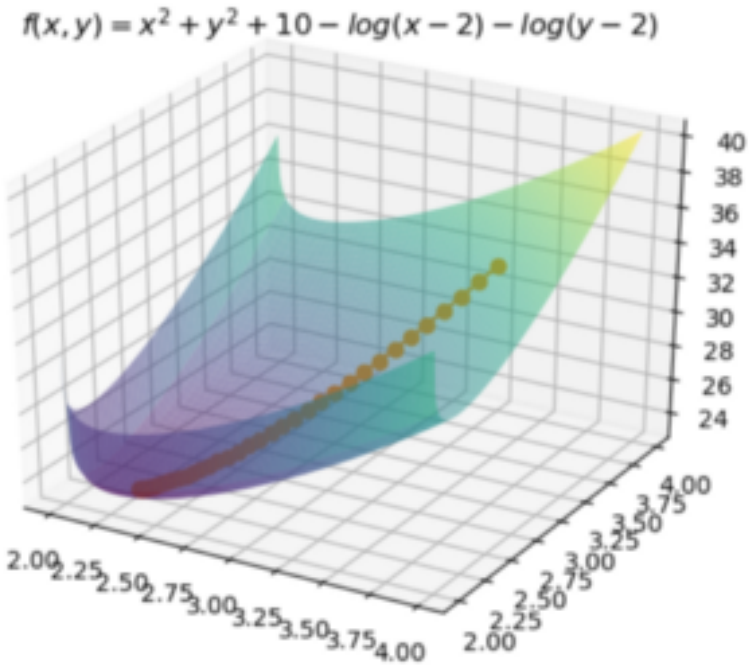
Gradient Descent

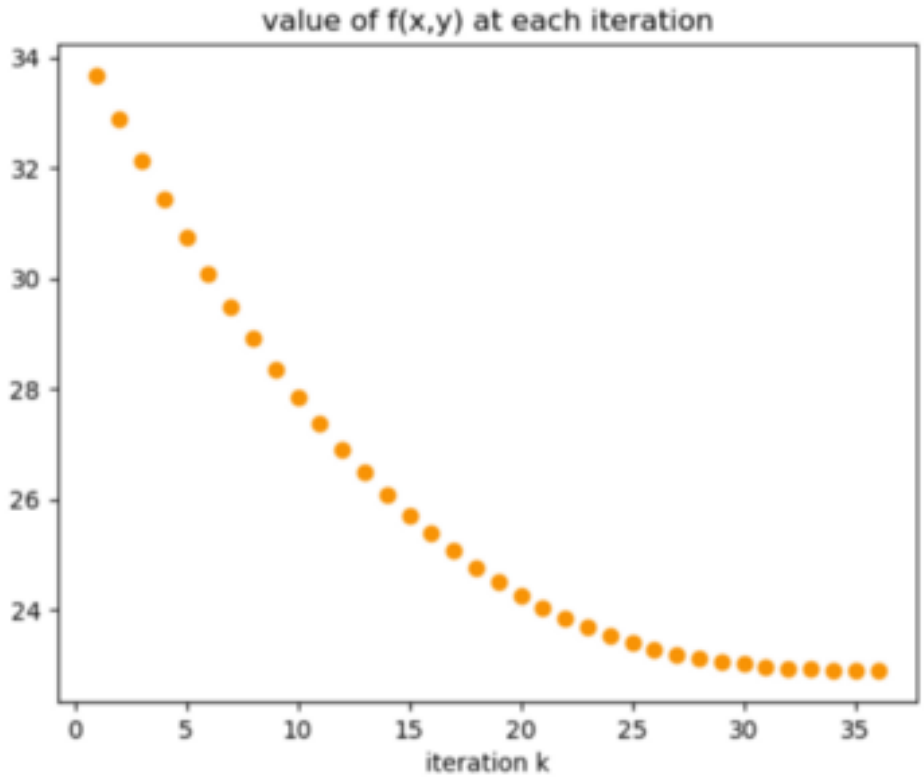
Gradient Descent is one example of a Gradient based optimization technique:

- The search direction is set to $a_k = -\nabla f(x_k)$
- the step size is typically fixed and set by the user at the start of the algorithm but there are all kinds of variations here
- Stopping criteria can vary but generally come in two flavors:
  - $|f(x_{k+1}) - f(x_k)| < \epsilon$, *i.e.*, diminishing returns
  - $||\nabla f(x_k)|| < \epsilon$, *i.e.*, gradient is close enough to zero.

Gradient Descent example:

Code to reproduce and intelectual source of the following image can be found (here)





Section 4

Review of Probability

Joint Probability (Visually)

$p(A, B)$

Marginal and Conditional Probability (Visually)

Marginal Probability, e.g., $p(A)$

Conditional Probability, e.g., $p(B|A)$

Simple Rules of Probabilty

The joint probability $p(A, B)$ can be factored into a conditional probability $p(A|B)$ and a marginal probability $p(B)$

$$p(A, B) = p(A|B)p(B).$$

The marginal probability $p(B)$ is related to the joint $p(A, B)$ by summing over all the possible values of A

$$p(B) = \int_{A \in \mathcal{A}} p(A, B) dA$$

where $\mathcal{A}$ is just notation for "all the possible values of A".

Baye's rule follows from the factoring identity above:

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}.$$

Expectations, the Mean and Variance

Let x be a random variable with probability density (if continuous) or probability mass (if discrete) function $p(x|\theta)$ (where θ just refers to some parameters of the function). We could have equivalently written $x \sim p(\theta)$, they are just different notations.

Expectations

We define the expectation of a function f of x to be

$$E_p[f(x)] = \int f(x)p(x|\theta)dx$$

Expectations, the Mean and Variance

Two important expectations are the mean and variance: The mean of x is defined as:

$$\text{mean}(x) = E_p[x] = \int x p(x|\theta) dx$$

Expectations, the Mean and Variance

Two important expectations are the mean and variance: The mean of x is defined as:

$$\text{mean}(x) = E_p[x] = \int x p(x|\theta) dx$$

The variance of x is defined as:

$$\text{var}(x) = E_p(x^2) - (E_p(x))^2$$

which is equivalent to $E_p[(x - E_p[x])^2]$.

From its definition it follows that variance must be positive. You can think of variance as the spread about the mean.

Mean and Variance of Finite Samples

Don't get confused. The above definitions of mean and variance are the calculations/definitions if you knew the true distribution of x. If instead we only have a finite set of samples of x then we must estimate the mean and variance. Common estimators for mean and variance are

$$\widehat{\text{mean}}(x) = \frac{1}{N}\sum_i x_i$$

and

$$\widehat{\text{var}}(x) = \frac{1}{N}\sum_i (x_i - \widehat{\text{mean}}(x))^2$$

But these are just one choice of estimators!

Covariance

Covariance is a generalization of variance to more than one random variable.

$$\text{Cov}(x, y) = E[(x - E[x])(y - E[y])]$$

Covariance is symmetric such that $\text{Cov}(x, y) = \text{Cov}(y, x)$. This is why covariance matricies are symmetric. Notice that $\text{Cov}(x, x) = \text{Var}(x)$

A common estimator for covariance is

$$\widehat{\text{Cov}}(x, y) = \frac{1}{N-1} \sum_i (y_i - \widehat{\text{mean}}(y))(x_i - \widehat{\text{mean}}(x))$$

Covariance Matrix

The covariance matrix is a generalization of covariance to more than 2 variables. Consider a random variable $\mathbf{x} = (x_1, \ldots, x_k)$. The covariance matrix $Cov(\mathbf{x}) = \Sigma$ is defined such that the elements of Σ are given by

$$\Sigma_{ij} = \text{Cov}(x_i, x_j).$$

Covariance matrices are symmetric positive semi-definite: They are symmetric and have eigenvalues greater than or equal to zero.

Correlation

Correlation can be defined as a normalized covariance. Let $\Sigma = Cov(X)$ be a covariance matrix. Let $D = \text{diag}(\Sigma)$ (e.g., a matrix with just the diagonal elements of Σ and all off-diagonals equal to zero).

$$Corr(X) = D^{-\frac{1}{2}} \Sigma D^{-\frac{1}{2}}$$

It follows from this definition that the diagonal elements of a correlation matrix have to be equal to 1.

The Multivarite Normal

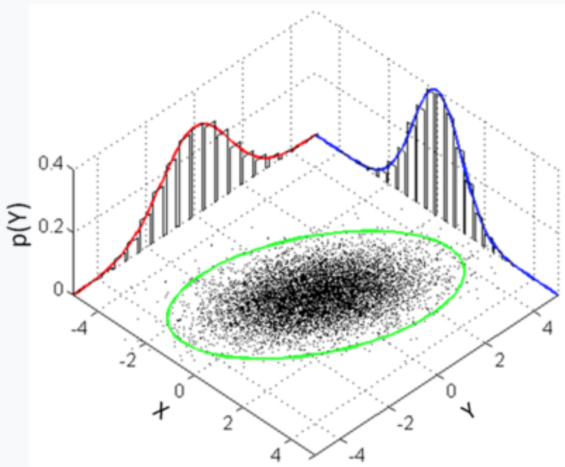
Most statistics and machine learning is based on the multivariate normal distribution distribution in one way or another.

$$\mathbf{x} = (x_1, \ldots, x_k) \sim N(\mu, \Sigma)$$

Here μ is a k-vector (the mean $E[x_i] = \mu_i$) and Σ is a $k \times k$ covariance matrix such that $\mathrm{Cov}(x_i, x_j) = \Sigma_{ij}$.

Multivariate normal

Probability density function



Many sample points from a multivariate normal distribution with $\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 3/5 \\ 3/5 & 2 \end{bmatrix}$, shown along with the 3-sigma ellipse, the two marginal distributions, and the two 1-d histograms.

Notation	$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Properties of the Multivariate Normal

There are many reasons why the multivariate normal is central to data analysis. One reason is that **the marginals and conditionals of the multivariate normal are also multivariate normal.**

Set up to the properties on the next two slides

Let the dimensions of $\mathbf{x}$ be partitioned into k_1 and k_2 dimensional subsets where $k = k_1 + k_2$ such that we can write $\mathbf{x} = (\mathbf{x_1}, \mathbf{x}_2)$ and we can write

$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$

Marginal Property

$$p(\mathbf{x}_1) = \int p(\mathbf{x}_1, \mathbf{x}_2) d\mathbf{x}_2$$

$$p(\mathbf{x}_1) \sim N(\mu_1, \Sigma_{11})$$

Conditional Property

$$p(\mathbf{x}_1|\mathbf{x}_2 = a) \sim N(\mu^*, \Sigma^*)$$

where

$$\mu^* = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(a - \mu_2)$$

$$\Sigma^* = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

Affine Transformation Property

If $Y = c + BX$ is an affine transformation of $X \sim N(\mu, \Sigma)$ then the distribution of Y is given by:

$$Y \sim N(c + B\mu, B\Sigma B^T)$$

Sum of Two Normal Random Variables

If $x_1 \sim N(\mu_1, \Sigma_1)$ and $x_2 \sim N(\mu_2, \Sigma_2)$ [7]

$$(x_1 + x_2) \sim N(\mu_1 + \mu_2, \Sigma_1 + \Sigma_2)$$

[7] where Σ_1 and Σ_2 are covariance matrices

Sampling from the Multivariate Normal

The following relationship is essential for efficiently sampling from the multivariate normal.

Three Equivalent Parameterizations

$$x \sim N(\mu, \Sigma)$$

$$x = \mu + e \text{ where } e \sim N(0, \Sigma)$$

$$x = \mu + \Sigma^{\frac{1}{2}} e \text{ where } e \sim N(0, I)$$

In the last relationship have reduced the problem of sampling from the multivariate normal to sampling from k univariate normal with mean of 0 and variance of 1. This is one of the core reasons that the Cholesky factorization is so useful (it lets us efficiently calculate the square root of Σ).

Univariate Normal / Multivariate Normal Notation

Let $x \in \mathcal{R}^p$, the following two notations are equivalent:

1. $x \sim N(0, I)$
2. $x_i \overset{iid}{\sim} N(0, 1)$