### 12 Inequalities

**Markov’s Inequality**:

Proof:

1. Let indicator if and otherwise.  
   Then i.e. .

**Chebyshev’s Inequality**:

*reminder*:

Proof:

1. Since is a non-negative RV, we can apply Markov’s inequality with :
2. , so

### 13 Sample Mean, Weak Law of Large Numbers & CLT

**Distribution of Sample Mean**:

Random variable

Suppose the s are independent and identically distributed.

Each has mean and variance .

We can calculate the mean of as:

We say is an **unbiased estimator** of since .

The variance of is:

**The Weak Law of Large Numbers**:

Consider independently identically distributed random variables. Let .

By Chebyshev’s inequality:

**The Central Limit Theorem (CLT)**:

When independent random variables are added, their properly normalized sum tends toward a normal distribution, even if the original variables themselves are **not** normally distributed.

Histogram of as increases, but we normalise to keep the area of the curve fixed.

The curve narrows as increases.

**Normal / Gaussian Distribution**:

CLT says that as increases, the distribution of converges to a normal / Gaussian distribution.

### 14 Confidence Intervals

**Bootstrapping** is a **resampling** method:

1. Draw a sample of N data points uniformly at random from the data, with **replacement**.
2. Using this sample, estimate the mean
3. Repeat to generate a set of estimates
4. The distribution of these estimates approximates the distribution of the sample mean.

|  |
| --- |
| Central Limit Theorem: ()   * Gives full distribution of X. * Only requires the mean and variance to fully describe the distribution. * However, it is an approximation when N is finite, and is hard to be sure how accurate it is and how big N should be. |
| Chebyshev and other inequalities:   * Provides an actual bound. * Works for all N. * However is loose in general. |
| Bootstrapping:   * Gives a full distribution of X, but doesn’t assume normality. * However it is an approximation when N is finite, and it’s hard to be sure of how accurate it is and how big N should be. * Requires availability of all N measurements. |

### 15 Continuous Random Variables

**Cumulative Distribution Function**:

Suppose is a random variable, then is the CDF.

**Probability Density Function**:

For a continuous-valued RV there exists a function such that:

is the PDF of .

The **area under the PDF** between two points is the **probability** .

**Uniform Random Variables**:

is a uniform random variable when it has a PDF:

**The Normal Distribution**:

is a **Normal / Gaussian random variable**  when it has a PDF:

**Joint Cumulative Distribution Function**:

Suppose and are two random variables:

is the CDF.

When and are independent then:

When and are two discrete random variables then:

When and are two continuously-valued random variables with PDF :

**Conditional Probability Density Function**:

Suppose and are two continuously-valued random variables with joint PDF .

Define conditional PDF:

**Chain Rule for PDFs**:

**Bayes’ Rule for PDFs**:

**Independence**:

### 16 Linear Regression

**Hypothesis**:

**Least Squares Case**:

**Gradient Descent**:

1. Start with some and .
2. *Repeat*: Update and to new values which makes smaller.

Therefore:

**Linear Regression with Multiple Variables**:

Feature vector

Parameter vector

**Gradient Descent for Multiple Variables**:

### 17 Estimation

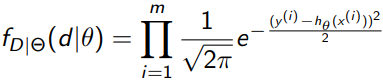
**Probabilistic Interpretation of Linear Regression**:

where:

* is Gaussian noise with mean 0 and variance 1.

Assuming:

The likelihood of the training data is therefore:



In linear regression, we select to maximise.

i.e. we minimise .

**Closed-Form Solution**:

1. Select to maximise
2. Compute derivative with respect to :
3. Set derivative equal to zero and solve for :

**Fitting Non-Linear Curves**:

Hypothesis:

Define a feature vector with and apply previously discussed ideas.

**Overfitting**:

As we add more parameters to the hypothesis, we start to fit the noise in the training data.

**Underfitting**:

If we use too few parameters in the hypothesis we will get a poor fit.

Overfitting vs. underfitting = **bias-variance trade-off**.

**Bayesian Estimation**:

Estimate the posterior rather than the likelihood . A **distribution** rather than a single value.

**MAP (Maximum a posteriori) Estimation**:

Select that maximises posterior back to a single value, rather than a distribution.

Taking logs, select to maximise .

i.e. to minimise .

This is called **ridge regression**.

* When , we are certain .
* When is large, we know little about prior to observations.

Map estimation:

1. Select to maximise the log-posterior:
2. Differentiate with respect to :
3. Set derivative equal to zero:

**MAP vs. Maximum Likelihood Estimation**:

As our number of observations grows, the impact of prior on posterior tends to decline.

### 18 Logistic Regression

**Classification with Two Classes**:

Output variable takes values or . is often referred to as the **label**.

Predict output when and when .

**Logistic regression** tries to fit a plane that separates the and data.

Data is **linearly separable** if the heuristic divides the data so that all s are separated from all s by the line. Not all data is linearly separable.

**0-1 Loss Function**:

For **logistic regression** we use:

noting that or .

**Gradient Descent**:

1. Start with some .
2. *Repeat*:

For :

* , , chain rule:

Therefore:

1. Start with some .
2. *Repeat*:

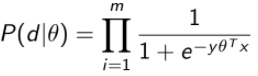
is convex, and has a single minimum. Iteration moves until it reaches the minimum.

**Probabilistic Interpretation of Logistic Regression**:

The label takes only values or .

Assume

The **likelihood**  of the training data is:



Taking logs:

And the maximum likelihood estimate of minimises:

since .

is our estimate of our confidence in the prediction. When it is close to , we are confident in out prediction.