Probabilistic analysis: Quick reference

vishvAs vAsuki

December 2, 2011

Based on [1]. Also see complex analysis reference sheet for important functions and approximations.

Contents

Part I

Probability Theory

1 Sample space and Events

1.1 Sample/ state space

1.1.1 States/ sample points

Consider the output of a stochastic system. This output can be considered to be characterized as the state of a system. If two outputs are distinguishable from one another, then they can be interpreted as correspond to different states of the system.

1.1.2 Specification with state variables

Two states or outputs are distinguishable by differing assignments (x_i) to certain features or variables (X_i).

1.1.3 Sample space notation

Sample space of a statistical experiment: Ω or S.

1.2 Events and Sigma algebra

1.2.1 Events

A certain event is a set of outputs, which share some common characteristic. Usually, this shared characteristic is concise: Eg: fixing a certain variable $X_i = x_i$ specifies an event. Eg: Sun rises.

Specification using Iverson brackets, indicator functions See algebra ref. Eg:[Sun rises]. They can also be specified naturally using indicator random variables.

1.2.2 Sigma algebra

Sigma algebra F over Ω: $\bigcup$ closed set of sets (events) which includes $\emptyset$ and Ω.

Inclusion of negation If event $A \in \Omega$, so is $\neg A = \Omega - A$.

Representation by variables Each $S \in F$ can be represented by the values assigned to a variable $X = x$.

1.3 Space of events

(Ω, F) describes a space of events.

1.3.1 Product probability space

Take spaces $(S_1, F_1), (S_2, F_2)$; thence get product space $(S_1 \times S2, F_1 \times F_2)$.

1.3.1.1 Over the same sample space S

If $S_1 = S_2 = S$, product space is just $(S, F_1 \times F_2)$, as the $(s_i, s_j) \forall j \neq i$ will have 0 probability measure, if defined over that set.

It is useful to think of this as including events $f_i \cap g_j$ and $f_i \cup g_j$ for $f_i \in F_1, g_j \in F_2$.

1.4 Counting

1.4.1 Count action sequence

Multiplicative principle. If action A can be made in a ways, and action B can be made in b ways, there are ab ways of doing action AB.

1.4.2 'Picking' phenomena: varieties

Many phenomena are modeled as picking k items from a set of n items. The picking may or mayn't be done with repetition. The order of picking may or mayn't matter (the corresponding phenomena are called permutation/ arrangement and choosing/ combination respectively).

1.4.2.1 Permutation

Any permutation is writable as a set of directed cycles in a graph whose nodes are the items chosen. So, counting is equivalent to counting the number of 'legal' cycles.
With repetition, using the multiplicative principle, we have n^k permutations, without repetition, we have $n..(n-k+1)$ permutations. When $n = k$, we use the notation $n!$.

Dealing with identical items Suppose, there are t identical items of a kind among the k selected. In this case, as these t items can be indistinguishably arranged in $t!$ ways, the previous argument cannot be used.

1.4.2.2 Combinations without repetition

Here, order does not matter. Observing that, for each selection of k items, there are $k!$ permutations, we see that the number of combinations is $\binom{n}{k} = n..(n-k+1)/k!$
Useful property: $\binom{n}{r} = \binom{n}{n-r}$. For extension to real n and general properties, see complex analysis ref.

1.4.2.3 Choosing with repetition

$\binom{n+k-1}{k-1}$ choices. [**Proof**]: select places to put k-1 | in $n+k-1$ spots. □
Error alert: One may try to use the argument, as in the case of choosing without repetition, that for each selection of k items, there are $k!$ permutations and conclude that there are $\frac{n^k}{k!}$ choices. But this is false (too low) because, when there are repetitions, each selection of k items corresponds to fewer than $k!$ items. △

1.4.3 Combinatorial proofs of equality

Take a process, count it in two ways, conclude that expressions corresponding to the two counting procedures are equal.

1.4.4 Cardinality of sets

General properties concerning set sizes are considered elsewhere.

2 Probability of events

Example applications, modeling intricacies, various interpretations of probability are discussed in the probabilistic modeling survey.

2.1 Probability measure

The following axiomatization is common to both frequentist and subjective interpretation of probability.
Take sample space S, sigma algebra F. The probability measure is a special measure $v : F \to [0, 1]$; so operates over sets, like CDF. Additionally, the general additivity property (described in algebra survey) is usually assumed.
Other properties of the measure: should be countably additive over disjoint sets: σ additive; $v(\Omega) = 1$.
So, v specifies event probability. (S, F, v) is called a probability space.

2.1.1 Importance

Viewing probability as the measure of a set of events makes many notions [even the simple Union bound] much more intuitive!

2.1.2 Visualization: An area of atomic events

Use a spotted 2-d compact set, whose area represents the randomness (cross product of all distributions) involved in the probability; the spots represent an event of interest.

2.1.3 Subscript notation

Consider the subscript in $Pr_{X \in_D S}(E)$. In this case, the subscript indicates the variables which need to be specified in order to specify a point in the sample space; in doing so, it tells us where the randomness lies. It also tells us something about the range S of the random variable X, and its distribution D.

2.1.3.1 Other notations

Also often used: $Pr_X(E), Pr_D(E)$, or $Pr_t(E)$
if the distribution D is parametrized by t.

2.1.3.2 Importance

This notation/ representation is very valuable; using it, we can clearly manipulate and reason about probability quantities, seeing for example how the sample space shrinks as we consider condition probability distributions.

2.1.4 Empirical measure

$v_n(E) = n^{-1}\sum_i I_E(x_i)$, where I is the indicator function. This is useful in deducing the actual measure v using experiments/ sampling.

2.2 Conditional and unconditional probabilities

Conditional (posterior) probability $Pr(E_1|E_2)$ considers the evidence that E_1 has occurred. This conditioning alters the measure, so that $Pr(E_1|E_1) = 1$, and $\frac{Pr(E_1,E_2)}{Pr(E_1)} = Pr(E_2|E_1)$ (aka product rule). So, the sample space S can be thought of as now being constricted to E_1.
The unconditioned measure $Pr(E_1)$ is called prior (marginal) probability.

2.2.1 Common errors

2.2.1.1 Equal weight error

Instead of calculating $\frac{Pr(E_1,E_2)}{Pr(E_1)}$, one common error is to use unweighted counts: $\#(E_1 \cap E_2)/\#(E_1)$, which leads to the wrong result. Hence, this should be avoided and proper formalism used.
See examples provided later.

2.2.1.2 Misidentified prior error

Another common problem is the misidentification of the prior event with another, which leads to a different weight being assigned to the probabilities involved. This can lead to the equal weight error.
See examples provided later.

2.2.1.3 Illustrations

Example: Warden problem. Of 3 prisoners ($P1, P2, P3$) scheduled to be executed, one is pardoned. The identity of the spared prisoner is known only to the warden. $P1$ tries to find out about his fate.

On being pressed, the warden, reasoning that he is not leaking any information relevant to $P1$, only says that $P2$ is executed. But, $P1$ is now happy that the probability of his being pardoned is increased from 1/3 to 1/2.
The warden is correct and $P1$ is wrong. Reason follows.
Let Pi also represent event where Pi is pardoned. Let W represent the event where warden tells $P1$ that $P2$ is being executed. Now, $Pr(P1) = 1/3$. We want to find $Pr(P1|W)$. $Pr(W) = \sum_i Pr(W \cap Pi) = 1/6 + 0 + 1/3 = 1/2$, and $Pr(P1 \cap W) = 1/6$, so $Pr(P1|W) = 1/3$. So, $P1$ has learned nothing about his fate.
Another source of error in this example is confusing W with event $P2$. ▲
Example: Monty Hall problem. In a game show conducted by Monty Hall, there are 3 doors (() $P1, P2, P3$), one of which has a reward. Only Monty Hall knows where. A player chooses a door, say $P1$. Monty Hall opens one of the other doors, say $P2$, and reveals it to contain no reward. Should the player switch to $P3$?
Same rigorous reasoning as in the case of Warden problem can be applied to reveal that he should switch. The source of errors are also the same. ▲

2.2.2 Independence of events

E_1 and E_2 are independent if the $Pr(E_2) = Pr(E_2|E_1)$: so, the evidence E_1 does not change the probability measure as applied to E_2. This is same as saying: $Pr(E_1 \wedge E_2) = Pr(E_1)Pr(E_2)$.

2.3 Properties of the measure

Important properties such as the inclusion/ exclusion principle, union and intersection measure bounds follow from those described for general measures.

2.3.1 Connection with expectation

Consider the measure v. $Pr(E_i) = E_v[I_{E_i}(x)]$.

2.4 Probability with Multiple variables/ sigma algebras

Consider the product $(S_1 \times S_2, F_1 \times F_2, v)$ of the probability spaces (S_i, F_i, v_i) for $i \in 1, 2$. Consider the specific event $E \in F_1$.

2.4.1 The product measure

The resulting product measure v always obeys the following constraints:
$\forall E \in F_1 : v(E, S_2) = v_1(E)$. A symmetrical condition holds for all $G \in F_2$.

2.4.1.1 Marginalization

Aka Law of total probability, marginalization. $Pr(E) = \sum_{G \in F_2} Pr(E \wedge G)$: like adding up rows in a table of joint probabilities.

2.4.2 Conditional probability inversion

Aka Bayes' theorem, Bayes's rule.
$Pr(G|E) = \frac{Pr(E|G)Pr(G)}{Pr(E)}$
$= \frac{Pr(E|G)Pr(G)}{\sum_{G \in F_2} Pr(E|G)Pr(G)}$.
Fixing E, this becomes a function $F_2 \to [0,1]$, we can write $Pr(G|E) =\propto Pr(E|G)Pr(G)$.

2.4.2.1 Likelihood function

What is the likelihood of a hypothesis E given the evidence G? We can use $f_G(E) = Pr(G|E)$, a function of E alone, as a measure of this.
For use in statistical inference, see statistics ref.

2.5 Associated quantities

2.5.1 Odds and log odds

Odds: $Pr(E_1)/Pr(\neg E_1)$. Thence is defined log odds or logit. In logistic models, this is modeled, rather than the probability itself.

Part II

Random variables

3 Random variable (RV) X

3.1 Map sample space to measurable space

Consider the probability space $(S, \sigma(S), v)$ and a measurable space R with an associated measure μ and sigma algebra G, aka state space.
$X : S \to R$, where X is a $(\sigma(S), \sigma(R))$ measurable function is a random variable (RV). To emphasize the (sigma algebra membership) structure preserving properties, we write: $X : (S; F, v) \to R$.
Note that $(R, \sigma(R))$ is usually $(\Re, B)$, where B is the union and complement closure of the set of (semi)open intervals.
So, the correct way to write a RV is: $X(o) = x$, value of X over $o \in S$.

3.2 Induced Probability measure Pr

X induces a probability measure over the space $(R, \sigma(R))$. This is aka probability distribution.
Never write $Pr(X)$***, but*** $Pr(X = x)$ ***is fine!***

3.2.1 Probability density function (pdf) wrt measure m

Consider a measure m (such that $Pr << m$) over the measurable space $(R, \sigma(R))$. The pdf wrt m is the inter-measure (Radon/ Nikodym) derivative between (v, m), if it exists.
So, it is any function f such that $Pr(X \in E) = \int_E f(x)dm = \int_{X^{-1}(E)} dv$.
Note that f has the property that
$\int_R f(x)dm = Pr(X \in R) = 1$.

3.2.1.1 Notation

The pdf associated with a random variable X is often denoted by f_X.

3.2.1.2 Not probability measure

Note that f is not a probability measure: Together with m, it only helps specify Pr. Specifically, $f(x) \neq Pr(X = x)$ in general.
Note that $f(x) \to \infty$ as $x \to t$ is possible - but this would be impossible for a probability.

3.2.1.3 Probability mass function (pmf)

Consider the case where range(X) is discrete. Then, the pdf $f(x) = Pr(X = x)$ when used with the counting measure. Such a pdf is called a pmf.

3.2.1.4 Support

$\{x : f_X(x) \neq 0\}$ support of the distribution of X.

3.2.1.5 Improper densities

Aka pseudo-density. Sometimes, the pdf is specified in a form which does not sum to 1. $f_X(x) = \frac{p_X(x)}{Z}$, where the constant Z is not specified.

3.2.1.6 In terms of cdf

Derivation trick for some pdf's defined in terms of CDF's:
$\int_{-\infty}^{y} F(x)^a f(x)dx = \frac{F(y)^a}{a}$.

3.2.2 Cumulative density functions (CDF)

Take any real valued RV. $F(x) = Pr(X \in [-\infty, x])$. If the pdf exists, this is $\int_{-\infty}^{x} f(x)dx$. Can by itself describe distribution - pdf need not exist; but important for describing continuous distributions.

3.2.2.1 Notation

The CDF associated with a random variable X is often denoted by F_X.

3.2.2.2 Properties

The CDF is monotonically increasing. It is right continuous.

3.2.2.3 Connection to discreteness

X is discrete (ie its range is discrete) iff F_X is a step function.
X is continuous iff F_X is continuous.

3.2.2.4 Multidimensional case

Just $F(x) = \int_{(-\infty)^d}^{x} f(x)dx$ if $x \in R^d$!

3.2.2.5 Quantiles

points taken at regular intervals from CDF. Types: Percentiles, deciles etc..

3.2.3 Entropy

See information and coding theory ref.

3.3 Importance

Random variables allow us to express probability measures simply using pdf's and pmf's.
Furthermore, they allow us to study models where an underlying random process (probability space) results in observations in a different space (the range of a random variable).

3.4 Random variable for probability space

For analyzing arbitrary probability spaces using properties/ notation of random variables, one can simply add on a measurable space R' and a measurable function X: R' could even be the probability space itself!

3.4.1 Indicator RV

The indicator function corresponding to the event set can be used as a 2-range random variable: see algebra ref.

3.5 RV from a map

Let X be an RV. Consider a measurable function $h : ran(X) \to ran(Y)$. $Y = h(X)$ is a random variable itself.

3.5.1 Monotonic maps

Let h be monotonic.

3.5.1.1 CDF

If h is increasing: $F_X(x) = F_Y(h(x))$.
If h is decreasing: $F_Y(h(x)) = 1 - F_X(x)$.

3.5.1.2 pdf

Consider $f_Y(y) = \frac{dF_y(y)}{dy} = \frac{dF_x(h^{-1}(y))}{dy}$. By chain rule, and using the CDF relationships from earlier, $f_Y(y) = f_X(h^{-1}(y)) \left|\frac{dh^{-1}(y)}{dy}\right|$.

3.5.1.3 Utility

This is useful in sampling complex distributions by transforming random variables with easy to sample distributions. The monotonicity is useful because we often 'stretch' parts of range(X) to form range(Y) in order to arrive at the more complex distribution of Y.

4 Multiple random variables

4.1 Random vector

A random vector is an n-dim vector $X = (A_i)$, which are a bunch of jointly distributed random variables. Similarly, X can be a $m \times n$ random matrix.
Below, we consider $X = (X_1, X_2)$, where $X_1 : (S_1; F_1, v_1) \to R_1$ and $X_2 : (S_2; F_2, v_2) \to R_2$.
A random vector is itself a random variable $X : (S_1 \times S_2; F_1 \times F_2, v) \to (R_1 \times R_2)$.

4.1.1 Marginalization

The marginalization properties of the joint/ product probability space leads to: $Pr(X_1 \in E_1, X_2 \in S_2) = Pr(X_1 \in E_1)$, so
$\int_{E_1 \times S_2} f_X(x)dv = \int_{E_1} \int_{x_2 \in S_2} f_X(x_1, x_2)dv_2 dv_1 = \int_{E_1} f_{X_1}(x_1)dv_1$.
Hence, $\int_{x_2 \in S_2} f_X(x_1, x_2)dv_2 = f_{X_1}(x_1)$.

4.2 Conditional pdf

Definition (described elsewhere)
of conditional probabilities of the form $Pr(A|B)$ breaks down if $Pr(B)$, the probability measure of the event B is 0.
One can craft a similar definition to cover events $X_2 = b$ with $v_2(X_2 = b) = 0$.
Then,
$Pr(X_1 \in E_1|X_2 = b) = \frac{Pr(X_1 \in E_1 \land X_2 = b)}{f_{X_2}(b)} =$
$\int_{E_1} f_X(x_1, b) f_{X_2}(b)^{-1} dv_1$.
$f_X(x_1, b) f_{X_2}(b)^{-1} = f_{X_1|X_2=b}(x_1)$ is aka conditional pdf.

4.2.1 Inversion

Similar to the Bayes's rule, using the definition, one can invert the conditional pdf.
$f_{X_2|X_1=x_1}(x_2) = \frac{f_{X_1|X_2=x_2}(x_1) f_{X_2}(x_2)}{f_{X_1}(x_1)} = \frac{f_{X_1|X_2=x_2}(x_1) f_{X_2}(x_2)}{\int_{S_2} f_{X_1|X_2=x_2}(x_1) f_{X_2}(x_2) dv_2}$.

4.2.1.1 Improper densities

Note that the construction of $f_{X_2|X_1=x_1}(x_2)$ works even if the prior pdf $f_{X_1}(x_1)$ is an improper density which does not sum to 1! This sometimes makes the task of modeling random processes easier.

4.3 Independence

One can extend the notion of independence of events to random variables, which represent a pair of algebras of events.
Suppose that $f_X(x) = f_{X_1}(x_1) f_{X_2}(x_2)$. Then, $\forall E_1, E_2 : Pr(X_1 \in E_1, X_2 \in E_2) = Pr(X \in E_1) Pr(X_2 \in E_2)$. In such a case, X_1 and X_2 are independent. This is denoted by $I(X_1, X_2)$.
Also independence of events corresponds to independence of corresponding Indicator random variables: $A \perp B$ if $I_A \perp I_B$.

4.4 Conditional Independence

Conditional:
$X \perp Y|Z \equiv f_{XY|Z}(x, y|z) = f_{X|Z}(x|z) f_{Y|Z}(y|z) \equiv f_{X|Y,Z}(x|y, z) = f_{X|Z}(x|z)$

$\equiv f_{X,Y,Z}(x,y,z) = \frac{f_{X,Z}(x,z)f_{Y,Z}(y,z)}{f_Z(z)}$.
Marginal: $X \perp Y$ when $Z = \phi$.
Amongst sets of vars: $\{X_i\} \perp \{Y_i\} \,|\, \{Z_i\}$ iff
$f_{(X_i|Y_j,\{Z_k\})}(x_i|y_j,\{z_k\}) = f_{(X_i|\{Z_k\})}(x_i|\,\{z_k\})\forall i,j$.
Marginal independence without conditional independent:
$X \nvdash Y|X+Y$. Conditional independent sans marginal independent: consider suitable Bayesian network.
Graphical models can be used to specify this.

5 Averaging using the pdf

Consider the real valued random variable $X : (S, B) \to (R, B_r, m)$, whose pdf is f_X defined relative to the reference measure m.

5.1 Mean/ Expectation of real valued RV

Aka Expected value. $E : \{RV\} \to R$. $E[X] = \mu = \int_X x f_X(x)dm = E_X[X]$.
This is the weighted average of $range(X)$. $E[X]$ is actually a convex combination of points in range(X).

5.1.1 Subscript notation

See probability section.

5.1.2 Conditional Expectation

Conditional expectation of X wrt event A: $E_X[X|A]$ is computed using the conditional pdf $f_{X|A}(x)$. Sometimes, this is considered as a function of variable A.
For events with non-0 probability measure, this is $= \frac{E[XI_A(X)]}{Pr(A)}$.

5.2 Expectation: Properties

$E_Y[E_X[X|Y]] = E_Y[f(Y)] = E_X[X]$.

5.2.1 Connection to probability measure

$Pr(B) = E_X[Pr(B|X)]$.
If X is an indicator RV, E[X] = Pr(X = 1).

5.2.2 Products of independent RV

If expectations are finite: $I(X,Y) \equiv E_{X,Y}[XY] = E_X[X]E_Y[Y]$ as $I(X,Y) \equiv f_{X,Y}(x,y) = f_X(x)f_Y(y)$.

5.2.3 Linearity properties

$E[k] = k$.

5.2.3.1 Linearity in X

This follows from the linearity of integration.
$E[\sum_i X_i] = \int_X (\sum_i x_i) f_X(x) dm = \sum_i \int_X x_i f_X(x) dm$
$= \sum_i \int_{X_i} x_i f_{X_i}(x) dm = \sum_i E[X_i]$
Expectation is linear, even if the summed RVs are dependent.
E[X] is convex!

5.2.3.2 Importance

Powerful, unintuitive! 10 folks go to a ghoShTi, leave their hats; retrieve some random hat after the ghoShTi. What is the expected number of people to retrieve the hat they came with? Linearity greatly simplifies calculation.

5.2.3.3 Linearity in pdf

Consider $E[X]$, where $f_X(x)$ is a convex combination of $(f_i(x))$ with coefficients (a_i). Because of the linearity of integration, $E[X]$ is linear in the pdf components: $E[X] = \sum_i a_i E_{f_i}[X]$.

5.2.4 Convex function of E[] inequality

Aka Jensen's inequality.
If f is convex, $E[f(X)] \geq f(E[X])$: [**Proof**]: E[X] is actually a convex combination of points in range(X). So, follows directly from definition of convexity (see vector spaces ref). □
So: $E[X^2] - (E[X])^2 \geq 0$.

5.3 Expectation: Analysis tricks

If X and Y are not independent, fix the factor in X and Y which causes the dependence, and ye may have independence.
The fact that $\max X \geq E[X]$ is useful in getting lower bounds on some quantities.

5.4 Variance from mean

$Var[X] = \sigma^2 = E[X - E[X]]^2 = E[X^2] - (E[X])^2$. Weighted avg of square deviation from mean of function over points in sample space.

5.4.1 Concavity in p

Consider discrete distributions; let x be a vector of values X takes. $var[X] = p^T x - x^T pp^T x$.

5.4.2 Other properties

$Var[\sum a_i X_i] = \sum a_i^2 Var[X_i] + 2\sum_i \sum_j a_i a_j Cov[X_i, X_j]$: using linearity of expectations on $E[(\sum_i (a_i X_i - a_i \mu_i))^2]$.

5.4.2.1 Translation invariance

var[X + c] = var[X].

5.5 Moments of RV X

5.5.1 kth moment

$E[X^k]$ is the kth Moment of X. Empirical kth moment is $\frac{\sum_{i=1}^n X_i^k}{n}$

5.5.2 kth moment about the mean

kth Moment of X about 0 : $E[X^k]$ vs moment of X about μ aka central moment: $E[(X-\mu)^k]$.
Normalized moments: $\frac{E[(X-\mu)^k]}{\sigma^k}$.

5.5.3 Important moments

Central moment immune to translation; describes shape of pdf. 2nd central moment, aka variance, measures fatness.
Skewness: $\gamma = \frac{E[X-\mu]^3}{\sigma^3}$: putting more weight on farther points; pdf has left skew if $\gamma < 0$.
Kurtosis: measure tallness/ leanness vs shortness/ squatness: $\frac{E[X-\mu]^4}{\sigma^4} - 3$: 3 term ensures that Normal distribution has Kurtosis 0.
Can easily determine statistics corresponding to these parameters from a sample.
The moments of X completely describe pdf of X [**Find proof**]. $N(\mu, \sigma^2)$ has only 2 moments [**Check**].

5.5.4 Moment generating function

$M_X(t) = E[e^{tX}]$.
$\frac{dE[f(x,t)]}{dt} = \frac{d\int f(x,t)dx}{dt} = lt_{\delta\int(f(x,t+\delta t)-f(x,t))dx\to 0\,\frac{}{\delta t}} = E[\frac{df(x,t)}{dt}]$. So, use to find nth moment of X about 0, $E[X^n] = d^n M_X(t)/(dt)^n|_{t=0}$: can also use Taylor series for e^{tX} with linearity of expectations.

It is unique: If $M_X(t) = M_Y(t)$, X and Y have same distribution: as it generates all possible moments of X and Y identically. [**Find proof**]
$M_{\sum X_i} = E[e^{t\sum X_i}] = \prod M_{X_i}$ if $\{X_i\} \perp$.

5.5.5 Characteristic function of X

$f_X = E_X[e^{itX}]$. Useful as sometimes, the moment generating function is not well defined.

5.5.6 For Poisson trials

$M_{X_i}(t) = 1+p_i(e^t-1) \leq e^{p_i(e^t-1)}$. So, if $X = \sum X_i, \mu = \sum p_i; M_X \leq e^{\mu(e^t-1)}$.

6 Random Vector properties

6.1 Mean

$E[X] := (E[X_i])$.

6.1.1 Linearity

If X is a random matrix, A, B, C are constant matrices: $E[AXB + C] = AE[X]B + C$. Proof: by using $(AXB)_{i,j} = A_{i,:}XB_{:,j}$, which is a linear combination of $X_{k,l}$.
Also, if X is random vector, $E[a^TX] = a^TE[X]$.

6.2 Covariance

6.2.1 Definition

How correlated are deviations of X, Y from their means?
$cov(X, Y) = E_{x,y}[(X - E[X])(Y - E[Y])]$.

6.2.1.1 Extension to vectors

$cov(X, Y) = E_{x,y}[(X - E[X])^T(Y - E[Y])]$:
corresponds to measuring $cov(X_i, Y_j)$.

6.2.2 Correlation

(Pearson) correlation coefficient: $corr(X, Y) = \frac{cov(X,Y)}{\sigma_X\sigma_Y}$: normalized covariance.

6.2.2.1 Correlation vs Independence

If $X_i \perp X_j, Cov[X_i, X_j] = 0$: even if they are only pairwise independent. But, $cov(X, X^2) = 0$ even if (X, X^2) not $\perp$.

If $Cov[X_i, X_j] = 0$ holds, then Xi and Xj are uncorrelated. If they are independent, they are uncorrelated; but not necessarily vice versa.

6.3 Covariance matrix

$\Sigma = var[X] = cov(X, X) = E[(X - \mu)(X - \mu)^T] = E[XX^T] - \mu\mu^T$.
Diagonal has variances of (X_i). It is diagonal if (X_i) are independent.

6.3.1 Effect of linear transformation

$Var[BX + a] = E[(BX - BE[X])(BX - BE[X])^T] = BVar[X]B^T$. As in the scalar case, constant shifts have no effect.
Special case: $var[a^T X] = a^T var(X)a$.

6.3.2 Nonnegative definiteness

$\Sigma \succeq 0$ as $a^T E[(X - \mu)(X - \mu)^T]a \geq 0$.
If $a^T \Sigma a = 0$, with probability 1, $a^T X - a^T \mu = 0$, so some $\{X_i\}$ are linearly dependent. So, X lies on the hyperplane/ subspace with normal a.

6.3.3 Precision matrix

$V = \Sigma^{-1}$. Consider partial correlation deduced in case of multidimensional normal distribution.

6.3.4 Moment generating function

$E[e^{t^T X}]$ is the moment generating function.

7 Random variable sequence

(X_i) with CDF (F_i).

7.1 Convergence in distribution to X

Aka weak convergence. If $\forall x : \lim_{n \to \infty} F_n(x) = F(x)$, then $X_n \to^d X$. Comments about limit of CDF's.

7.2 Convergence in probability to X

If $\forall \epsilon : \lim_{n \to \infty} Pr(|X_n - X| > \epsilon) = 0$, say $X_n \to^p c$; so limit of sequence of probabilities. Probability of deviation from X grows smaller and smaller, but doesn't necessarily hit 0. Eg: Weak law of large numbers.
Implies convergence in distribution.

7.3 Almost sure convergence

$Pr(\lim_{n\to\infty}(X_n = X)) = 1$: Note that lim is inside Pr(); so limit of sequence of boolean events. Eventually, X_n will behave exactly like X. Eg: Animal's daily consumption will some day hit 0 and stay 0.
Implies convergence in probability.

7.4 Sure convergence

All X_n are over exactly same sample space Ω. $\forall w \in \Omega : \lim_{n\to\infty} X_n(w) = X(w)$. Implies all other forms of convergence.

Part III

Inference and comparison of distributions

8 Statistics

See statistics ref.

8.1 Estimating a probability with accuracy ϵ

See statistics ref.

9 Distances between distributions

9.1 Total variation distance between distributions

Aka Statistical distance. Sample space X. $\Delta(D, D') = 2^{-1}\sum_{x\in X}|D(x) - D'(x)|$: $\in [0,1]$. But, $\sum_{x\in X}(D(x) - D'(x)) = 0$.
Visualize as space between probability curves. Total prob under either curve is 1.

9.1.1 Largest deviation in event probability

For event $E \subseteq X : \max_{E\subseteq X}|Pr_D(x \in E) - Pr_{D'}(x \in E)| = \Delta(D, D')$. Or, max (signed) area between curves covered by E is at most half the total area. Useful in bounding probability of events.

9.2 Code-length divergence

(Kullback Leibler) Aka information divergence, information gain, relative entropy. A particular Bregman divergence. General case specified in vector spaces ref. For connection with entropy and cross entropy, see information theory ref. $K(D||D') = E_{x\sim D}[\log \frac{D(x)}{D'(x)}] =$
$\sum D(x)\log \frac{D(x)}{D'(x)} = \sum D(x)\log \frac{1}{D'(x)} - H(D) = H_c(D') - H(D)$. Expected number of extra bits used to code samples in D using code based on D'.

9.2.1 Nonnegativity

See wiki diagram: Puts greater weight D(x), often for cases where $\frac{D(x)}{D'(x)} \geq 1$.
$K(D, D') \geq 0$ (aka Gibbs inequality): Take probability distributions p, q; get $\sum p_i \log(p_i/q_i) \geq 0$ using $\ln x \leq x - 1$. $K(D||D') = 0$ only if $D = D'$ using same idea.

9.2.2 Other properties

Not a metric as it is asymmetric and does not satisfy the triangular inequality.
$\exists x : D(x) \neq 0, D'(x) = 0 :\implies K(D||D') = \infty$.

9.2.3 Connection with variation distance

(Pinsker's inequality) $\sqrt{KL(P||Q)/2} \geq \Delta P, Q$.[**Find proof**]

10 Inferences about distributions of function(RV)

Y = g(X).
Use $Pr(g(X) \in A) = Pr(X \in g^{-1}(A))$. So, given CDF, PDF of X, can deduce CDF of g(X) and thence derive PDF of g(X).

10.1 Using $\frac{dg^{-1}(Y)}{dY}$

If g is monotone in $(x, x+\delta x)$: $p_X(x)\delta x \approx p_Y(y)\delta y$, taking $(x, x+\delta x)$ to $(y, y+\delta y)$ using g: So $p_Y(y) = p_X(x)|\frac{dx}{dy}| = p_X(g^{-1}(y))|\frac{dg^{-1}(y)}{dy}|$: so maximum probability density changes with variable change.
If g is not continuous, but $\exists$ partition $A_0, ..A_k$ with $Pr(X \in A_0) = 0$, with $\{g_i\} = g$ over $\{A_i\}$ monotone; then $p_Y(y) = \sum_i p_X(g^{-1}(y))|\frac{dg_i^{-1}(y)}{dy}|$; where $\sum$ appears to account for the probability that Y=y over various domains of X.

10.1.1 Extension to multidimensional distributions

$Y = g(X_1, X_2)$;
$X_1 = h(Y, X_2)$. Fix $X_2 = x_2$; get $p(Y, x_2) = p_{X_1,X_2}(X_1$

$= h^{-1}(Y, x_2)|x_2)|\frac{dh^{-1}(Y,x_2)}{dY}|$; then do $p_Y(y) = \int p(Y, x_2)dx_2$.

10.2 Using moment generating functions

Given $m_X(t)$, find
$m_Y(t) = E[e^{f(X)t}]$; thence deduce pdf of Y.

11 Bounds on deviation probability

Aka concentration of measure inequalities.

11.1 Expectation based deviation bound

(Aka Markov's inequality). If $X \geq 0$: $Pr(X \geq a) \leq \frac{E[X]}{a}$: $Pr(X \geq a)$ is max when X is 0 or a.
Averaging argument. If $X \leq k$, $c\mu Pr(X \leq c\mu) + k(1 - Pr(X \leq c\mu)) \geq \mu$; so $Pr(X \leq c\mu) \leq \frac{k-\mu}{k-c\mu}$.
This technique is used repeatedly in other deviation bounds based on variance and moment generating functions.

11.2 Variance based deviation bound

(Aka Chebyshev's inequality). By Markov's inequality: $Pr((X - E[X])^2 \geq a^2) \leq \frac{Var[X]}{a^2}$.

11.2.1 Use in estimation of mean

$Pr(n^{-1}(\sum X_i - E[X_i])^2 \geq a^2) = Pr((\sum X_i - E[X_i])^2 \geq na^2) \leq \frac{Var[\sum X_i]}{na^2}$.
Applicable for pair-wise independent Bernoulli trials.

11.3 Exponentially small deviation bounds

11.3.1 General technique

(Chernoff) $Pr(e^{tX} \geq e^{ta}) \leq E[e^{tX}]/e^{ta}$: applying Markov. Used to bound both $Pr(X > a)$ and $Pr(X < a)$ with $t > 0$ or $t < 0$. Get a bound exponentially small in μ, deviation.

11.3.2 For random variable sequences

$\mu = \sum E[X_i]$. For $X = \sum_{i=1}^{n} X_i$. Note that RVs are not necessarily identically distributed.

11.3.2.1 Pairwise independent RVs

Use variance based deviation bounds, as variance of pairwise independent RVs is an additive function.

11.3.3 Sum of n-wise independent RVs

11.3.3.1 Bounds from MGF's.

$Pr(e^{tX} \geq e^{ta}) \leq E[e^{tX}]/e^{ta} = (\prod E[e^{tX_i}])/e^{ta}$: here ye have used independence.
If $d > 0$, $Pr(X \geq (1+d)\mu) \leq \frac{e^{\mu(e^t-1)}}{e^{t(1+d)\mu}} \leq \frac{e^{d\mu}}{(1+d)^{(1+d)\mu}}$: using $t = \ln(1+d)$ and M_X bound.
So, if $R = (1+d)\mu > 6\mu : d = \frac{R}{\mu} - 1 \geq 5, Pr(X \geq (1+d)\mu) \leq (\frac{e}{6})^R \leq 2^{-R}$.
If d in (0,1], $Pr(X \geq (1+d)\mu) \leq e^{\frac{-\mu d^2}{3}}$: As $\frac{e^d}{(1+d)^{(1+d)}} \leq 2^{\frac{-d^2}{3}}$: as $f(d) = d - (1+d)\ln(1+d) + \frac{d^2}{3} \leq 0$: as $f(0) \leq 0$ and $f'(d) < 0$.
If d in (0,1], $Pr(X \leq (1-d)\mu) < \frac{e^{-d\mu}}{(1-d)^{(1-d)\mu}}$; $Pr \leq e^{\frac{-\mu d^2}{2}}$.
So, $Pr(|X - \mu| \geq d\mu) \leq 2e^{-\mu d^2/3}$. ***So, probability of deviation from mean decreases exponentially with deviation from mean.***
Can be used in both additive and multiplicative forms.

11.3.3.2 Goodness of empirical mean

Now, $E[X_i] = p$. Using $X/n = \sum X_i/n$ to estimate mean p. So, $Pr(|\frac{\sum X_i}{n} - p| \geq dp) \leq 2e^{-npd^2/3}$. ***So, probability of erroneous estimate decreasing exponentially with number of samples!***

11.3.3.3 Code length divergence bound

Let D_p and D_q be probability distributions of binary random variables with probabilities p and q of being 1 respectively.
$D_p(\sum_i X_i \geq qn) \leq (n - qn)e^{-nKL(D_p||D_q)}$.
[**Proof**]: Suppose that $X_i \sim D_p$ and that $p < q, k \geq qn$.
$D_p(\sum_i X_i = k) \leq \frac{D_p(\sum_i X_i = k)}{D_q(\sum_i X_i = k)} = (\frac{p}{q})^k(\frac{1-p}{1-q})^{n-k} \leq (\frac{p}{q})^{qn}(\frac{1-p}{1-q})^{n(1-q)} = e^{-nKL(D_p||D_q)}$.
So, taking the union bound over all $k \geq qn$, we have the result. □
Using the connection between the code length divergence and the total variation distance: $KL(D_p||D_q) \geq 2(p-q)^2$. This can be used to derive other deviation bounds.

11.3.3.4 Additive deviation bounds

See Azuma inequality section.

11.3.4 iid RV: Tightness of the Chernoff bound

(Cramer) Take $l(a) = \max_t ta - M(a)$. For large n: $Pr(\frac{\sum X_i}{n} \geq a) \geq e^{-n(l(a)+\epsilon)}$ [**Find proof**]. Combining with Chernoff, $Pr(\frac{\sum X_i}{n} \geq a) = e^{-n(l(a)+\epsilon_n)}$ for some seq $(\epsilon_n) \to 0$.

Part IV

Probabilistic Analysis Techniques

12 Existence proofs

$Pr(X \geq E[X]) > 0$, $Pr(X \leq E[X]) > 0$.

12.1 For sparse dependency graphs

Aka Lovasz local lemma.
For Dependency graph with $Pr(E_i) < p, 4dp < 1$: $Pr(\cap \bar{E}_i) > 0$.
Lovasz local lemma: general case: Dependency. graph G=(V,E),
$x_i \in [0,1]$: $Pr(E_i) \leq x_i \prod_{(i,j)\in E}(1-x_j)$: $Pr(\cap \bar{E}_i) \geq \prod_i (1-x_i) > 0$

12.2 Threshold behavior

$X > 0$: Second moment method: $Pr(X = 0) \leq Pr((X - E[X])^2 \geq (E[X])^2)$.
Conditional expectation inequality for $\sum$ indicators:
$Pr(X > 0) \geq \sum_{i=1}^n \frac{Pr(X_i=1)}{E[X|X_i=1]}$.

12.3 Explicit constructions

[**Incomplete**]

12.4 Make Existence proofs

Design sample space, show $Pr(E) > 0$, maybe modify to get final object. Use Expectation argument. Make non-negative RV X, use Chebyshev to bound Pr(X=0). Make dependency graph, use Lovasz local lemma.

13 Extremal combinatorics

Prove extremal statistic about some extremal set. [**Incomplete**]

14 Analysis strategies

14.1 General strategies

Pick things randomly. Be able to specify the random process.
Analyzing X and Y; If there is uniform symmetry in X, set X to be any value without loss of generality.
Cast the problem into a stochastic process : Eg: Markov chain/ Random walk problem, Martingale.

14.2 Bound probabilities and expectations

14.2.1 Break up big events into smaller cases

One can analyze $Pr(A)$ using $Pr(A) = \sum_b Pr(A|G)Pr(G)$.

14.2.1.1 One of many events

Use the Union bound.

14.2.1.2 Co-occurring events

$Pr(\wedge_{i \in S} E_i) = Pr(E_1)Pr(E_2|E_1)Pr(E_3|E_2, E_1)...$ Taking conditional probabilities, one analyzes events with smaller sample spaces. If this decomposition is done considering the 'causation', as in the case of directed graphical models, we can take advantage of conditional independence.
Or one can use $Pr(\wedge_{i \in S} E_i) \leq Pr(E_i)$ for some suitable i.

14.2.1.3 Principle of deferred decisions

Find or lower bound $Pr(F(X))$ using principle of deferred decisions: Let RVs $X_1 \ldots X_n$ decide X's value: suppose $X_1 \ldots X_{n-1}$ happen: What is the probability that X_n takes the right value for $F(X)$?

14.2.2 Identify independent events

Identifying independent events often helps in bounding probabilities: whether in being able to apply suitable tail bounds, or in decomposing $E[XY]$ or $var[X+Y]$ or $Pr(XY)$.
To do this, one can consider causation by constructing a graphical model; or identify it algebraically by keeping track of the sample space connected with the expectation/ probability/ variance using suitable subscripts.

14.2.3 Use super-events

If $E_1 \implies E_2$, then $Pr(E_1) \leq Pr(E_2)$.

14.2.4 Use concentration of measure around the mean

Find mean, use tail bounds.

14.2.4.1 Means and variances

In finding mean, we can often use properties such as the linearity of expectation. Similarly, we can decompose the problem of finding $var[X+Y]$ or $E[XY]$ the random variables involved have suitable independence properties.
If it is difficult to find the mean directly (Eg: $E[X] = \sum E[X_i]$), one can find an upper bound to it by other means, and use it.

14.2.4.2 Dealing with lack of independence

If you're analyzing a probability distribution of $f(X_1, ..X_i)$ or $\sum X_i$, and $\{X_i\}$ are not independent, use a martingale. If $\exists$ pairwise independence, use Chebyshev.

15 Results

Max load Y when hash function from k-universal family used: $Pr(Y > \sqrt[k]{2n}) < 2^{-1}$ (bounding expected number of collisions, use Markov).

Bibliography

[1] *Probability and Computing: Randomized Algorithms and Probabilistic Analysis.* Cambridge University Press, 2005.