0.1 Frequentists vs. Bayesians

প্রবাবিলিটি বা সম্ভাব্যতা কাকে বলে?

একদিকে প্রবাবিলিটিকে ফ্রিকোয়েন্টিস্ট এর আলোকে ব্যাখ্যা করা হয়; যেখানে সম্ভাব্যতা (probability) কোনো ঘটনার দীর্ঘমেয়াদি পুনরাবৃত্তির হারকে বোঝায়। উদাহরণস্বরূপ, আমরা যদি একটা কয়েনকে অনেকবার ছুঁড়ি, তবে ধারণা করা হয় এটি প্রায় অর্ধেক সময় "হেডস" পড়বে।

প্রবাবিলিটির আরেকটা ব্যাখ্যা দাড় করানো হয় বায়েসিয়ান(Bayesian) ব্যাখ্যার ভিত্তিতে। এই ব্যাখ্যায়, সম্ভাব্যতা আমাদের কোনো ঘটনার প্রতি অনিশ্চয়তা বোঝাতে ব্যবহৃত হয়; অর্থাৎ, এটি পুনরাবৃত্তি করা পরীক্ষার উপর নির্ভর না করে ডাটা বা ইনফর্মেশনের সঙ্গে সম্পর্কিত। বায়েসিয়ান সংজ্ঞার ভিত্তিতে কয়েন পরবর্তী বার ছুড়ে মারলে "হেডস" বা "টেলস" পড়ার সমান সম্ভাবনা রয়েছে।

বায়েসিয়ান ব্যাখ্যার একটা বড় সুবিধা হল, এটি এমন সব ঘটনার অনিশ্চয়তা(uncertainty) মডেল করতে পারে যেগুলোর পুনরাবৃত্তি নাও হতে পারে। উদাহরণস্বরূপ, আমরা ২০২০ সালের মধ্যে মেরু বরফ গলে যাবে কিনা তা নিয়ে সম্ভাব্যতা নির্ণয় করতে চাই; এই ঘটনা ঘটলে সর্বোচ্চ একবার হতে পারে বা একদম নাও হতে পারে; কিন্তু বারবার হবে না। কিন্তু তবুও এরকম শূন্য/একবার ঘটে যাওয়া ঘটনার অনিশ্চয়তা নির্ণয় করতে হতে পারে। মেশিন লার্নিং ভিত্তিক আরেকটা ঘটনার আলোকে ব্যাপারটা ব্যাখ্যা করা যাক। ধরি, আমরা রাডারে একটি "ব্লিপ" দেখেছি এবং এর ভিত্তিতে আমরা লক্ষ্যবস্তুর অবস্থান (যা হয়তো একটি পাখি, বিমান বা ক্ষেপণাস্ত্র হতে পারে) সম্পর্কে probability distribution নির্ণয় করতে চাই। এই ক্ষেত্রে, পুনরাবৃত্তি করা পরীক্ষার ধারণা প্রাসঙ্গিক নয়, কিন্তু বায়েসিয়ান ব্যাখ্যা স্বাভাবিক এবং যথাযথ।

এই বইয়ে আমরা বায়েসিয়ান ব্যাখ্যার ভিত্তিতেই সব আলোচনা করব।

0.2 A brief review of probability theory

ধরি কোনো অজানা পরিমাণকে নির্দেশ করছে, যেমন একটা লুডুর ডাইস গড়ালে সেটা কোন দিক পড়বে তা বের করতে হবে। এরকম একটা random ঘটনার সম্ভাব্য ফলাফল বোঝাতে X চিহ্নের ব্যবহার করা হয়; যেখানে X ডিসক্রিট(Discrete) বা কন্টিনিউয়াস(Continuous) হতে পারে।

0.2.1 Basic concepts

Discrete random variable: X একটা সীমিত(finite) গণনাযোগ্য অসীম সেট(countably infinite set) থেকে মান গ্রহণ করে। যেমন কয়েন ছুড়ে টস করার পর কতবার হেডস এসেছে সেটা ডিসক্রিট নাম্বার (10,13,78,..)

Continuous random variable: X এর মান একটি নির্দিষ্ট সীমার মধ্যে যেকোনো বাস্তব সংখ্যা (Real numbers) হবে। যেমন একটা স্থানের মানুষদের উচ্চতা কন্টিনিউয়াস (5.3 ft, 6.0 ft)

0.2.1.1 CDF: cumulative distribution function

একটা random variable X এর ভিন্ন মান পাওয়ার প্রবাবিলিটি ডিস্ট্রিবিউশন কেমন হবে সেটা জানার জন্যে CDF ব্যবহার করা যায় যাকে () এর মাধ্যমে প্রকাশ করা হয়।

$$F(x) \triangleq P(X \leq x) = \begin{cases} \sum_{u \leq x} p(u) & \text{, discrete} \\ \int_{-\infty}^{x} f(u)\mathrm{d}u & \text{, continuous} \end{cases} \tag{0.1}$$

- $P(X \leq x)$ X -এর মান x-কম বা সমান হওয়ার সম্ভাবনা।
- $p(u)$ হচ্ছে ডিসক্রিট ভ্যারিয়েবেলের জন্যে প্রবাবিলিটি মাস ফাংশন (Probability Mass Function, PMF), যা u এর একটি নির্দিষ্ট মান পাওয়ার সম্ভাব্যতা প্রকাশ করে
- $f(u)$ হচ্ছে কন্টিনিউয়াস ভ্যারিয়েবেলের জন্যে প্রবাবিলিটি ডেনসিটি ফাংশন (probability density function), u এর সম্ভাব্যতা।

0.2.1.2 PMF and PDF

PMF: Probability Mass Function

Random Variable X এর নির্দিষ্ট মান পাওয়ার সম্ভাবনা কতটুকু, সেটা নির্ধারন করা হয় Probability Mass Function (PMF) এর মাধ্যমে। উদাহরণস্বরূপ, যদি একটি ৬-পাশের ডাইস ফেলা হয় , প্রবাবিলিটি মাস ফাংশন (Probability Mass Function, PMF) প্রতিটি পাশের সম্ভাবনা নির্দেশ করবে, যার মান ১ থেকে ৬ এর মধ্যে আসবে।

- PMF, $p(x) = P(X = x)$ হিসাবে প্রকাশ করা হয়, যা নির্দেশ করে random ভেরিয়েবল X একটি নির্দিষ্ট মান x নেওয়ার সম্ভাবনা।
- বৈশিষ্ট্য:
  - $0p(x)1$ (সম্ভাবনা ০ এবং ১-এর মধ্যে থাকে)
  - $\sum_{x \in X} p(x) = 1$ (সব সম্ভাবনার যোগফল ১ হয়)

PDF: Probability Density Function

Probability Density Function (PDF) continuous random ভেরিয়েবলের probability density নির্দেশ করে। এটা ব্যবহার করা হয় একটি নির্দিষ্ট সীমার মধ্যে সম্ভাবনা বের করার জন্য।

- $P(a \leq X \leq b) = \int_a^b f(x)\,dx$ (এখানে $f(x)$ হল PDF, যা probability density নির্দেশ করে)
- বৈশিষ্ট্য:
  - $f(x) \geq 0$ (density শূন্য বা তার বেশি হয়)
  - $\int_{-\infty}^{\infty} f(x)\,dx = 1$ (পুরো density value ১ হয়)

0.2.2 Mutivariate random variables

Joint Distribution দুইটি random ভ্যারিয়েবলের একসাথে ঘটার সম্ভাবনাকে জয়েন্ট প্রবাবিলিটি বলে, যেমন একটা ডাইসের 2 পড়ার এবং কয়েনের হেডস পড়ার সম্ভাবনা।

Marginal Distribution অন্য ভ্যারিয়েবলের মান বিবেচনা না করে, একটা random ভ্যারিবেলের নির্দিষ্ট মান আসার সম্ভাবনা,যেমন ডাইসে ২ আসার সম্ভাবনা, কয়েন টসের ফল কি হবে সেটা বিবেচনা না করে। ডিসক্রিট এর ক্ষেত্রে -

$$P(X=x)=\sum_{y}P(X=x,Y=y) \tag{0.2}$$

কনটিনউয়াস এর ক্ষেত্রে -

$$P(X=x)=\int_{-\infty}^{+\infty}f(x,y)\,dy \tag{0.3}$$

0.2.2.1 Joint CDF

দুটি random ভ্যারিয়েবল X এবং Y এর জন্যে CDF-

$$F(x,y)\triangleq P(X\leq x\cap Y\leq y)=P(X\leq x,Y\leq y)$$

$$F(x,y)\triangleq P(X\leq x,Y\leq y)=\begin{cases}\sum_{u\leq x,v\leq y}p(u,v) & \text{(discrete)}\\ \int_{-\infty}^{x}\int_{-\infty}^{y}f(u,v)\mathrm{d}u\mathrm{d}v & \text{(continuous)}\end{cases} \tag{0.4}$$

- $F(x,y)$ হচ্ছে joint CDF, যেখানে X এবং Y এর মান x এবং y এর কম বা সমান হওয়ার সম্ভাবনা কত নির্ধারন করে ।
- $\sum_{u\leq x,v\leq y}p(u,v)$ হচ্ছে ডিসক্রিট দুইটি র্যান্ডম ভেরিয়েবলের PMF
- $\int_{-\infty}^{x}\int_{-\infty}^{y}f(u,v)\,du\,dv$ হচ্ছে কন্টিনিউয়াস ভ্যারিয়েবেলের ক্ষেত্রে pdf ইন্টিগ্রেশন।

0.2.2.2 Product Rule

দুটি ঘটনা একসাথে ঘটার সম্ভাবনাকে প্রকাশ করা যায় Product Rule এর মাধ্যমে। প্রোডাক্ট রুলকে কন্ডিশনাল প্রবাবিলিটি ($P(X,Y)$) এবং মার্জিনাল প্রবাবিলিটি ($P(Y)$) এর গুণফলের মাধ্যমে প্রকাশ করা হয়।

$$p(X,Y)=P(X|Y)P(Y) \tag{0.5}$$

- $P(X\cap Y)$ বোঝায় X এবং Y একসাথে ঘটার সম্ভাবনা।
- $P(X\mid Y)$ বোঝায় X ঘটার সম্ভাবনা, শর্ত হল Y ঘটে গেছে
- $P(Y)$ হলো Y ঘটার সম্ভাবনা।

0.2.2.3 Chain Rule

চেইন রুল প্রোডাক্ট রুল এর একটি সম্প্রসারণ যা একাধিক ঘটনার সম্ভাবনাকে একত্রে প্রকাশ করে। কমপ্লেক্স প্রবাবিলিটি বা মেশিন লার্নিং (backpropagation) এর ক্ষেত্রে চেইন রুল ব্যবহার করা হয়।

$$p(X_{1:N})=p(X_1)p(X_2|X_1)p(X_3|X_2,X_1)...p(X_N|X_{1:N-1}) \tag{0.6}$$

- এখানে n সংখ্যক ঘটনার একসাথে ঘটার সম্ভাবনা নির্ণয় করতে আমরা প্রতিটি ঘটনার কন্ডিশনাল প্রবাবিলিটি গুণ করছি

0.2.2.4 Marginal Distribution

0.2.2.5 Marginal Distribution

মার্জিনাল ডিস্ট্রিবিউশন ভ্যারিয়েবেলের সেট থেকে কেবলমাত্র একটা ভ্যারিয়েবলকে ফোকাসে রেখে তার প্রবাবিলিটি ডিস্ট্রিবিউশন নির্ণয় করে অন্য সকল ভ্যারিয়েবল বাদ রেখে। যেমন, random variable X এর marginal CDF নির্ণয় করার সময় X এর মান শুধুই x এর সমান বা কম বিবেচনা করা হবে, Y এর মান যাই থাকুক না কেন।

Marginal CDF for X: Discrete case continuous case এর জন্যে marginal CDF-

$$F_X(x)\triangleq F(x,+\infty)=\begin{cases}\sum_{x_i\leq x}P(X=x_i)=\sum_{x_i\leq x}\sum_{j=1}^{+\infty}P(X=x_i,Y=y_j)\\ \int_{-\infty}^{x}f_X(u)du=\int_{-\infty}^{x}\int_{-}^{+\infty}f(u,v)dudv\end{cases} \tag{0.7}$$

- প্রথম ক্ষেত্রে, যখন random ভেরিয়েবল discrete হয়, তখন আমরা X এর সকল x_i মানের জন্য $P(X=x_i)$ সম্ভাবনা নির্ণয় করে summation বের করছি, যেখানে $x_i<_x$; দ্বিতীয় summation এর ক্ষেত্রে, Y এর সকল মানের জন্যে পাওয়া জয়েন্ট প্রবাবিলিটি $P(X=x_i,Y=y_j)$ sum up করছে।
- দ্বিতীয় ক্ষেত্রে, যখন র্যান্ডম ভেরিয়েবল continuous হয়, তখন আমরা joint probability density function $f(u,v)$ কে ইন্টিগ্রেট করছি প্রথমে v এর সাপেক্ষে (অর্থাৎ Y এর সম্ভাব্য সকল মান নিয়ে), অতপর u এর সাপেক্ষে । ।

Marginal CDF for Y:

$$F_Y(y)\triangleq F(+\infty,y)=\begin{cases}\sum_{y_j\leq y}P(Y=y_j)=\sum_{i=1}^{+\infty}\sum_{y_j\leq y}P(X=x_i,Y=y_j)\\ \int_{-\infty}^{y}f_Y(v)dv=\int_{-\infty}^{+\infty}\int_{-\infty}^{y}f(u,v)dudv\end{cases} \tag{0.8}$$

- প্রথম ক্ষেত্রে, আমরা Y এর মান $y_j\leq y$ এর জন্য $P(Y=y_j)$ এর সম্ভাবনা বের করি। এই সম্ভাবনাটি $X=x_i$ এবং $Y=y_j$ এর joint probability নির্ণয় করে, এবং এটি সব $y_j\leq y$ এর প্রবাবিলিটি sum up করে।
- দ্বিতীয় ক্ষেত্রে, যখন Y একটি continuous random ভেরিয়েবল হয়, তখন আমরা joint probability definitions function $f(u,v)$ ইন্টিগ্রেট করে $Y\leq y$ এর সম্ভাবনা বের করি। এটি সমস্ত X এর জন্য এবং Y এর নির্দিষ্ট মান পর্যন্ত সম্ভাবনাকে ইন্টিগ্রেট করে।

Marginal PMF and PDF:

$$\begin{cases} P(X=x_i)=\sum_{j=1}^{+\infty} P(X=x_i,Y=y_j) & \text{, discrete} \\ f_X(x)=\int_{-\infty}^{+\infty} f(x,y)\mathrm{d}y & \text{, continuous} \end{cases} \tag{0.9}$$

$$\begin{cases} p(Y=y_j)=\sum_{i=1}^{+\infty} P(X=x_i,Y=y_j) & \text{, discrete} \\ f_Y(y)=\int_{-\infty}^{+\infty} f(x,y)\mathrm{d}x & \text{, continuous} \end{cases} \tag{0.10}$$

0.2.2.6 Conditional distribution

একটা ঘটনার প্রভাবে আরেকটি ঘটনা ঘটার সম্ভাব্যতাকে conditional distribution দ্বারা প্রকাশ করা হয়। যদি $Y=y_j$ হয়, তবে $X=x_i$ হওয়ার সম্ভাবনাকে প্রকাশ করা যায় - Conditional PMF:

$$p(X=x_i|Y=y_j)=\frac{p(X=x_i,Y=y_j)}{p(Y=y_j)} \text{ if } p(Y)>0 \tag{0.11}$$

- $\frac{p(X=x_i,Y=y_j)}{p(Y=y_j)}$, XY এর জয়েন্ট প্রবাবিলিটি
- $p(Y=y_j)$, Y এর মার্জিনাল প্রবাবিলিটি

The pmf $p(X|Y)$ is called conditional probability.
Conditional PDF:

$$f_{X|Y}(x|y)=\frac{f(x,y)}{f_Y(y)} \tag{0.12}$$

- $p(X=x_i,Y=y_j)$, XY এর জয়েন্ট প্রবাবিলিটি
- $p(Y=y_j)$, Y এর মার্জিনাল প্রবাবিলিটি

0.2.3 Bayes rule

কন্ডিশনাল প্রবাবিলিটির মধ্যে সম্পর্ক প্রকাশ করার জন্যে Bayes Rule ব্যবহার করা হয় যেখানে নতুন ইনফর্মেশনের ভিত্তিতে একটা ঘটনার প্রববিলিটিকে আপডেট করা হবে। মেশিন লার্নিং এ, বিশেষত প্রোবাবিলিস্টিক মডেল যেমন Naive Bayes ক্লাসিফায়ার এই নিয়ম ব্যবহার করা হয়।

$$\begin{aligned} p(Y=y|X=x) &= \frac{p(X=x,Y=y)}{p(X=x)} \\ &= \frac{p(X=x|Y=y)p(Y=y)}{\sum_{y'} p(X=x|Y=y')p(Y=y')} \end{aligned} \tag{0.13}$$

- $p(X=x,Y=y)$ হল $Y=y$ হওয়ার সম্ভাবনা, যখন $X=x$ হবে ; এখানে $p(X=x,Y=y)$ কন্ডিশনাল প্রবাবিলিটি এবং $p(X=x)$ হল মার্জিনাল প্রবাবিলিটি ।
- দ্বিতীয় সমীকরণে, আমরা প্রোডাক্ট রুল ব্যবহার করেছি, যেখানে $p(X=x|Y=y)$ যখন $Y=y$ হবে তখন $X=x$ হওয়ার সম্ভাবনা এবং $p(Y=y)$ হলো X এর কোন ইনফরমেশন ছাড়া মার্জিনাল প্রবাবিলিটি ।
- ডিনোমিনেটরে থাকা $\sum_{y'} p(X=x|Y=y')p(Y=y')$ হলো নরমালাইজেশন ফ্যাক্টর যেখানে $X=x$ এর জন্য Y-এর সম্ভাব্য মান নিয়ে summation করা হয় ।

0.2.4 Independence and conditional independence

We say X and Y are unconditionally independent or marginally independent, denoted $X \perp Y$, if we can represent the joint as the product of the two marginals, i.e.,

$$X \perp Y = P(X,Y)=P(X)P(Y) \tag{0.14}$$

We say X and Y are conditionally independent(CI) given Z if the conditional joint can be written as a product of conditional marginals:

$$X \perp Y|Z = P(X,Y|Z)=P(X|Z)P(Y|Z) \tag{0.15}$$

0.2.5 Quantiles

Since the cdf F is a monotonically increasing function, it has an inverse; let us denote this by F^{-1}. If F is the cdf of X , then $F^{-1}(\alpha)$ is the value of x_α such that $P(X \le x_\alpha)=\alpha$; this is called the α quantile of F. The value $F^{-1}(0.5)$ is the median of the distribution, with half of the probability mass on the left, and half on the right. The values $F^{-1}(0.25)$ and $F^{1}(0.75)$are the lower and upper quartiles.

0.2.6 Mean and variance

The most familiar property of a distribution is its mean,or expected value, denoted by μ. For discrete rv's, it is defined as $\mathbb{E}[X] \triangleq \sum_{x\in\mathcal{X}} xp(x)$, and for continuous rv's, it is defined as $\mathbb{E}[X] \triangleq \int_{\mathcal{X}} xp(x)\mathrm{d}x$. If this integral is not finite, the mean is not defined (we will see some examples of this later).

The variance is a measure of the "spread" of a distribution, denoted by σ^2. This is defined as follows:

$$var[X] = \mathbb{E}[(X-\mu)^2] \tag{0.16}$$
$$= \int (x-\mu)^2 p(x)\mathrm{d}x$$
$$= \int x^2 p(x)\mathrm{d}x + \mu^2 \int p(x)\mathrm{d}x - 2\mu \int x p(x)\mathrm{d}x$$
$$= \mathbb{E}[X^2] - \mu^2 \tag{0.17}$$

from which we derive the useful result

$$\mathbb{E}[X^2] = \sigma^2 + \mu^2 \tag{0.18}$$

The standard deviation is defined as

$$std[X] \triangleq \sqrt{var[X]} \tag{0.19}$$

This is useful since it has the same units as X itself. https://en.wikipedia.org/wiki/Standardized_moment

0.3 Some common discrete distributions

In this section, we review some commonly used parametric distributions defined on discrete state spaces, both finite and countably infinite.

0.3.1 The Bernoulli and binomial distributions

Definition 0.1. Now suppose we toss a coin only once. Let $X \in \{0,1\}$ be a binary random variable, with probability of "success" or "heads" of θ. We say that X has a Bernoulli distribution. This is written as $X \sim \mathrm{Ber}(\theta)$, where the pmf is defined as

$$\mathrm{Ber}(x|\theta) \triangleq \theta^{\mathbb{I}(x=1)}(1-\theta)^{\mathbb{I}(x=0)} \tag{0.20}$$

Definition 0.2. Suppose we toss a coin n times. Let $X \in \{0,1,\cdots,n\}$ be the number of heads. If the probability of heads is θ, then we say X has a binomial distribution, written as $X \sim \mathrm{Bin}(n,\theta)$. The pmf is given by

$$\mathrm{Bin}(k|n,\theta) \triangleq \binom{n}{k}\theta^k(1-\theta)^{n-k} \tag{0.21}$$

0.3.2 The multinoulli and multinomial distributions

Definition 0.3. The Bernoulli distribution can be used to model the outcome of one coin tosses. To model the outcome of tossing a K-sided dice, let $\vec{x} = (\mathbb{I}(x=1),\cdots,\mathbb{I}(x=K)) \in \{0,1\}^K$ be a random vector(this is called dummy encoding or one-hot encoding), then we say X has a multinoulli distribution(or categorical distribution), written as $X \sim \mathrm{Cat}(\theta)$. The pmf is given by:

$$p(\vec{x}) \triangleq \prod_{k=1}^{K} \theta_k^{\mathbb{I}(x_k=1)} \tag{0.22}$$

Definition 0.4. Suppose we toss a K-sided dice n times. Let $\vec{x} = (x_1,x_2,\cdots,x_K) \in \{0,1,\cdots,n\}^K$ be a random vector, where x_j is the number of times side j of the dice occurs, then we say X has a multinomial distribution, written as $X \sim \mathrm{Mu}(n,\vec{\theta})$. The pmf is given by

$$p(\vec{x}) \triangleq \binom{n}{x_1\cdots x_k}\prod_{k=1}^{K}\theta_k^{x_k} \tag{0.23}$$

where $\binom{n}{x_1\cdots x_k} \triangleq \frac{n!}{x_1!x_2!\cdots x_K!}$

Bernoulli distribution is just a special case of a Binomial distribution with $n=1$, and so is multinoulli distribution as to multinomial distribution. See Table ?? for a summary.

Table 0.1: Summary of the multinomial and related distributions.

Name	K	n	X
Bernoulli	1	1	$x \in \{0,1\}$
Binomial	1	-	$\vec{x} \in \{0,1,\cdots,n\}$
Multinoulli	-	1	$\vec{x} \in \{0,1\}^K, \sum_{k=1}^K x_k = 1$
Multinomial	-	-	$\vec{x} \in \{0,1,\cdots,n\}^K, \sum_{k=1}^K x_k = n$

0.3.3 The Poisson distribution

Definition 0.5. We say that $X \in \{0,1,2,\cdots\}$ has a Poisson distribution with parameter $\lambda > 0$, written as $X \sim \mathrm{Poi}(\lambda)$, if its pmf is

$$p(x|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!} \tag{0.24}$$

The first term is just the normalization constant, required to ensure the distribution sums to 1.

The Poisson distribution is often used as a model for counts of rare events like radioactive decay and traffic accidents.

0.3.4 The empirical distribution

The empirical distribution function[1], or empirical cdf, is the cumulative distribution function associated with the empirical measure of the sample. Let $\mathcal{D} = \{x_1, x_2, \cdots, x_N\}$ be a sample set, it is defined as

$$F_n(x) \triangleq \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(x_i \leq x) \tag{0.25}$$

0.4 Some common continuous distributions

In this section we present some commonly used univariate (one-dimensional) continuous probability distributions.

0.4.1 Gaussian (normal) distribution

If $X \sim N(0,1)$,we say X follows a standard normal distribution.

The Gaussian distribution is the most widely used distribution in statistics. There are several reasons for this.

1. First, it has two parameters which are easy to interpret, and which capture some of the most basic properties of a distribution, namely its mean and variance.
2. Second,the central limit theorem (Section TODO) tells us that sums of independent random variables have an approximately Gaussian distribution, making it a good choice for modeling residual errors or "noise".
3. Third, the Gaussian distribution makes the least number of assumptions (has maximum entropy), subject to the constraint of having a specified mean and variance, as we show in Section TODO; this makes it a good default choice in many cases.
4. Finally, it has a simple mathematical form, which results in easy to implement, but often highly effective, methods, as we will see.

See (Jaynes 2003, ch 7) for a more extensive discussion of why Gaussians are so widely used.

0.4.2 Student's t-distribution

where $\Gamma(x)$ is the gamma function:

$$\Gamma(x) \triangleq \int_0^\infty t^{x-1} e^{-t} \mathrm{d}t \tag{0.26}$$

μ is the mean, $\sigma^2 > 0$ is the scale parameter, and $\nu > 0$ is called the degrees of freedom. See Figure ?? for some plots.

The variance is only defined if $\nu > 2$. The mean is only defined if $\nu > 1$.

As an illustration of the robustness of the Student distribution, consider Figure ??. We see that the Gaussian is affected a lot, whereas the Student distribution hardly changes. This is because the Student has heavier tails, at least for small ν(see Figure ??).

If $\nu = 1$, this distribution is known as the Cauchy or Lorentz distribution. This is notable for having such heavy tails that the integral that defines the mean does not converge.

To ensure finite variance, we require $\nu > 2$. It is common to use $\nu = 4$, which gives good performance in a range of problems (Lange et al. 1989). For $\nu \gg 5$, the Student distribution rapidly approaches a Gaussian distribution and loses its robustness properties.

0.4.3 The Laplace distribution

Here μ is a location parameter and $b > 0$ is a scale parameter. See Figure ?? for a plot.

Its robustness to outliers is illustrated in Figure ??. It also put mores probability density at 0 than the Gaussian. This property is a useful way to encourage sparsity in a model, as we will see in Section TODO.

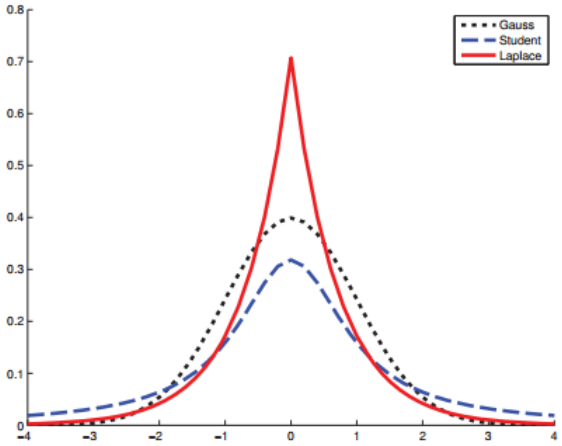
0.4.4 The gamma distribution

Here $a > 0$ is called the shape parameter and $b > 0$ is called the rate parameter. See Figure ?? for some plots.

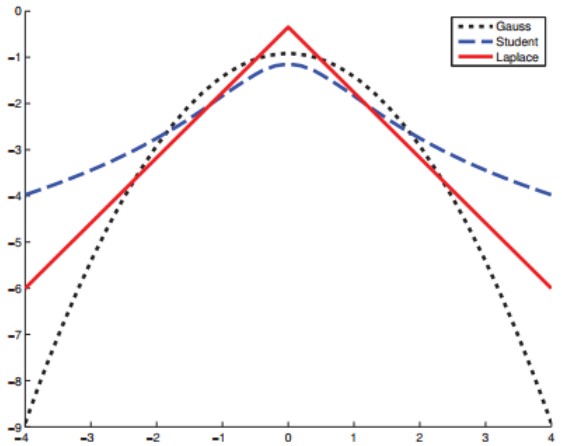
[1] http://en.wikipedia.org/wiki/Empirical_distribution_function

Table 0.2: Summary of Bernoulli, binomial multinoulli and multinomial distributions.

Name	Written as	X	$p(x)$(or $p(\vec{x})$)	$\mathbb{E}[X]$	var$[X]$
Bernoulli	$X \sim \text{Ber}(\theta)$	$x \in \{0,1\}$	$\theta^{\mathbb{I}(x=1)}(1-\theta)^{\mathbb{I}(x=0)}$	θ	$\theta(1-\theta)$
Binomial	$X \sim \text{Bin}(n,\theta)$	$x \in \{0,1,\cdots,n\}$	$\binom{n}{k}\theta^k(1-\theta)^{n-k}$	$n\theta$	$n\theta(1-\theta)$
Multinoulli	$X \sim \text{Cat}(\vec{\theta})$	$\vec{x} \in \{0,1\}^K, \sum_{k=1}^K x_k = 1$	$\prod_{k=1}^K \theta_j^{\mathbb{I}(x_j=1)}$	-	-
Multinomial	$X \sim \text{Mu}(n,\vec{\theta})$	$\vec{x} \in \{0,1,\cdots,n\}^K, \sum_{k=1}^K x_k = n$	$\binom{n}{x_1 \cdots x_k}\prod_{k=1}^K \theta_j^{x_j}$	-	-
Poisson	$X \sim \text{Poi}(\lambda)$	$x \in \{0,1,2,\cdots\}$	$e^{-\lambda}\frac{\lambda^x}{x!}$	λ	λ

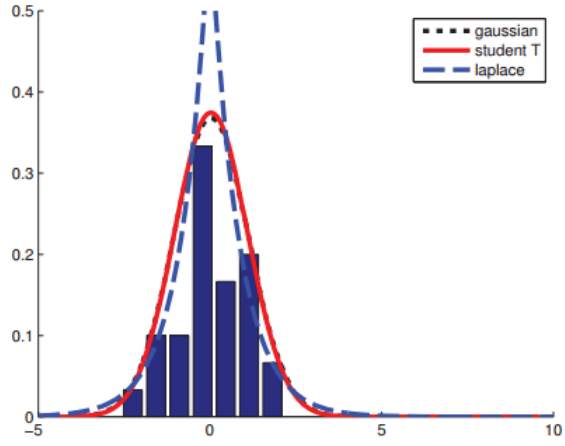


(a)

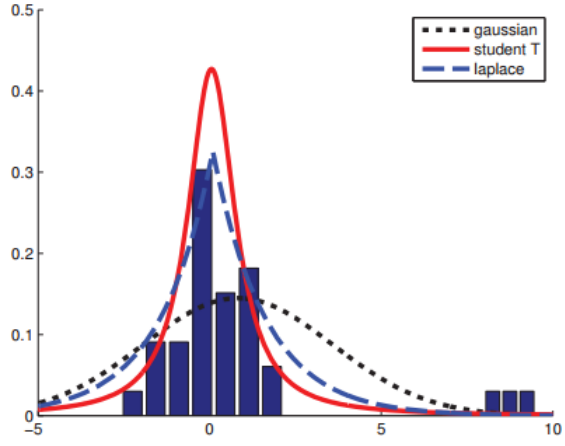


(b)

Figure 0.1: (a) The pdf's for a $\mathcal{N}(0,1)$, $\mathcal{T}(0,1,1)$ and $Lap(0,1/\sqrt{2})$. The mean is 0 and the variance is 1 for both the Gaussian and Laplace. The mean and variance of the Student is undefined when $\nu = 1$.(b) Log of these pdf's. Note that the Student distribution is not log-concave for any parameter value, unlike the Laplace distribution, which is always log-concave (and log-convex...) Nevertheless, both are unimodal.



(a)



(b)

Figure 0.2: Illustration of the effect of outliers on fitting Gaussian, Student and Laplace distributions. (a) No outliers (the Gaussian and Student curves are on top of each other). (b) With outliers. We see that the Gaussian is more affected by outliers than the Student and Laplace distributions.

Table 0.3: Summary of Gaussian distribution.

Written as	$f(x)$	$\mathbb{E}[X]$	mode	$\text{var}[X]$
$X \sim \mathcal{N}(\mu, \sigma^2)$	$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$	μ	μ	σ^2

Table 0.4: Summary of Student's t-distribution.

Written as	$f(x)$	$\mathbb{E}[X]$	mode	$\text{var}[X]$
$X \sim \mathcal{T}(\mu, \sigma^2, \nu)$	$\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})} \left[1 + \frac{1}{\nu}\left(\frac{x-\mu}{\nu}\right)^2\right]$	μ	μ	$\frac{\nu\sigma^2}{\nu-2}$

Table 0.5: Summary of Laplace distribution.

Written as	$f(x)$	$\mathbb{E}[X]$	mode	$\text{var}[X]$
$X \sim \text{Lap}(\mu, b)$	$\frac{1}{2b}\exp\left(-\frac{\lvert x-\mu\rvert}{b}\right)$	μ	μ	$2b^2$

Table 0.6: Summary of gamma distribution

Written as	X	$f(x)$	$\mathbb{E}[X]$	mode	$\text{var}[X]$
$X \sim \text{Ga}(a, b)$	$x \in \mathbb{R}^+$	$\frac{b^a}{\Gamma(a)} x^{a-1} e^{-xb}$	$\frac{a}{b}$	$\frac{a-1}{b}$	$\frac{a}{b^2}$

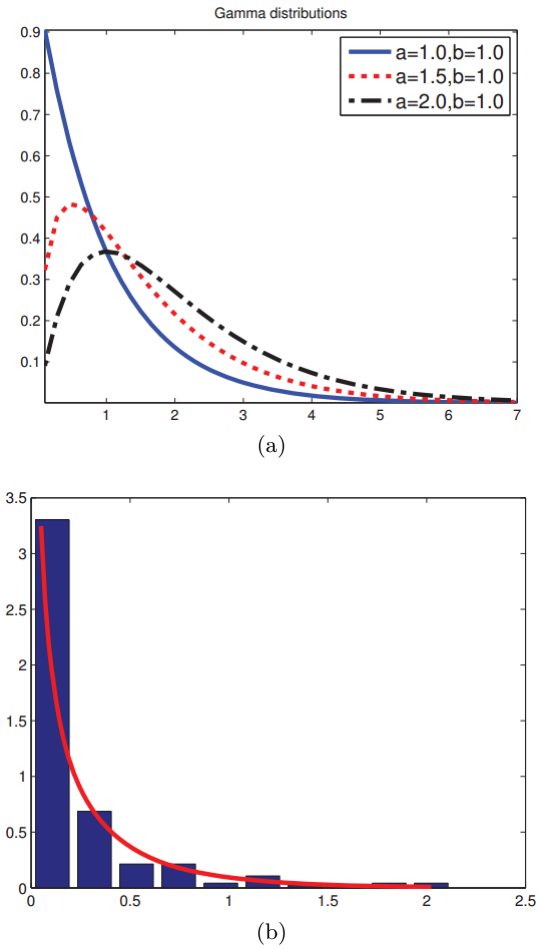


Figure 0.3: Some $\text{Ga}(a, b = 1)$ distributions. If $a \leq 1$, the mode is at 0, otherwise it is > 0.As we increase the rate b, we reduce the horizontal scale, thus squeezing everything leftwards and upwards. (b) An empirical pdf of some rainfall data, with a fitted Gamma distribution superimposed.

0.4.5 The beta distribution

Here $B(a, b)$is the beta function,

$$B(a, b) \triangleq \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \tag{0.27}$$

See Figure ?? for plots of some beta distributions. We require $a, b > 0$ to ensure the distribution is integrable (i.e., to ensure $B(a, b)$ exists). If $a = b = 1$, we get the uniform distirbution. If a and b are both less than 1, we get a bimodal distribution with "spikes" at 0 and 1; if a and b are both greater than 1, the distribution is unimodal.

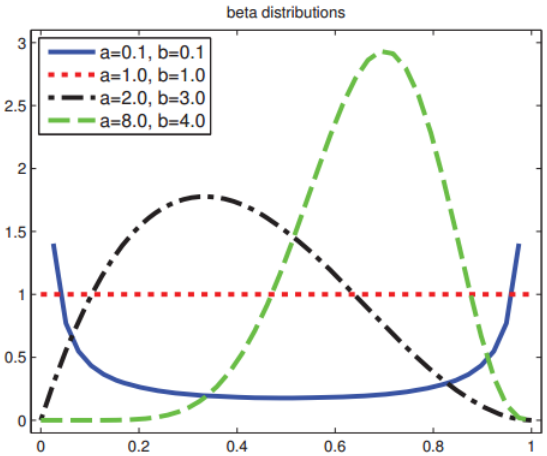


Figure 0.4: Some beta distributions.

0.4.6 Pareto distribution

The Pareto distribution is used to model the distribution of quantities that exhibit long tails, also called heavy tails.

As $k \to \infty$, the distribution approaches $\delta(x - m)$. See Figure ??(a) for some plots. If we plot the distribution on a log-log scale, it forms a straight line, of the form $\log p(x) = a \log x + c$ for some constants a and c. See

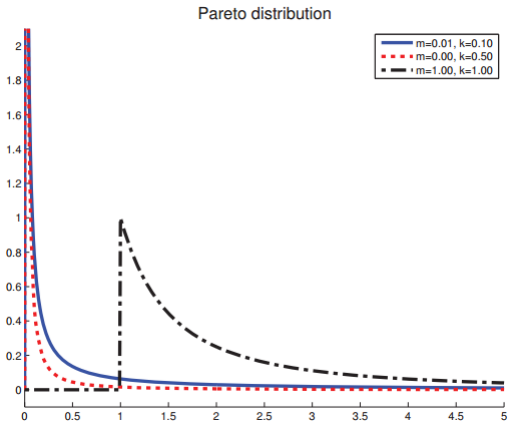
Table 0.7: Summary of Beta distribution

Name	Written as	X	$f(x)$	$\mathbb{E}[X]$	mode	var$[X]$
Beta distribution	$X \sim \text{Beta}(a,b)$	$x \in [0,1]$	$\frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}$	$\frac{a}{a+b}$	$\frac{a-1}{a+b-2}$	$\frac{ab}{(a+b)^2(a+b+1)}$

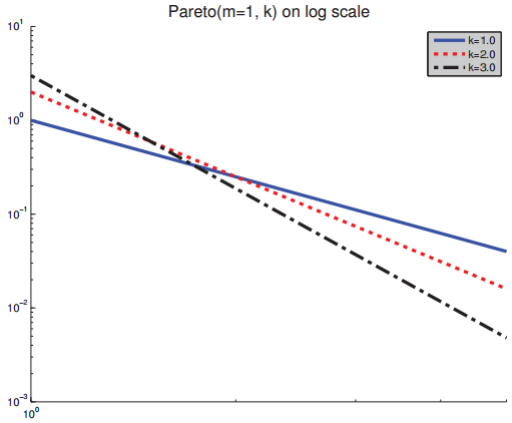
Table 0.8: Summary of Pareto distribution

Name	Written as	X	$f(x)$	$\mathbb{E}[X]$	mode	var$[X]$
Pareto distribution	$X \sim \text{Pareto}(k,m)$	$x \geq m$	$km^k x^{-(k+1)} \mathbb{I}(x \geq m)$	$\frac{km}{k-1}$ if $k>1$	m	$\frac{m^2 k}{(k-1)^2(k-2)}$ if $k>2$

Figure ??(b) for an illustration (this is known as a power law).



(a)



(b)

Figure 0.5: (a) The Pareto distribution Pareto$(x|m,k)$ for $m=1$. (b) The pdf on a log-log scale.

0.5 Joint probability distributions

Given a multivariate random variable or random vector [2] $X \in \mathbb{R}^D$, the joint probability distribution[3] is a probability distribution that gives the probability that each of $X_1, X_2, \cdots, X_D$ falls in any particular range or discrete set of values specified for that variable. In the case of only two random variables, this is called a bivariate distribution, but the concept generalizes to any number of random variables, giving a multivariate distribution.

The joint probability distribution can be expressed either in terms of a joint cumulative distribution function or in terms of a joint probability density function (in the case of continuous variables) or joint probability mass function (in the case of discrete variables).

0.5.1 Covariance and correlation

Definition 0.6. The covariance between two rv's X and Y measures the degree to which X and Y are (linearly) related. Covariance is defined as

$$\begin{aligned} \text{cov}[X,Y] &\triangleq \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \end{aligned} \tag{0.28}$$

Definition 0.7. If X is a D-dimensional random vector, its covariance matrix is defined to be the following symmetric, positive definite matrix:

[2] http://en.wikipedia.org/wiki/Multivariate_random_variable

[3] http://en.wikipedia.org/wiki/Joint_probability_distribution

$$\text{cov}[X] \triangleq \mathbb{E}\left[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T\right] \tag{0.29}$$

$$= \begin{pmatrix} \text{var}[X_1] & \text{Cov}[X_1, X_2] & \cdots & \text{Cov}[X_1, X_D] \\ \text{Cov}[X_2, X_1] & \text{var}[X_2] & \cdots & \text{Cov}[X_2, X_D] \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}[X_D, X_1] & \text{Cov}[X_D, X_2] & \cdots & \text{var}[X_D] \end{pmatrix} \tag{0.30}$$

Definition 0.8. The (Pearson) correlation coefficient between X and Y is defined as

$$\text{corr}[X, Y] \triangleq \frac{\text{Cov}[X, Y]}{\sqrt{\text{var}[X], \text{var}[Y]}} \tag{0.31}$$

A correlation matrix has the form

$$\mathbf{R} \triangleq \begin{pmatrix} \text{corr}[X_1, X_1] & \text{corr}[X_1, X_2] & \cdots & \text{corr}[X_1, X_D] \\ \text{corr}[X_2, X_1] & \text{corr}[X_2, X_2] & \cdots & \text{corr}[X_2, X_D] \\ \vdots & \vdots & \ddots & \vdots \\ \text{corr}[X_D, X_1] & \text{corr}[X_D, X_2] & \cdots & \text{corr}[X_D, X_D] \end{pmatrix} \tag{0.32}$$

The correlation coefficient can viewed as a degree of linearity between X and Y, see Figure ??.

Uncorrelated does not imply independent. For example, let $X \sim U(-1, 1)$ and $Y = X^2$. Clearly Y is dependent on X(in fact, Y is uniquely determined by X), yet one can show that $\text{corr}[X, Y] = 0$. Some striking examples of this fact are shown in Figure ??. This shows several data sets where there is clear dependence between X and Y, and yet the correlation coefficient is 0. A more general measure of dependence between random variables is mutual information, see Section TODO.

0.5.2 Multivariate Gaussian distribution

The multivariate Gaussian or multivariate normal(MVN) is the most widely used joint probability density function for continuous variables. We discuss MVNs in detail in Chapter 4; here we just give some definitions and plots.

The pdf of the MVN in D dimensions is defined by the following:

$$\mathcal{N}(\vec{x}|\vec{\mu}, \Sigma) \triangleq \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\vec{x} - \vec{\mu})^T \Sigma^{-1} (\vec{x} - \vec{\mu})\right] \tag{0.33}$$

where $\vec{\mu} = \mathbb{E}[X] \in \mathbb{R}^D$ is the mean vector, and $\Sigma = \text{Cov}[X]$ is the $D \times D$ covariance matrix. The normalization constant $(2\pi)^{D/2}|\Sigma|^{1/2}$ just ensures that the pdf integrates to 1.

Figure ?? plots some MVN densities in 2d for three different kinds of covariance matrices. A full covariance matrix has A $D(D+1)/2$ parameters (we divide by 2 since Σ is symmetric). A diagonal covariance matrix has D parameters, and has 0s in the off-diagonal terms. A spherical or isotropic covariance,$\Sigma = \sigma^2 \vec{I}_D$, has one free parameter.

0.5.3 Multivariate Student's t-distribution

A more robust alternative to the MVN is the multivariate Student's t-distribution, whose pdf is given by

$$\mathcal{T}(x|\vec{\mu}, \Sigma, \nu) \triangleq \frac{\Gamma(\frac{\nu+D}{2})}{\Gamma(\frac{\nu}{2})} \frac{|\Sigma|^{-\frac{1}{2}}}{(\nu\pi)^{\frac{D}{2}}} \left[1 + \frac{1}{\nu}(\vec{x} - \vec{\mu})^T \Sigma^{-1} (\vec{x} - \vec{\mu})\right]^{-\frac{\nu+D}{2}} \tag{0.34}$$

$$= \frac{\Gamma(\frac{\nu+D}{2})}{\Gamma(\frac{\nu}{2})} \frac{|\Sigma|^{-\frac{1}{2}}}{(\nu\pi)^{\frac{D}{2}}} \left[1 + (\vec{x} - \vec{\mu})^T \vec{V}^{-1} (\vec{x} - \vec{\mu})\right]^{-\frac{\nu+D}{2}} \tag{0.35}$$

where Σ is called the scale matrix (since it is not exactly the covariance matrix) and $\vec{V} = \nu\Sigma$. This has fatter tails than a Gaussian. The smaller ν is, the fatter the tails. As $\nu \to \infty$, the distribution tends towards a Gaussian. The distribution has the following properties

$$\text{mean} = \vec{\mu} \ , \ \text{mode} = \vec{\mu} \ , \ \text{Cov} = \frac{\nu}{\nu - 2}\Sigma \tag{0.36}$$

0.5.4 Dirichlet distribution

A multivariate generalization of the beta distribution is the Dirichlet distribution, which has support over the probability simplex, defined by

$$S_K = \left\{\vec{x} : 0 \le x_k \le 1, \sum_{k=1}^{K} x_k = 1\right\} \tag{0.37}$$

The pdf is defined as follows:

$$\text{Dir}(\vec{x}|\vec{\alpha}) \triangleq \frac{1}{B(\vec{\alpha})} \prod_{k=1}^{K} x_k^{\alpha_k - 1} \mathbb{I}(\vec{x} \in S_K) \tag{0.38}$$

where $B(\alpha_1, \alpha_2, \cdots, \alpha_K)$ is the natural generalization of the beta function to K variables:

$$B(\alpha) \triangleq \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\alpha_0)} \text{ where } \alpha_0 \triangleq \sum_{k=1}^{K} \alpha_k \tag{0.39}$$

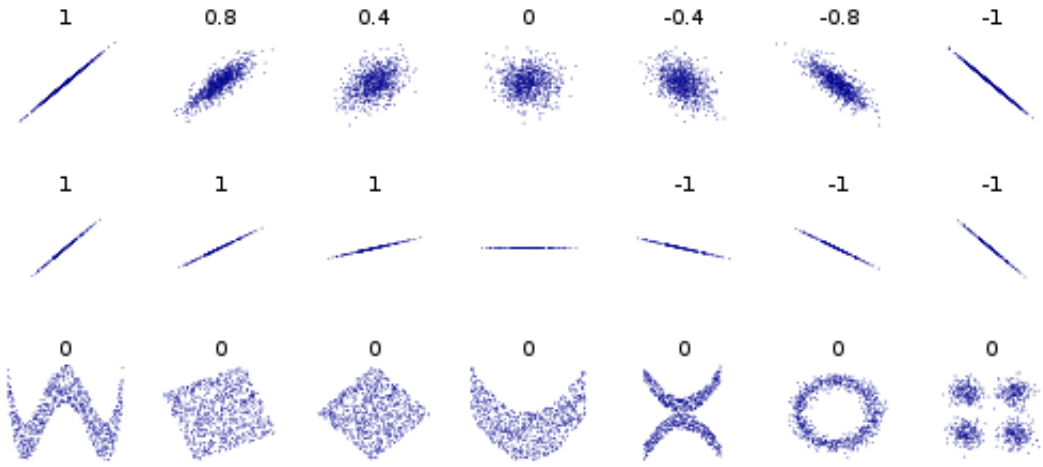


Figure 0.6: Several sets of (x,y) points, with the Pearson correlation coefficient of x and y for each set. Note that the correlation reflects the noisiness and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom). N.B.: the figure in the center has a slope of 0 but in that case the correlation coefficient is undefined because the variance of Y is zero.Source:http://en.wikipedia.org/wiki/Correlation

Figure ?? shows some plots of the Dirichlet when $K=3$, and Figure ?? for some sampled probability vectors. We see that α_0 controls the strength of the distribution (how peaked it is), and the□kcontrol where the peak occurs. For example, $\text{Dir}(1,1,1)$ is a uniform distribution, $\text{Dir}(2,2,2)$ is a broad distribution centered at $(1/3,1/3,1/3)$, and $\text{Dir}(20,20,20)$ is a narrow distribution centered at $(1/3,1/3,1/3)$.If $\alpha_k < 1$ for all k, we get "spikes" at the corner of the simplex.

For future reference, the distribution has these properties

$$\mathbb{E}(x_k) = \frac{\alpha_k}{\alpha_0},\ \text{mode}[x_k] = \frac{\alpha_k - 1}{\alpha_0 - K},\ \text{var}[x_k] = \frac{\alpha_k(\alpha_0 - \alpha_k)}{\alpha_0^2(\alpha_0 + 1)} \tag{0.40}$$

0.6 Transformations of random variables

If $\vec{x} \sim P()$ is some random variable, and $\vec{y} = f(\vec{x})$, what is the distribution of Y? This is the question we address in this section.

0.6.1 Linear transformations

Suppose $g()$ is a linear function:

$$g(\vec{x}) = A\vec{x} + b \tag{0.41}$$

First, for the mean, we have

$$\mathbb{E}[\vec{y}] = \mathbb{E}[A\vec{x} + b] = A\mathbb{E}[\vec{x}] + b \tag{0.42}$$

this is called the linearity of expectation.

For the covariance, we have

$$\text{Cov}[\vec{y}] = \text{Cov}[A\vec{x} + b] = A\Sigma A^T \tag{0.43}$$

0.6.2 General transformations

If X is a discrete rv, we can derive the pmf for y by simply summing up the probability mass for all the x's such that $f(x) = y$:

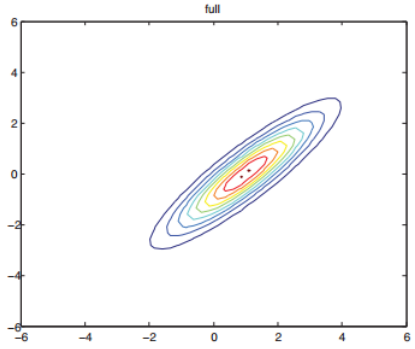
$$p_Y(y) = \sum_{x:g(x)=y} p_X(x) \tag{0.44}$$

If X is continuous, we cannot use Equation ?? since $p_X(x)$ is a density, not a pmf, and we cannot sum up densities. Instead, we work with cdf's, and write

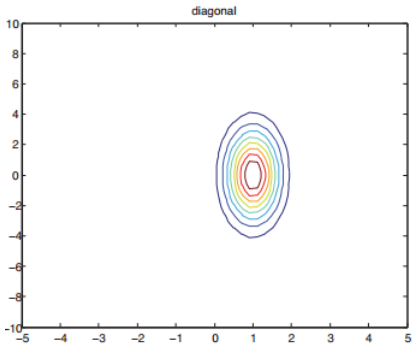
$$F_Y(y) = P(Y \le y) = P(g(X) \le y) = \int_{g(X)\le y} f_X(x)\text{d}x \tag{0.45}$$

We can derive the pdf of Y by differentiating the cdf:

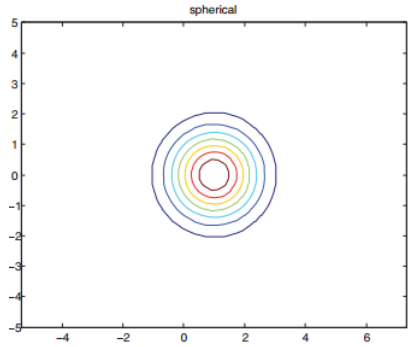
$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right| \tag{0.46}$$



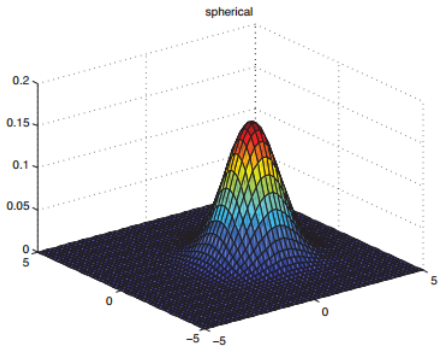
(a)



(b)

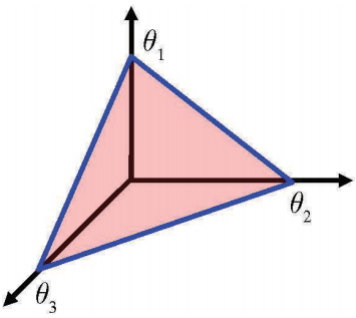


(c)

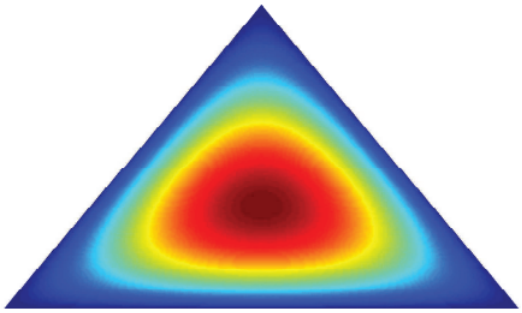


(d)

Figure 0.7: We show the level sets for 2d Gaussians. (a) A full covariance matrix has elliptical contours.(b) A diagonal covariance matrix is an axis aligned ellipse. (c) A spherical covariance matrix has a circular shape. (d) Surface plot for the spherical Gaussian in (c).

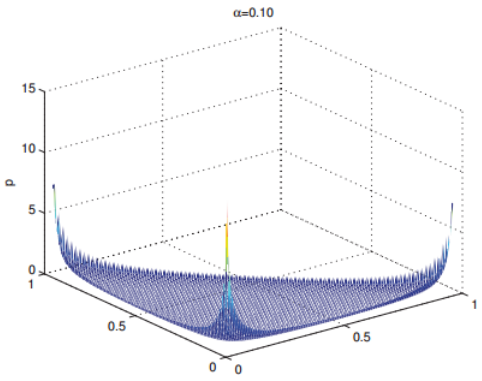


(a)



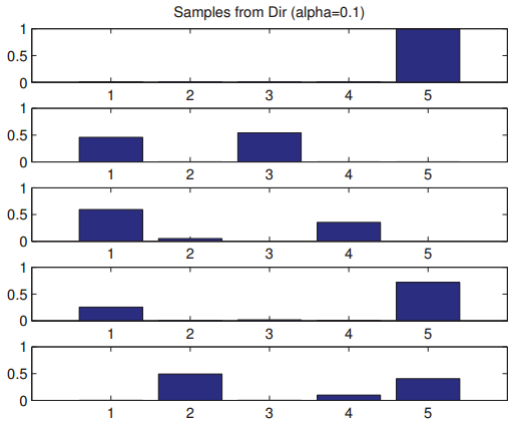
(b)

(c)

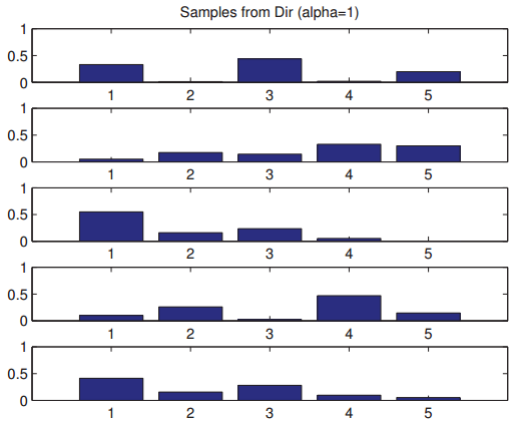


(d)

Figure 0.8: (a) The Dirichlet distribution when $K = 3$ defines a distribution over the simplex, which can be represented by the triangular surface. Points on this surface satisfy $0 \leq \theta_k \leq 1$ and $\sum_{k=1}^{K} \theta_k = 1$. (b) Plot of the Dirichlet density when $\vec{\alpha} = (2,2,2)$. (c) $\vec{\alpha} = (20,2,2)$.



(a) $\vec{\alpha} = (0.1, \cdots, 0.1)$. This results in very sparse distributions, with many 0s.



(b) $\vec{\alpha} = (1, \cdots, 1)$. This results in more uniform (and dense) distributions.

Figure 0.9: Samples from a 5-dimensional symmetric Dirichlet distribution for different parameter values.

This is called change of variables formula. We leave the proof of this as an exercise.

For example, suppose $X \sim U(1,1)$, and $Y = X^2$. Then $p_Y(y) = \frac{1}{2} y^{-\frac{1}{2}}$.

0.6.2.1 Multivariate change of variables *

Let f be a function $f : \mathbb{R}^n \to \mathbb{R}^n$, and let $\vec{y} = f(\vec{x})$. Then its Jacobian matrix $\vec{J}$ is given by

$$\vec{J}_{\vec{x} \to \vec{y}} \triangleq \frac{\partial \vec{y}}{\partial \vec{x}} \triangleq \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \vdots & \vdots \\ \frac{\partial y_n}{\partial x_1} & \cdots & \frac{\partial y_n}{\partial x_n} \end{pmatrix} \tag{0.47}$$

$|\det(\vec{J})|$ measures how much a unit cube changes in volume when we apply f.

If f is an invertible mapping, we can define the pdf of the transformed variables using the Jacobian of the inverse mapping $\vec{y} \to \vec{x}$:

$$p_y(\vec{y}) = p_x(\vec{x}) |\det(\frac{\partial \vec{x}}{\partial \vec{y}})| = p_x(\vec{x}) |\det(\vec{J}_{\vec{y} \to \vec{x}})| \tag{0.48}$$

0.6.3 Central limit theorem

Given N random variables $X_1, X_2, \cdots, X_N$, each variable is independent and identically distributed[4](iid for short), and each has the same mean μ and variance σ^2, then

$$\frac{\sum_{i=1}^{n} X_i - N\mu}{\sqrt{N}\sigma} \sim \mathcal{N}(0,1) \tag{0.49}$$

this can also be written as

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{N}} \sim \mathcal{N}(0,1) \quad \text{, where } \bar{X} \triangleq \frac{1}{N} \sum_{i=1}^{n} X_i \tag{0.50}$$

0.7 Monte Carlo approximation

In general, computing the distribution of a function of an rv using the change of variables formula can be difficult. One simple but powerful alternative is as follows. First we generate S samples from the distribution, call them $x_1, \cdots, x_S$. (There are many ways to generate such samples; one popular method, for high dimensional distributions, is called Markov chain Monte Carlo or MCMC; this will be explained in Chapter TODO.) Given the samples, we can approximate the distribution of $f(X)$ by using the empirical distribution of $\{f(x_s)\}_{s=1}^{S}$. This is called a Monte Carlo approximation[5], named after a city in Europe known for its plush gambling casinos.

We can use Monte Carlo to approximate the expected value of any function of a random variable. We simply draw samples, and then compute the arithmetic mean of the function applied to the samples. This can be written as follows:

$$\mathbb{E}[g(X)] = \int g(x) p(x) \mathrm{d}x \approx \frac{1}{S} \sum_{s=1}^{S} f(x_s) \tag{0.51}$$

[4] http://en.wikipedia.org/wiki/Independent_identically_distributed

[5] http://en.wikipedia.org/wiki/Monte_Carlo_method

where $x_s \sim p(X)$.

This is called Monte Carlo integration[6], and has the advantage over numerical integration (which is based on evaluating the function at a fixed grid of points) that the function is only evaluated in places where there is non-negligible probability.

0.8 Information theory

0.8.1 Entropy

The entropy of a random variable X with distribution p, denoted by $\mathbb{H}(X)$ or sometimes $\mathbb{H}(p)$, is a measure of its uncertainty. In particular, for a discrete variable with K states, it is defined by

$$\mathbb{H}(X) \triangleq -\sum_{k=1}^{K} p(X=k)\log_2 p(X=k) \tag{0.52}$$

Usually we use log base 2, in which case the units are called bits(short for binary digits). If we use log base e , the units are called nats.

The discrete distribution with maximum entropy is the uniform distribution (see Section XXX for a proof). Hence for a K-ary random variable, the entropy is maximized if $p(x=k) = 1/K$; in this case, $\mathbb{H}(X) = \log_2 K$.

Conversely, the distribution with minimum entropy (which is zero) is any delta-function that puts all its mass on one state. Such a distribution has no uncertainty.

0.8.2 KL divergence

One way to measure the dissimilarity of two probability distributions, p and q , is known as the Kullback-Leibler divergence(KL divergence)or relative entropy. This is defined as follows:

$$\mathbb{KL}(P||Q) \triangleq \sum_x p(x)\log_2 \frac{p(x)}{q(x)} \tag{0.53}$$

where the sum gets replaced by an integral for pdfs[7]. The KL divergence is only defined if P and Q both sum to 1 and if $q(x) = 0$ implies $p(x) = 0$ for all x(absolute continuity). If the quantity $0\ln 0$ appears in the formula, it is interpreted as zero because $\lim_{x\to 0} x\ln x$. We can rewrite this as

$$\begin{aligned}\mathbb{KL}(p||q) &\triangleq \sum_x p(x)\log_2 p(x) - \sum_{k=1}^{K} p(x)\log_2 q(x) \\ &= \mathbb{H}(p,q) - \mathbb{H}(p)\end{aligned} \tag{0.54}$$

where $\mathbb{H}(p,q)$ is called the cross entropy,

$$\mathbb{H}(p,q) = -\sum_x p(x)\log_2 q(x) \tag{0.55}$$

One can show (Cover and Thomas 2006) that the cross entropy is the average number of bits needed to encode data coming from a source with distribution p when we use model q to define our codebook. Hence the "regular" entropy $\mathbb{H}(p) = \mathbb{H}(p,p)$, defined in section □??,is the expected number of bits if we use the true model, so the KL divergence is the diference between these. In other words, the KL divergence is the average number of extra bits needed to encode the data, due to the fact that we used distribution q to encode the data instead of the true distribution p.

The "extra number of bits" interpretation should make it clear that $\mathbb{KL}(p||q) \geq 0$, and that the KL is only equal to zero if $q = p$. We now give a proof of this important result.

Theorem 0.1. (Information inequality) $\mathbb{KL}(p||q) \geq 0$ with equality iff $p = q$.

One important consequence of this result is that the discrete distribution with the maximum entropy is the uniform distribution.

0.8.3 Mutual information

Definition 0.9. Mutual information or MI, is defined as follows:

$$\begin{aligned}\mathbb{I}(X;Y) &\triangleq \mathbb{KL}(P(X,Y)||P(X)P(Y)) \\ &= \sum_x \sum_y p(x,y)\log \frac{p(x,y)}{p(x)p(y)}\end{aligned} \tag{0.56}$$

We have $\mathbb{I}(X;Y) \geq 0$ with equality if $P(X,Y) = P(X)P(Y)$. That is, the MI is zero if the variables are independent.

To gain insight into the meaning of MI, it helps to re-express it in terms of joint and conditional entropies. One can show that the above expression is equivalent to the following:

[6] http://en.wikipedia.org/wiki/Monte_Carlo_integration

[7] The KL divergence is not a distance, since it is asymmetric. One symmetric version of the KL divergence is the Jensen-Shannon divergence, defined as $JS(p_1,p_2) = 0.5\mathbb{KL}(p_1||q) + 0.5\mathbb{KL}(p_2||q)$,where $q = 0.5p_1 + 0.5p_2$

$$\mathbb{I}(X;Y) = \mathbb{H}(X) - \mathbb{H}(X|Y) \quad (0.57)$$
$$= \mathbb{H}(Y) - \mathbb{H}(Y|X) \quad (0.58)$$
$$= \mathbb{H}(X) + \mathbb{H}(Y) - \mathbb{H}(X,Y) \quad (0.59)$$
$$= \mathbb{H}(X,Y) - \mathbb{H}(X|Y) - \mathbb{H}(Y|X) \quad (0.60)$$

where $\mathbb{H}(X)$ and $\mathbb{H}(Y)$ are the marginal entropies, $\mathbb{H}(X|Y)$ and $\mathbb{H}(Y|X)$ are the conditional entropies, and $\mathbb{H}(X,Y)$ is the joint entropy of X and Y, see Fig. ??[8].

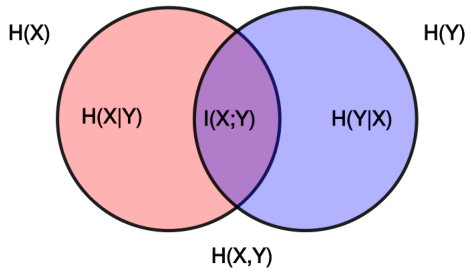


Figure 0.10: Individual $\mathbb{H}(X), \mathbb{H}(Y)$, joint $\mathbb{H}(X,Y)$, and conditional entropies for a pair of correlated subsystems X,Y with mutual information $\mathbb{I}(X;Y)$.

Intuitively, we can interpret the MI between X and Y as the reduction in uncertainty about X after observing Y, or, by symmetry, the reduction in uncertainty about Y after observing X.

A quantity which is closely related to MI is the pointwise mutual information or PMI. For two events (not random variables) x and y, this is defined as

$$PMI(x,y) \triangleq \log \frac{p(x,y)}{p(x)p(y)} = \log \frac{p(x|y)}{p(x)} = \log \frac{p(y|x)}{p(y)} \quad (0.61)$$

This measures the discrepancy between these events occuring together compared to what would be expected by chance. Clearly the MI of X and Y is just the expected value of the PMI. Interestingly, we can rewrite the PMI as follows:

$$PMI(x,y) = \log \frac{p(x|y)}{p(x)} = \log \frac{p(y|x)}{p(y)} \quad (0.62)$$

This is the amount we learn from updating the prior $p(x)$ into the posterior $p(x|y)$, or equivalently, updating the prior $p(y)$ into the posterior $p(y|x)$.

[8] http://en.wikipedia.org/wiki/Mutual_information