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 Probability theory- এর একটি সংক্ষিপ্ত পর্যালোচনা

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 ব্যাসিক কনসেপ্ট

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)

Probability Distribution: একটা random variable-এর সম্ভাব্য সকল মান পড়ার সম্ভাবনাকে একত্র করলে,সেটাকে প্রবাবিলিটি ডিস্ট্রিবিউশন বলছে। discrete value এর জন্যে একে PMF বলে এবং continious value এর ক্ষেত্রে PDF বলে, যেটা আমরা পরের সেকশনে বিস্তারিত জানতে পারবো ।

লুডুর ডাইসের কথা চিন্তা করা যাক, যেখানে একবার ৬ পড়ার সম্ভাবনা $P(X)$= ১/৬; সকল মানকে একসাথে যদি আমরা একটা গ্রাফে বসায় যেখানে x-অক্ষ বরর্বর ডাটা পয়েন্ট থাকবে, আর তাদের এককভাবে আসার সম্ভাবনা Y-অক্ষ বরাবর রেখে মান বসালে যেই সম্পূর্ণ স্পেইসটা পাবো সেটাই probabilty distribution

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।

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

- $P(X \leq x)$ X -এর মান x-

- $p(u)$ 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 এবং 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 variable-এর 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 Multivariate random variables

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 -

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

কন্টিনিউয়াস এর ক্ষেত্রে -

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

0.2.2.1 Joint CDF

দুটি random variable 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 random variable-এর 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 conditional probability ($P(X,Y)$) এবং marginal probability ($P(Y)$) এর গুণফলের মাধ্যমে প্রকাশ করা হয়।

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

- $P(X \cap Y)$ বোঝায় X এবং
- $P(X \mid Y)$ বোঝায় Y , X ঘটার সম্ভাবনা
- $P(Y)$ 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 concept) 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

Marginal Distribution variable এর সেট থেকে কেবলমাত্র একটা ভ্যারিয়েবলকে ফোকাসে রেখে তার probability distirbution 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।

X এর জন্যে Marginal CDF:

Discrete case এবং 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 variable 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 joint probability $P(X = x_i, Y = y_j)$ sum up করছে।
- দ্বিতীয় ক্ষেত্রে, যখন random variable 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।

Y এর জন্যে Marginal CDF:

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

- 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 PDF:

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

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

0.2.2.5 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

pmf $p(X|Y)$ কে বলা হয় 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 joint probability
- $p(Y = y_j)$, 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 ।

- denominator -এ 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 & conditional independence

X এবং Y unconditional বা marginally independent হবে, যদি তাদের joint probability কে তাদের marginal probability-র গুণফল এর মাধ্যমে প্রকাশ করা যায়। Unconditional independence $X \perp Y$ সিম্বলের মাধ্যমে প্রকাশ করা হয়; যেখানে -

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

- সমীকরণে , X এবং Y একে অপরের উপর নির্ভরশীল নয়, তাদের মধ্যে কোনো সরাসরি সম্পর্ক নেই। একসাথে X এবং Y-এর joint probability, তাদের marginal probability $P(X)$ এবং $P(Y)$-এর গুণফল দিয়ে প্রকাশ করা সম্ভব।

- অন্যদিকে , X এবং Y conditionally independent (CI) হবে, যদি আরও একটি variable Z এর উপস্থিতি থাকে । এটি প্রকাশ করা হয়:

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

- এখানে যদি Z ঘটার সম্ভাবনা থাকে, তবে X এবং Y এর মধ্যে কোনো সম্পর্ক থাকে না । Z এর সাথে তাদের conditional probability থাকার শর্তে, নিজেদের মধ্যে joint probability থাকছে; পরের ইকুয়েশনে এই শর্তে তাদের marginal probabilities দিয়ে প্রকাশ করা যায়।

0.2.5 Quantiles

যেহেতু cdf F একই প্যাটার্নে (monotonic) বাড়তে থাকা ফাংশন, এর একটি inverse আছে, যা F^{-1} মাধ্যমে প্রকাশ করা যায়। যদি X-এর cdf F হয় , তাহলে $F^{-1}(\alpha)$ হল x_α-এর সেই মান যেখানে $P(X \leq x_\alpha) = \alpha$। একে বলা হয় F-এর α quan-

tile। $F^{-1}(0.5)$-এর মানকে বলা হয় distribution-এর median, যেখানে বামপাশে এবং ডানপাশে probability সমানভাবে ভাগ করা থাকে। $F^{-1}(0.25)$ এবং $F^{-1}(0.75)$ lower এবং upper quartiles।

0.2.6 Mean variance

একটি distribution-এর সবচেয়ে পরিচিত বৈশিষ্ট্য হল তার mean (গড়) বা expected value, যা μ দিয়ে প্রকাশ করা হয়। discrete random variable-এর জন্য mean কে সংজ্ঞায়িত করা হয়:

$$\mathbb{E}[X] \triangleq \sum_{x \in \mathcal{X}} xp(x)$$

$\mathbb{E}[X]$ দ্বারা X এর expected value বা গড় প্রকাশ করা হয় এবং continuous variable-এর জন্য:

$$\mathbb{E}[X] \triangleq \int_{\mathcal{X}} xp(x)\mathrm{d}x$$

যদি এই ইন্টিগ্রাল finite না হয়, তাহলে mean নির্ধারণ করা যায় না।

Variance হল একটি distribution-এর "spread" বা বিচরণ পরিমাপ। এটি σ^2 দিয়ে প্রকাশ করা কতখানি দূরে আছে তা পরিমাপ করে । variance এর সংজ্ঞা দেয়া হয়:

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

- প্রথম অংশে $\mathbb{E}[(X-\mu)^2]$ মানে, আমরা X এর মান থেকে এর গড় μ বাদ দিয়ে তার স্কোয়ারের গড় নিচ্ছি।

- এরপর এইটাকে আমরা $\int (x-\mu)^2 p(x)\,\mathrm{d}x$ এর মাধ্যমে প্রকাশ করি, যেখানে $p(x)$ হলো X এর probability density function।

- তৃতীয় ক্ষেত্রে, $\int x^2 p(x)\,\mathrm{d}x$ হলো X এর স্কোয়ার করা মানগুলোর expected value।

- $\mu^2 \int p(x)\,\mathrm{d}x$ অংশটি গড়ের স্কোয়ারের গড় প্রকাশ করছে, আর $2\mu \int xp(x)\,\mathrm{d}x$ হলো গড় ও X এর মানের সম্পর্ক।

- সবশেষে, আমরা দেখতে পাই $\mathbb{E}[X^2] - \mu^2$, যা হলো variance এর সংক্ষিপ্ত ফর্ম

এখান থেকে আমরা একটি গুরুত্বপূর্ণ ফলাফল পাই:

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

Standard deviation কে সংজ্ঞায়িত করা হয়:

$$std[X] \triangleq \sqrt{var[X]} \qquad (0.19)$$

এটা কাজে লাগে কারণ এটি X-এর সাথে একই একক (units) থাকে।

0.3 কিছু সাধারণ ডিসক্রিট ডিস্ট্রিবিউশন

এই অংশে আমরা কিছু সাধারণত ব্যবহৃত প্যারামেট্রিক ডিস্ট্রিবিউশন আলোচনা করবো, যা ডিসক্রিট স্টেট স্পেসের উপর ভিত্তি করে, কিছু ফাইনাইট এবং কাউন্টেবল ইনফিনিট।

0.3.1 Barnouli এবং Binomial Distribution

Definition 0.1. ধরি আমরা একটি কয়েন একবার ছুঁড়লাম। $X \in \{0,1\}$ হলো একটি binary random variable, যেখানে "হেডস" পাওয়ার সম্ভাবনা θ। তখন আমরা বলি যে X এর Barnouli Distribution আছে। এটা লেখা হয় $X \sim \mathrm{Ber}(\theta)$, যেখানে pmf (Probability Mass Function) এর সংজ্ঞা দেয়া হয় এভাবে:

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

- X হলো র‍্যান্ডম ভ্যারিয়েবল, যা ০ অথবা ১ হতে পারে (মানে হেডস বা টেইলস পাওয়া গেল কিনা)।

- θ হলো হেডস পাওয়া সম্ভাবনা।

- $\mathbb{I}(x=1)$ একটি ইন্ডিকেটর ফাংশন, যা তখন ১ হবে যখন $x=1$, আর বাকি সব ক্ষেত্রে ০।

- $\theta^{\mathbb{I}(x=1)}$ এই টার্মটি তখন কাজ করবে যখন $x=1$, আর $1-\theta^{\mathbb{I}(x=0)}$ কাজ করবে যখন $x=0$।

Definition 0.2. n $X \in \{0,1,\cdots,n\}$ θ , X Binomial Distribution , $X \sim \mathrm{Bin}(n,\theta)$ pmf :

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

- $\binom{n}{k}$ এই টার্মটি কম্বিনেশন বোঝায়, মানে n বার ট্রায়াল থেকে k বার হেডস পাওয়ার সম্ভাব্য উপায়

- θ^k বোঝায় k বার হেডস পাওয়ার সম্ভাবনা।

- $(1-\theta)^{n-k}$ হলো বাকি $n-k$ বার টেইলস পাওয়ার সম্ভাবনা।

- পুরো ফর্মুলা বলতে চায় k বার সফলতা পাওয়ার সম্ভাবনা কী, যেখানে n বার কয়েন ছুঁড়া হয়েছে।

0.3.2 multinoulli এবং multinomial distribution

Definition 0.3. Barnouli Distribution একবার কয়েন ছুঁড়া মডেল করে। কিন্তু যদি K-পাশওয়ালা একটি লুডুর ডাইস ছুঁড়তে হয়, তখন $\vec{x} = (\mathbb{I}(x=1),\cdots,\mathbb{I}(x=K)) \in \{0,1\}^K$ হবে একটি random vector (এটা dummy encoding বা one-hot encoding নামে পরিচিত), তখন বলা হয় X এর multinoulli distribution (বা categorical distribution) আছে, প্রকাশ করা হয় $X \sim \mathrm{Cat}(\theta)$। pmf হলো:

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

- এক্ষেত্রে $\vec{x}$ একটি one-hot encoding vector, যেটা দেখায় কোন সাইডে (১ থেকে K) ডাইস পড়লো।
- θ_k হলো k-তম সাইডে ডাইস পড়ার সম্ভাবনা।
- $\prod_{k=1}^{K} \theta_k^{\mathbb{I}(x_k=1)}$ বোঝায় সেই সাইডে (কোন একটা নির্দিষ্ট k) ডাইস পড়ার সম্ভাবনা।

Definition 0.4. ধরি আমরা K-পাশওয়ালা ডাইস n বার ছুঁড়লাম। $\vec{x} = (x_1, x_2, \cdots, x_K) \in \{0, 1, \cdots, n\}^K$ হলো একটি random vector, যেখানে x_j হলো j-তম সাইডে কতবার ডাইস পড়েছে। তখন বলা হয় X এর মাল্টিনোমিয়াল ডিস্ট্রিবিউশন আছে, লেখা হয় $X \sim \text{Mu}(n, \vec{\theta})$। pmf হলো:

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

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

- $\binom{n}{x_1 \cdots x_k}$ বোঝায় n ট্রায়াল থেকে $x_1, x_2, \ldots, x_K$ হেডস পাওয়ার সম্ভাব্য উপায় - $\prod_{k=1}^{K} \theta_k^{x_k}$ হলো প্রতিটি সাইডের হেডস পাওয়ার সম্ভাবনা

Bernoulli Distribution হচ্ছে Binomial distribution- এর বিশেষ কেস, যেখানে $n = 1$। একইভাবে multinoulli distirbution হচ্ছে Multinomial distribution-এর একটি বিশেষ কেস। টেবিল ??-এ সংক্ষেপে দেখানো হয়েছে।

0.3.3 Poisson distribution

Definition 0.5. $X \in \{0, 1, 2, \cdots\}$,
$X \sim \text{Poi}(\lambda)$, pmf :

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

- প্রথম টার্মটি একটি নরমালাইজেশন কনস্ট্যান্ট, যাতে ডিস্ট্রিবিউশনের সম্ভাবনা ১ হয়।
- Poisson distribution সাধারণত বিরল ঘটনাগুলোর (rare events) সংখ্যা মডেল করতে ব্যবহার করা হয়, যেমন তেজস্ক্রিয় ক্ষয় (radioactive decay) বা ট্রাফিক অ্যাক্সিডেন্টের সংখ্যা।

0.3.4 Empirical distribution

Empirical distribution function (বা Empirical cdf) হলো Empirical measure-এর সাথে সম্পর্কিত cumulative distirbution function। যদি $\mathcal{D} = \{x_1, x_2, \cdots, x_N\}$ একটি স্যাম্পল সেট হয়, তাহলে এটা সংজ্ঞায়িত হয় এভাবে:

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

0.3.5 Gaussian (normal) distribution

Table 0.1: Summary of Gaussian distribution.

Written as	$f(x)$	$\mathbb{E}[X]$	mode	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

যদি $X \sim N(0, 1)$, আমরা বলি যে X একটি standard normal distribution অনুসরণ করে।

Gaussian distribution পরিসংখ্যানে সবচেয়ে বেশি ব্যবহৃত একটি distribution। এর কয়েকটি কারণ হলো-

1. প্রথমত, এর দুটি parameter আছে যেগুলো সহজে ব্যাখ্যা করা যায়, এবং যেমন mean এবং variance এর মতো সরল বৈশিষ্ট্য এই distirbution-এ পাওয়া যায়।
2. দ্বিতীয়ত, central limit theorem (Section TODO) অনুসারে, independent random variable গুলোর যোগফল প্রায় Gaussian distribution এর কাছাকাছি, যা residual errors বা "noise" মডেল করার জন্য ভালো
3. তৃতীয়ত, Gaussian distribution খুব কম assumption করে (maximum entropy ধারণ করে), নির্দিষ্ট mean এবং variance এর শর্তানুসারে, যেটি আমরা Section TODO তে দেখাবো; এটি অনেক ক্ষেত্রে একটি ভালো default অপশন।
4. সর্বশেষ, এর একটি সহজ গাণিতিক রূপ আছে, যা সহজে ইমপ্লিমেন্ট করা যায়

কেন Gaussian এত ব্যাপকভাবে ব্যবহৃত হয় তার বিস্তারিত আলোচনা দেখতে- (Jaynes 2003, ch 7) ,

0.3.6 Student's t-distribution

Table 0.2: Summary of Student's t-distribution.

Written as	$f(x)$	$\mathbb{E}[X]$	mode	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	$\frac{\nu\sigma^2}{\nu-2}$

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 হচ্ছে mean, $\sigma^2 > 0$ scale parameter, এবং $\nu > 0$ -কে বলা হয় degrees of freedom. Figure ?? দ্রষ্টব্য।

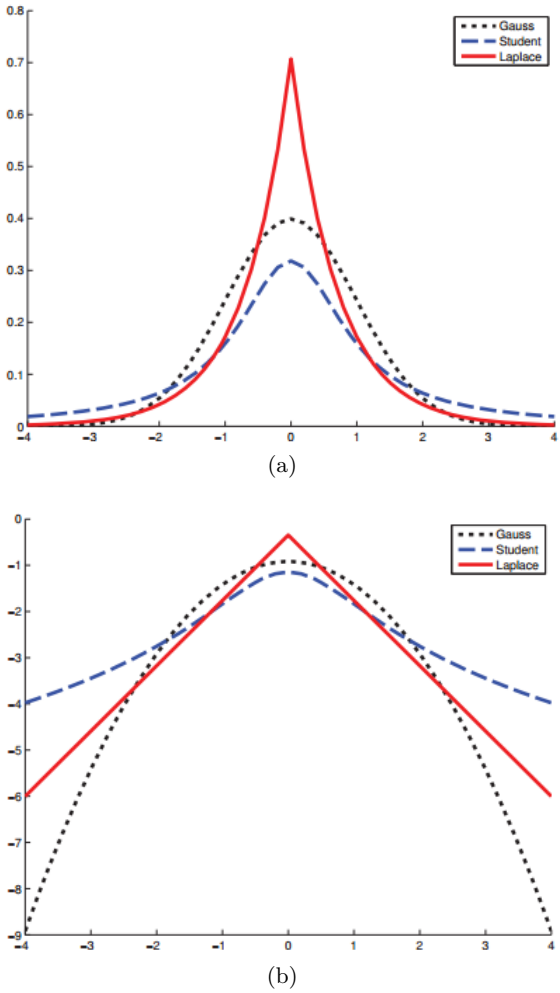


Figure 0.1: (a) $\mathcal{N}(0,1)$, $\mathcal{T}(0,1,1)$ $Lap(0,1/\sqrt{2})$ এর pdf, Gaussian এবং Laplace distribution এর জন্য mean 0 এবং variance 1। তবে যখন $\nu = 1$, তখন Student distribution এর mean এবং variance undefined থাকে। (b) এই pdf গুলোর log; Student distribution কোন parameter মানেই log-concave নয়, অন্যদিকে Laplace distribution সবসময়ই log-concave (এবং log-convex...)। তা সত্ত্বেও, উভয়ই unimodal।

variance শুধুমাত্র $\nu > 2$ হলে defined হয়। Mean শুধুমাত্র $\nu > 1$ হলে defined হয়।

Student distribution এর robustness-এর ক্ষেত্রে, Figure ??-এ আমরা দেখতে পাই যে Gaussian distribution অনেক বেশি প্রভাবিত হয়, কিন্তু Student distribution প্রায় অপরিবর্তিত থাকে। এর কারণ হল Student distribution এর তুলনামূলকভাবে ভারী tails (heavier tails) থাকে, বিশেষ করে যখন ν ছোট হয় (Figure ?? দ্রষ্টব্য)।

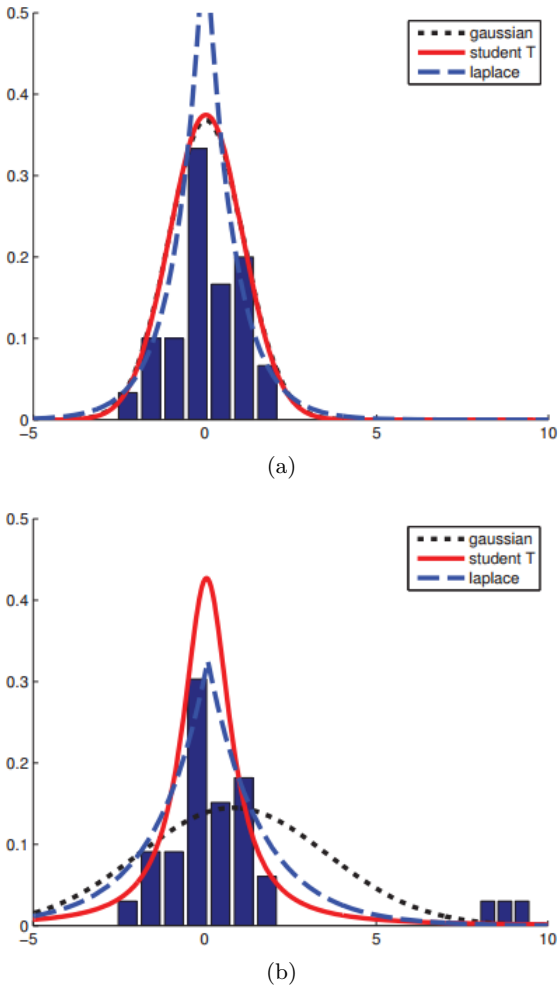


Figure 0.2: Outliers এর প্রভাব Gaussian, Student এবং Laplace distribution ফিট করার উপর কেমন তা দেখানো হয়েছে। (a) যখন কোন outliers নেই (Gaussian এবং Student curves একটির উপর আরেকটি থাকে)। (b) outliers সহ, Gaussian distribution outliers দ্বারা বেশি প্রভাবিত হয়, যেখানে Student এবং Laplace distributions তুলনামূলকভাবে কম প্রভাবিত হয়।

যদি $\nu = 1$ হয়, তখন এই distribution কে Cauchy বা Lorentz distribution বলে। ভারী tails (heavy tails) থাকার জন্য পরিচিত, কারণ যেই integral দ্বারা mean সংজ্ঞায়িত করা হয় তা converge করে না। finite variance নিশ্চিত করার জন্য, আমাদের $\nu > 2$ প্রয়োজন। সাধারণত $\nu = 4$ ব্যবহার করা হয়, যা বিভিন্ন সমস্যার ক্ষেত্রে ভালো performance দেয় (Lange et al. 1989)। যখন $\nu \gg 5$, তখন Student distribution দ্রুত Gaussian distribution এর কাছাকাছি চলে আসে এবং এর robustness এর সুবিধা হারিয়ে ফেলে।

0.3.7 The Laplace distribution

Table 0.3: Summary of Laplace distribution.

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

এখানে μ একটি location parameter এবং $b > 0$ একটি scale parameter। Figure ?? তে একটি plot Laplace distribution এর robustness outliers Figure ??-এ দেখানো হয়েছে। এটি Gaussian এর চেয়ে 0 এর দিকে বেশি probability density রাখে। এই বৈশিষ্ট্যটি একটি মডেলে sparsity (সম্প্রসারণ) বাড়াতে বেশ কার্যকর, যা আমরা Section TODO তে দেখতে পাবো।

0.3.8 The gamma distribution

Table 0.4: Summary of gamma distribution

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

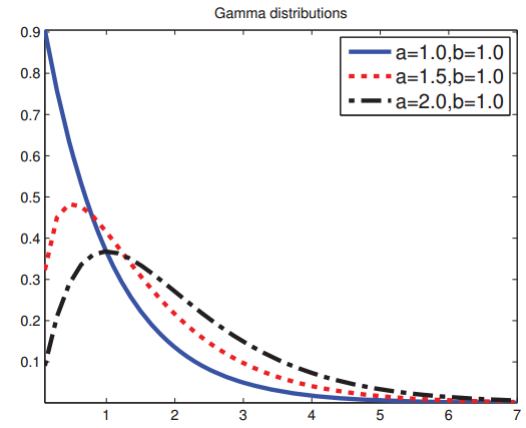
Here is called the and. See Figure for some plots.

0.3.9 The beta distribution

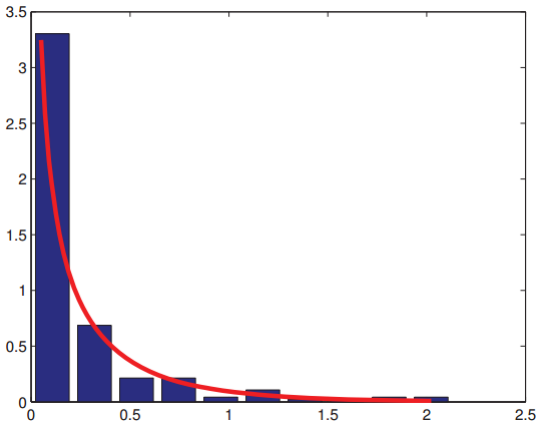
Here $B(a,b)$is the beta function,

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

কিছু beta distributions এর plot এর জন্য Figure ?? দ্রষ্টব্য। Distribution কে integrable করতে (অর্থাৎ $B(a,b)$ কে রাখার জন্যে), আমাদের $a,b > 0$ প্রয়োজন। যদি $a = b = 1$ হয়, আমরা uniform distribution পাই। যদি a এবং b উভয়ই 1 এর চেয়ে কম হয়, আমরা 0 এবং 1 এ "spikes" সহ একটি bimodal distribution পাই; যদি a এবং b উভয়ই 1 এর চেয়ে বড় হয়, distribution unimodal হয়।



(a)



(b)

Figure 0.3: কিছু Ga$(a, b = 1)$ distributions। যদি $a \leq 1$ হয়, তাহলে mode 0-তে থাকে, অন্যথায় এটি > 0 হয়। যখন আমরা b-এর rate বাড়াই, তখন আমরা horizontal scale কমাই, ফলে সবকিছু বাম দিকে এবং উপরের দিকে সঙ্কুচিত হয়। (b) কিছু বৃষ্টিপাতের ডেটার একটি empirical pdf, যেখানে একটি fitted Gamma distribution superimposed করা হয়েছে।

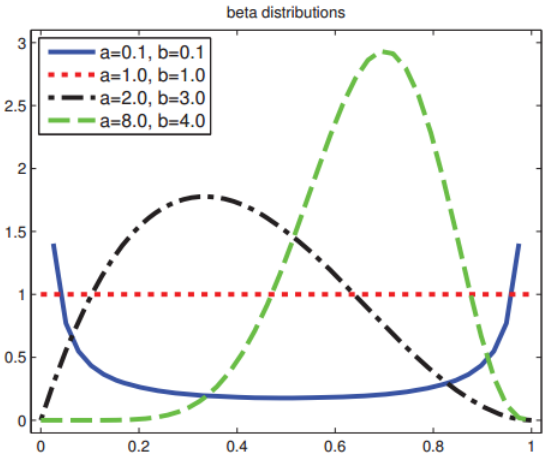


Figure 0.4: beta distribution

Table 0.5: Summary of Beta distribution

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

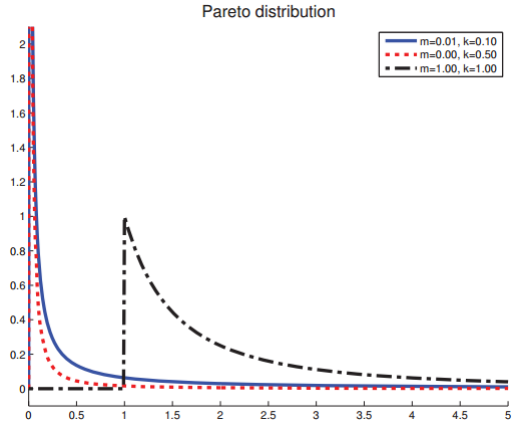
Table 0.6: Summary of Pareto distribution

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 $k>2$

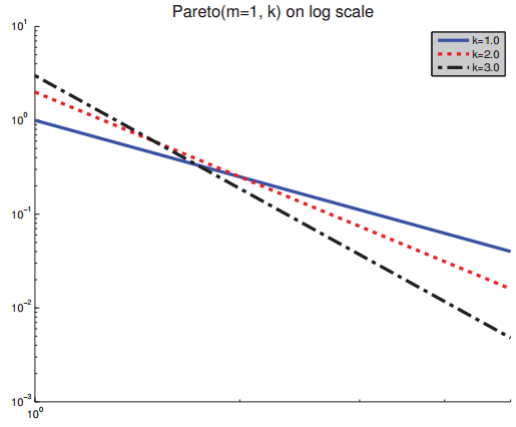
0.3.10 Pareto distribution

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 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.

0.4 Joint probability distributions

ধরি একটি multivariate random variable বা random vector [1] $X \in \mathbb{R}^D$, এর joint probability distribution [2] হল একটি probability distribution যা নির্দেশ করে যে $X_1, X_2, \cdots, X_D$ এর প্রতিটি নির্দিষ্ট ভ্যালু বা নির্দিষ্ট রেঞ্জে পড়ার সম্ভাবনা কত।

যখন মাত্র দুটি random variable থাকে, তখন একে bivariate distribution বলা হয়, কিন্তু যেকোন সংখ্যক random variables এর ক্ষেত্রে generalized করা হলে তাকে multivariate distribution বলে।

Joint probability distribution কে প্রকাশ করা যেতে পারে joint cumulative distribution function এর মাধ্যমে, অথবা joint probability density function এর মাধ্যমে (যদি variables continuous হয়) অথবা joint probability mass function এর মাধ্যমে (যদি variables discrete হয়)।

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

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

0.4.1 Covariance and correlation

Definition 0.6. Covariance random variables X Y ( ) linearity Covariance :

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

Definition 0.7. যদি X একটি D-dimensional random vector হয়, তাহলে এর covariance matrix কে নিম্নরূপ symmetric, positive definite matrix হিসেবে সংজ্ঞায়িত করা হয়:

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

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

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

Definition 0.8. X এবং Y এর মধ্যে(Pearson) correlation coefficient :

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

A correlation matrix has the form

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

Uncorrelated does not imply independent. উদাহরণস্বরূপ, যদি $X \sim U(-1,1)$ এবং $Y = X^2$, Y যদি X এর উপর নির্ভরশীল, একে দেখানো যায় যে $\mathrm{corr}[X,Y]=0$। Figure ?? এ এই বিষয়টি কিছু উদাহরণ সহ দেখানো হয়েছে যেখানে X এবং Y এর মধ্যে স্পষ্ট সম্পর্ক রয়েছে, কিন্তু correlation coefficient 0। Random variables এর মধ্যে আরও সাধারণ একটি dependence-এর পরিমাপ হল mutual information; বিস্তারিত Section TODO তে।

0.4.2 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 $\Sigma = \mathrm{Cov}[X]$ 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.

0.4.3 Multivariate Student's t-distribution

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

$$\begin{aligned}&\mathcal{T}(x|\vec{\mu},\Sigma,\nu) \\ &\triangleq \frac{\Gamma(\frac{\nu+D}{2})}{\Gamma(\frac{\nu}{2})}\frac{|\Sigma|^{-\frac{1}{2}}}{(\nu\pi)^{\frac{D}{2}}}\left[1+\frac{1}{\nu}(\vec{x}-\vec{\mu})^T\Sigma^{-1}(\vec{x}-\vec{\mu})\right]^{-\frac{\nu+D}{2}} \qquad (0.34)\\ &= \frac{\Gamma(\frac{\nu+D}{2})}{\Gamma(\frac{\nu}{2})}\frac{|\Sigma|^{-\frac{1}{2}}}{(\nu\pi)^{\frac{D}{2}}}\left[1+(\vec{x}-\vec{\mu})^T\vec{V}^{-1}(\vec{x}-\vec{\mu})\right]^{-\frac{\nu+D}{2}} \qquad (0.35)\end{aligned}$$

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

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

0.4.4 Dirichlet distribution

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

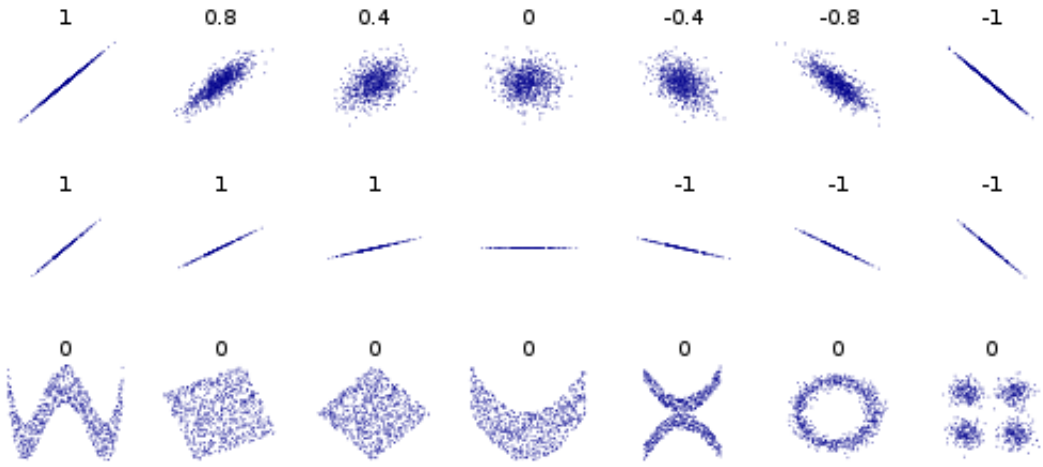


Figure 0.6: বেশ কিছু (x,y) points এর সেটে, x এবং y এর মধ্যে Pearson correlation coefficient নির্ধারণ করা হয়েছে। লক্ষ্য করি যে, correlation একটা linear relation direction এবং noisiness নির্দেশ করে, তবে সেই সম্পর্কের slope নির্দেশ করে না (মাঝের সারি), এবং নন-লিনিয়ার সম্পর্কের অনেক দিকও নির্দেশ করে না (নিচের সারি)। নোটঃ মাঝখানের ছবিটির slope 0, কিন্তু এই ক্ষেত্রে correlation coefficient নির্ধারিত হয়নি কারণ Y এর variance শূন্য। রেফারেন্স :http://en.wikipedia.org/wiki/Correlation

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

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 ?? 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

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

0.5 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.

0.5.1 Linear transformations

Suppose $g()$ is a linear function:

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

First, for the mean, we have

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

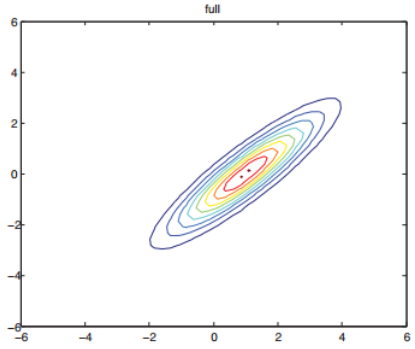
this is called the linearity of expectation.

For the covariance, we have

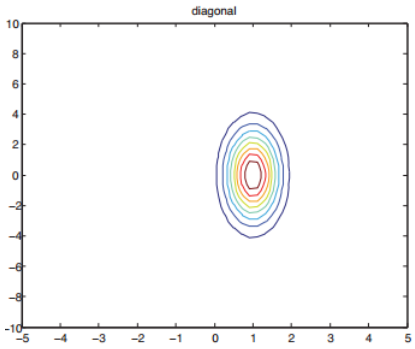
$$\text{Cov}[\vec{y}] = \text{Cov}[A\vec{x} + b] = A\Sigma A^T \tag{0.43}$$

0.5.2 General transformations

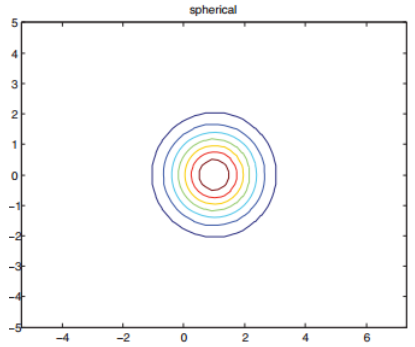
If X is a discrete rv, we can derive the pmf for y by simply summing up the probability mass for all the x's such that $f(x) = y$:



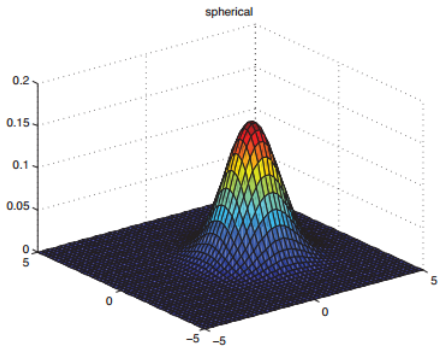
(a)



(b)

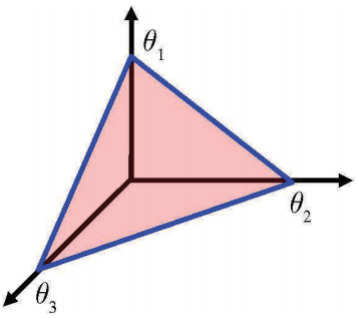


(c)

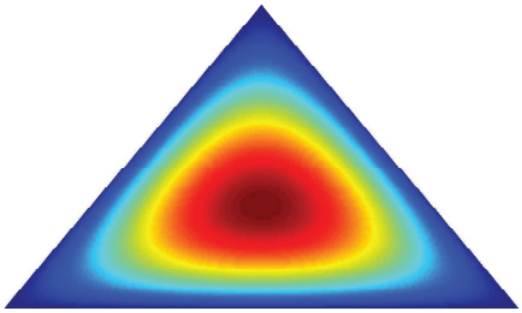


(d)

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).

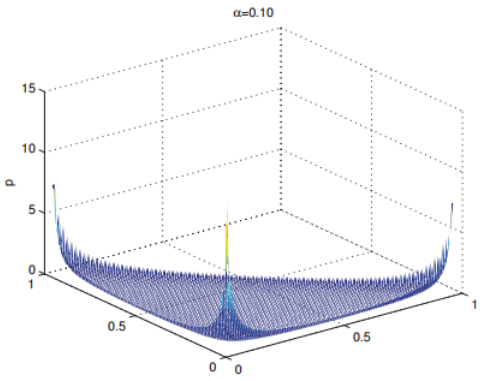


(a)



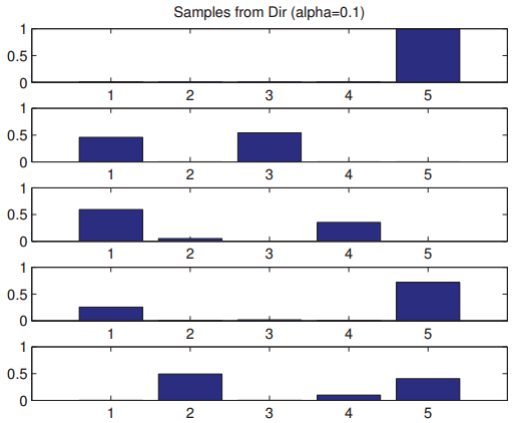
(b)

(c)

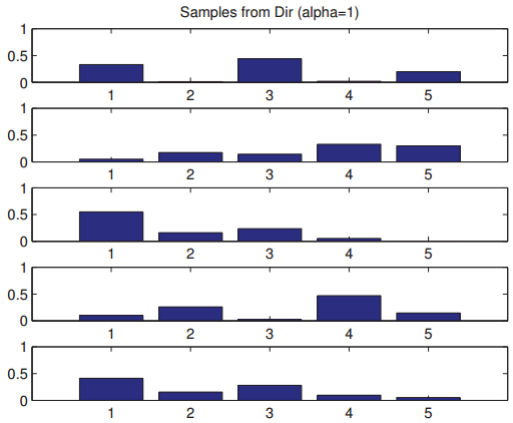


(d)

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.

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

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

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

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

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

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 $p_Y(y) = \frac{1}{2}y^{-\frac{1}{2}}$.

0.5.2.1 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 $\vec{y} \to \vec{x}$:

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

0.5.3 Central limit theorem

Given N random variables $X_1, X_2, \cdots, X_N$, each variable is independent and identically distributed[3](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

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

0.6 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 Chap-

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

ter 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[4], 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:

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

where $x_s \sim p(X)$.

This is called Monte Carlo integration[5], 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.

0.7 Information theory

0.7.1 Entropy

একটি random variable X এর entropy, যার distribution p দ্বারা নির্ধারিত, এবং যেটা $\mathbb{H}(X)$ বা কখনও কখনও $\mathbb{H}(p)$ দ্বারা চিহ্নিত হয়, এটি random variable- এর uncertainty এর একটি পরিমাপ। বিশেষ করে, যদি X একটি discrete ভেরিয়েবল হয় যার K-টি states আছে, তবে এটি নিম্নরূপ সংজ্ঞায়িত করা হয়:

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

- ইকুয়েশনটি একটি discrete random variable X এর জন্য entropy, যেখানে X এর একটি probability distribution p রয়েছে

- $\mathbb{H}(X)$: এটি random variable X এর entropy। Entropy দ্বারা X এর distribution এ কতটা uncertainty বা randomness আছে সেটা হিসাব করা হচ্ছে

- $\sum_{k=1}^{K}$: এই summation চিহ্নটি নির্দেশ করে যে আমরা X এর সমস্ত সম্ভাব্য states বা values এর উপর যোগফল নিচ্ছি, যেখানে K হল states এর সংখ্যা।

- $p(X=k)$: random variable X, k এর একটি নির্দিষ্ট মান গ্রহণ করার সম্ভাবনা। সুতরাং, $p(X=k)$ প্রতিটি state k এর জন্য probability নির্দেশ করে।

- $\log_2 p(X=k)$: এটি $p(X=k)$ এর base 2 logarithm। এখানে base 2 logarithm ব্যবহার করা হচ্ছে কারণ entropy কে "bits" এ পরিমাপ করা হচ্ছে, যা digital systems এর ক্ষেত্রে প্রযোজ্য।

- entropy-এর সম্পূর্ণ সমীকরণ: এই সমীকরণটি প্রতিটি probability $p(X=k)$ কে তার লগারিদম $\log_2 p(X=k)$ এর সাথে গুণ করে, তারপর সমস্ত states k এর জন্য যোগ করে, এবং শেষে এটিকে -1 দ্বারা গুণ করা হচ্ছে।

এর কাজ কী: Entropy পরিমাপ করে একটি probability distribution এর মধ্যে কতটা uncertainty বা information content রয়েছে। সহজ ভাষায়, এটি আমাদের বলে দেয় কতটা unpredictable হবে random variable X এর মান।

- যদি সমস্ত states সমান সম্ভাবনা নিয়ে থাকে, তাহলে entropy বেশি হবে, কারণ X এর মান সহজে পূর্বানুমান করা যায় না।

- যদি একটি state এর সমস্ত সম্ভাবনা থাকে (i.e., certaintly), তবে entropy কম বা শূন্য হবে, কারণ এখানে কোনও uncertainty নেই।

সাধারণত আমরা log এর base 2 ব্যবহার করি, এবং এই ক্ষেত্রে এর units কে bits (binary digits এর সংক্ষিপ্ত রূপ) বলা হয়। যদি আমরা log এর base e ব্যবহার করি, units গুলোকে nats বলা হয়।

সর্বোচ্চ entropy সহ একটি discrete distribution হল uniform distribution (প্রমাণের জন্য Section XXX দ্রষ্টব্য)। সুতরাং, যদি একটি K-ary random variable এর ক্ষেত্রে $p(x=k)=1/K$, তাহলে entropy সর্বাধিক হয়; এই ক্ষেত্রে $\mathbb{H}(X)=\log_2 K$।

বিপরীতভাবে, সর্বনিম্ন entropy সহ একটি distribution (যা শূন্য) হল যেকোনো delta-function, যা এর সব mass একটি state এ ধরে রাখে। এমন একটি distribution এ কোনো uncertainty থাকে না।

0.7.2 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[6]. 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

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

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

[6] 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 $q=0.5p_1+0.5p_2$

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

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.

0.7.3 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:

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

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. ??[7].

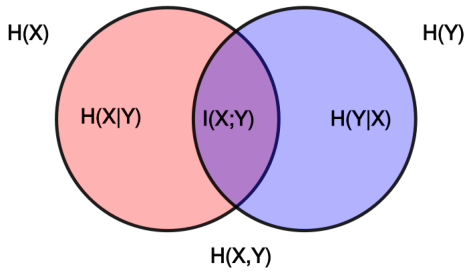


Figure 0.10: Individual $\mathbb{H}(X), \mathbb{H}(Y)$, joint $\mathbb{H}(X,Y)$, and conditional entropies for a pair of correlated subsystems X, Y with mutual information $\mathbb{I}(X;Y)$.

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

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

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:

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

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 $p(y|x)$.

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