Introduction to Probability and Statistics (Continued)

Prof. Nicholas Zabaras
Center for Informatics and Computational Science
https://cics.nd.edu/
University of Notre Dame
Notre Dame, Indiana, USA

Email: nzabaras@gmail.com
URL: https://www.zabaras.com/

August 29, 2018

Contents

- Markov and Chebyshev Inequalities
- The Law of Large Numbers, Central Limit Theorem, Monte Carlo Approximation of Distributions, Estimating π, Accuracy of Monte Carlo approximation, Approximating the Binomial with a Gaussian, Approximating the Poisson Distribution with a Gaussian, Application of CLT to Noise Signals
- Information theory, Entropy, KL divergence, Jensen's Inequality, Mutual information, Maximal Information Coefficient

References

- Following closely Chris Bishops' PRML book, Chapter 2
- Kevin Murphy's, Machine Learning: A probablistic perspective, Chapter 2
- Jaynes, E. T. (2003). Probability Theory: The Logic of Science. Cambridge University Press.
- Bertsekas, D. and J. Tsitsiklis (2008). Introduction to Probability. Athena Scientific. 2nd Edition
- Wasserman, L. (2004). All of statistics. A Concise Course in Statistical Inference. Springer.

Markov and Chebyshev Inequalities

- You can show (Markov's inequality) that if X is a non-negative integrable random variable and for any $a > 0$:

$$\Pr[X \geq a] \leq \frac{\mathbb{E}[X]}{a}$$

$$Indeed: \quad \mathbb{E}[X] = \int_0^\infty x\pi(x)dx \geq \int_a^\infty x\pi(x)dx \geq a\int_a^\infty \pi(x)dx = a\Pr[X \geq a]$$

- You can generalize this using any function of the random variable X as:

$$\Pr[f(X) \geq a] \leq \frac{\mathbb{E}[f(X)]}{a}$$

- Using $f(X) = (X - \mathbb{E}[X])^2$, we derive the following Chebyshev inequality:

$$a \equiv \varepsilon^2, \mathbb{E}\left[\left|X - \mathbb{E}[X]\right|^2\right] = \sigma^2 \qquad \Pr\left[\left|X - \mathbb{E}[X]\right| \geq \varepsilon\right] \leq \frac{\sigma^2}{\varepsilon^2} \forall \varepsilon$$

- In terms of std's, we can restate as : $\Pr\left[\left|X - \mathbb{E}[X]\right| \geq \varepsilon\sigma\right] \leq \frac{1}{\varepsilon^2} \forall \varepsilon$

- Thus the probability of X being more than 2σ away from $\mathbb{E}[X]$ is ≤ ¼.

The Law of Large Numbers (LLN)

- Let X_i for $i = 1, 2, \ldots, n$ be independent and identically distributed random variables (i.i.d.) with finite mean $\mathbb{E}(X_i) = \mu$ & variance $\text{Var}(X_i) = \sigma^2$.
- Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$
- Note that

$$\mathbb{E}[\bar{X}_n] = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

$$Var[\bar{X}_n] = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}$$

- Weak LLN: $\lim_{n\to\infty} \Pr[|\bar{X}_n - \mu| \geq \varepsilon] = 0 \; \forall \varepsilon > 0$
- Strong LLN: $\lim_{n\to\infty} \bar{X}_n = \mu \; almost\; surely$

This means that with probability one, the average of any realizations of $x_1, x_2, \ldots$ of the random variables $X_1, X_2, \ldots$ converges to the mean.

Statistical Inference: Parametric & Non-Parametric Approach

Assume that we have a set of observations

$$S = \{x_1, x_2, ...x_N\}, \quad x_j \in \mathbb{R}^n$$

The problem is to infer on the underlying probability distribution that gives rise to the data *S.*

- *Parametric problem: The underlying probability density has a specified known form that depends on a number of parameters. The problem of interest is to infer* those parameters.

- *Non-parametric problem: No analytical expression for the probability den*sity is available. Description consists of defining the dependency or independency of the data. This leads to numerical exploration.

A typical situation for the parametric model is when the distribution is the PDF of a random variable $X : \Omega \to \mathbb{R}^n$.

Example of the Law of Large Numbers

- Assume that we sample $S = \{x_1, x_2, \ldots x_N\}, \quad x_j \in \mathbb{R}^2$
- We consider a parametric model with x_j realizations of $X \sim \mathcal{N}(x_0, \Sigma)$

  where we take both the mean x_0 and the variance $\Sigma \in \mathbb{R}^{2\times 2}$ as unknowns.

- The probability density of X is:

$$\pi(x \mid x_0, \Sigma) = \frac{1}{2\pi\left(\det \Sigma\right)^{1/2}} \exp\left(-\frac{1}{2}(x - x_0)^T \Sigma^{-1}(x - x_0) \right)$$

- Our problem is to estimate x_0 and $\Sigma \in \mathbb{R}^{2\times 2}$

Empirical Mean and Empirical Covariance

- From the law of large numbers, we calculate:

$$x_0 = \mathbb{E}[X] \approx \frac{1}{N}\sum_{j=1}^{N} x_j = \bar{x}$$

- To compute the covariance matrix, note that if $X_1, X_2, \ldots$ are i.i.d. so are $f(X_1), f(X_2), \ldots$ for any function $f : \mathbb{R}^2 \to \mathbb{R}^k$

- Then we can compute:

$$\boldsymbol{\Sigma} = \mathrm{cov}(X) = \mathbb{E}\left[(x - \mathbb{E}[X])(x - \mathbb{E}[X])^T\right] \approx \mathbb{E}\left[(x - \bar{x})(x - \bar{x})^T\right]$$

$$\Rightarrow \quad \boldsymbol{\Sigma} \approx \frac{1}{N}\sum_{j=1}^{N}(x_j - \bar{x})\left(x_j - \bar{x}\right)^T = \overline{\boldsymbol{\Sigma}}$$

- The above formulas define **the empirical mean and empirical covariance.**

The Central Limit Theorem

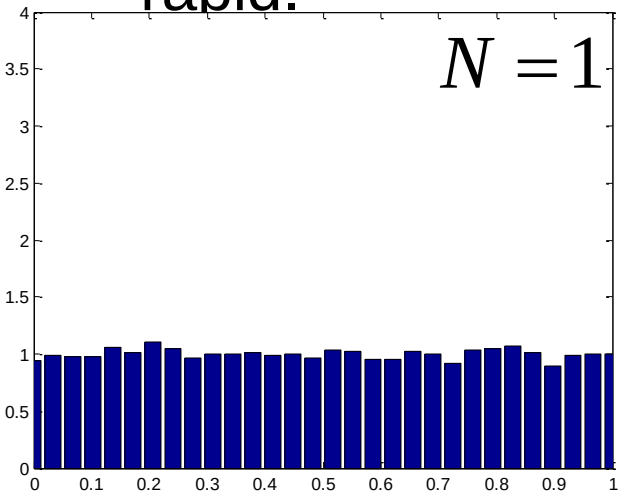
- Let $(X_1, X_2, \ldots X_N)$ be independent and identically distributed (i.i.d.) continuous random variables each with expectation μ and variance σ^2.
- Define: $Z_N = \frac{1}{\sigma\sqrt{N}}(X_1 + X_2 + \ldots + X_N - N\mu) = \frac{\bar{X}_N - \mu}{\frac{\sigma}{\sqrt{N}}}$, $\bar{X}_N = \frac{1}{N}\sum_{j=1}^{N} X_j$
- As $N -> \infty$, the distribution of Z_N converges to the distribution of a standard normal random variable

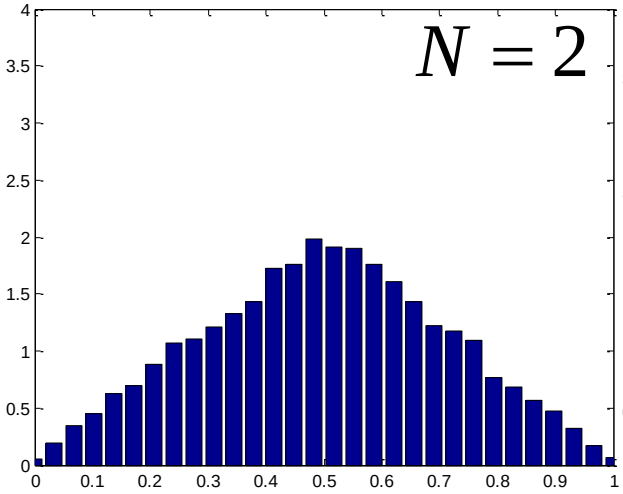
$$\lim_{N\to\infty} P\{Z_N \le x\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt$$

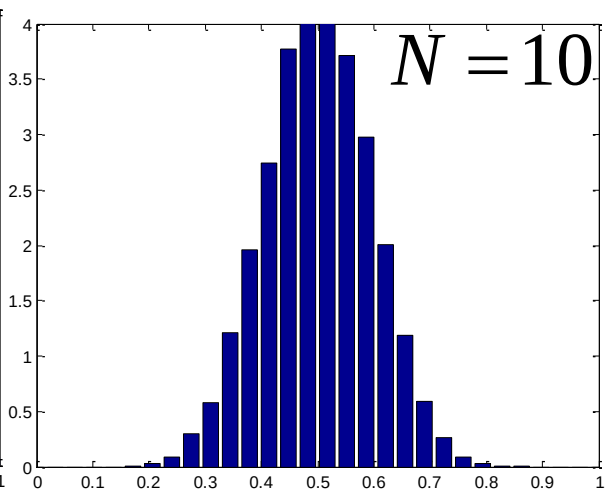
- If $\bar{X}_N = \frac{1}{N}\sum_{j=1}^{N} X_j$, for N large, $\bar{X}_N \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{N}\right)$ $as\ N \to \infty$
- Somewhat of *a justification for assuming Gaussian noise is common*

The CLT and the Gaussian Distribution

- As an example, assume N variables $(X_1, X_2, \ldots X_N)$ each of which has a uniform distribution over $[0,1]$ and then consider the distribution of the sample mean $(X_1 + X_2 + \ldots + X_N)/N$. For large N, this distribution tends to a Gaussian. The convergence as N increases can be rapid.



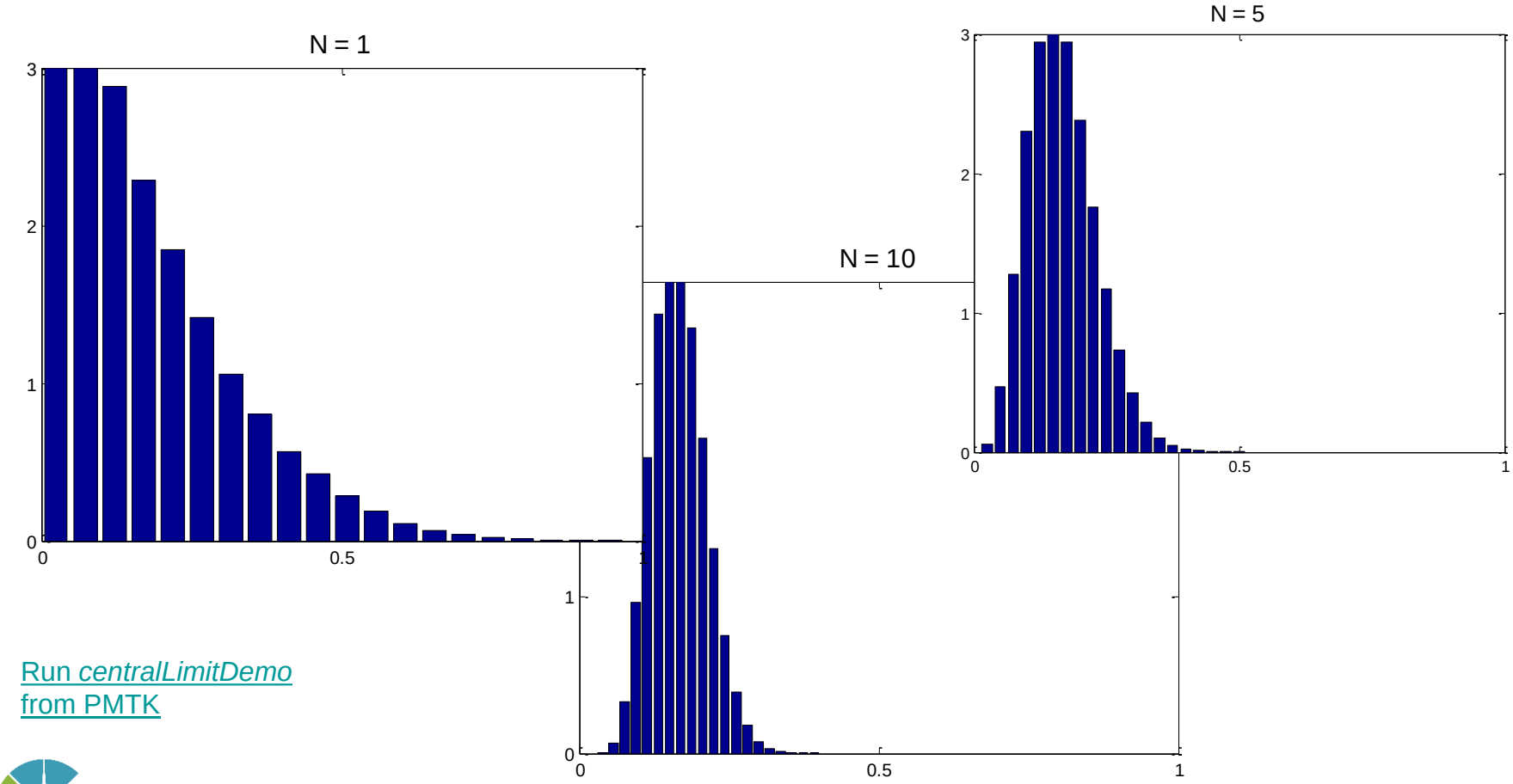




MatLab Code

The CLT and the Gaussian Distribution

- We plot a histogram of $\frac{1}{N}\sum_{i=1}^{N} x_{ij}, j = 1:10000$ *where* $x_{ij} \sim \mathcal{Beta}(1,5)$
- As $N \to \infty$, the distribution tends towards a Gaussian.



Run *centralLimitDemo* from PMTK

Accuracy of Monte Carlo Approximation

- In Monte Carlo approximation of the mean using the sample mean $\bar{\mu}$ approximation, we have:

$$For\ \mu = \mathbb{E}[f(x)]$$

$$\bar{\mu} - \mu \sim \mathcal{N}\left(0, \frac{\sigma^2}{N}\right)$$

$$\sigma^2 = \mathbb{E}[f^2(x)] - \mathbb{E}[\mathrm{f(x)}]^2 \approx \frac{1}{N}\sum_{s=1}^{N}(f(x_s) - \bar{\mu})^2 \equiv \bar{\sigma}^2$$

- We can now derive the following error bars (using central intervals):

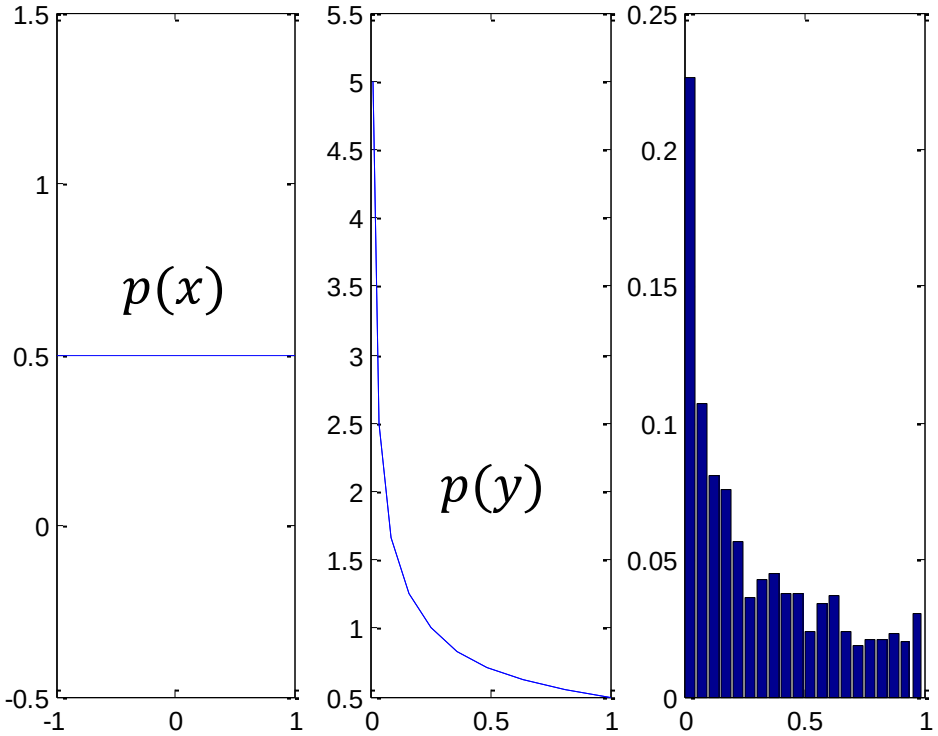
$$\Pr\left\{\mu - 1.96\frac{\bar{\sigma}}{\sqrt{N}} \leq \bar{\mu} \leq \mu + 1.96\frac{\bar{\sigma}}{\sqrt{N}}\right\} = 0.95$$

- The number of samples needed to drop the error within $\pm\varepsilon$ is then:

$$1.96\frac{\bar{\sigma}}{\sqrt{N}} \leq \varepsilon \Rightarrow N \geq \frac{4\bar{\sigma}^2}{\varepsilon^2}$$

Monte Carlo Approximation of Distributions

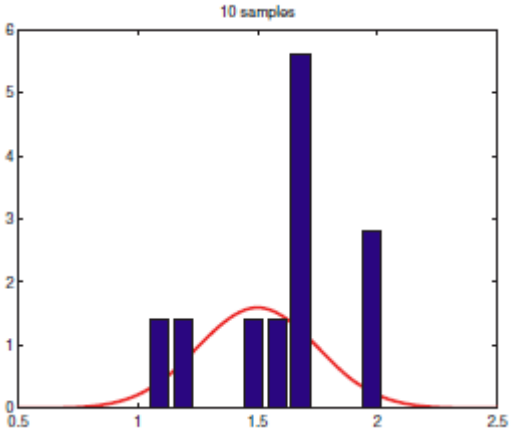
- Computing the distribution of $y = x^2$, $p(x)$ is uniform. The MC approximation is shown on the right. Take samples from $p(x)$, square them and then plot the histogram.

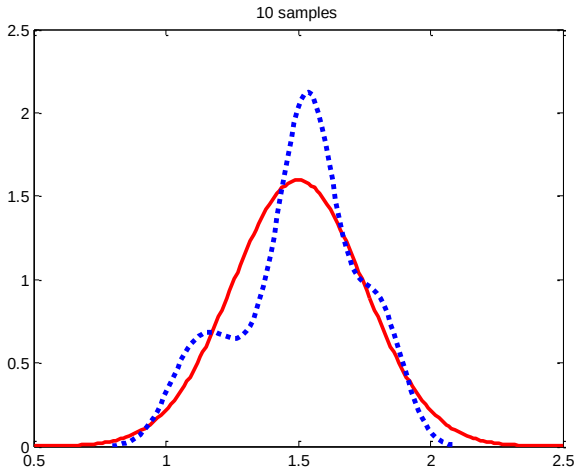


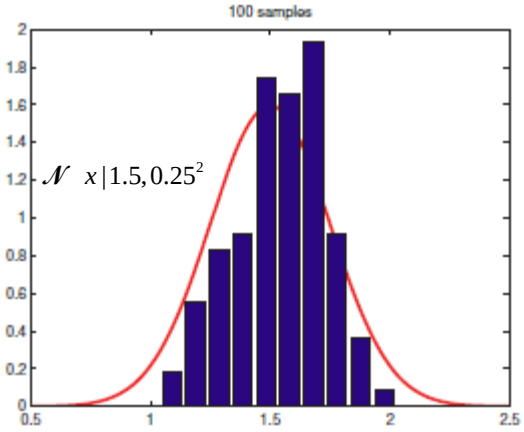
Run *changeofVarsDemo1d* from PMTK

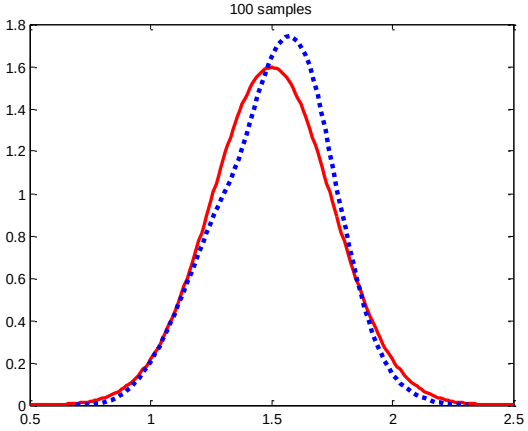
Accuracy of Monte Carlo Approximation

- Increase of the accuracy of MC with the number of samples. Histograms (on the left) and (on the right) pdfs using kernel density estimation.
- The actual distribution is shown on red.









Run *mcAccuracyDemo* from PMTK

Example of CLT: Estimating π by MC

- Use the CLT to approximate π. Let $x, y \sim \mathcal{U}[-r, r]$.

$$I = \int_{-r}^{r}\int_{-r}^{r} \mathbb{I}(x^2 + y^2 \le r^2)dxdy = \pi r^2 \Rightarrow$$

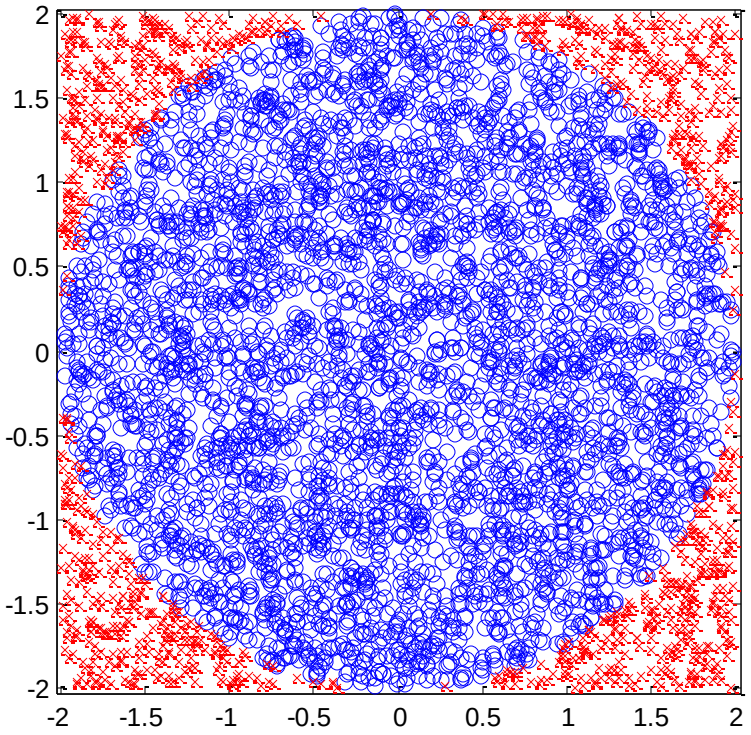
$$\pi = \frac{1}{r^2} I = \frac{1}{r^2} 4r^2 \int_{-r}^{r}\int_{-r}^{r} \mathbb{I}(x^2 + y^2 \le r^2)\, p(x)\, p(y)dxdy \Rightarrow$$

$$\pi = 4\int_{-r}^{r}\int_{-r}^{r} \mathbb{I}(x^2 + y^2 \le r^2)\, p(x)\, p(y)dxdy \Rightarrow$$

$$\bar{\pi} \approx 4\frac{1}{N}\sum_{s=1}^{N} \mathbb{I}(x_s^2 + y_s^2 \le r^2)\,, x_s, y_s \sim \mathcal{U}[-r, r]$$

- Here x, y are uniform random variables on $[-r, +r], r = 2$.

- We find $\bar{\pi} =$ 3.1416 with standard error 0.09.



$x, y \sim \mathcal{U}[-r, r]$

Run *mcEstimatePi* from PMTK

CLT: The Binomial Tends to a Gaussian

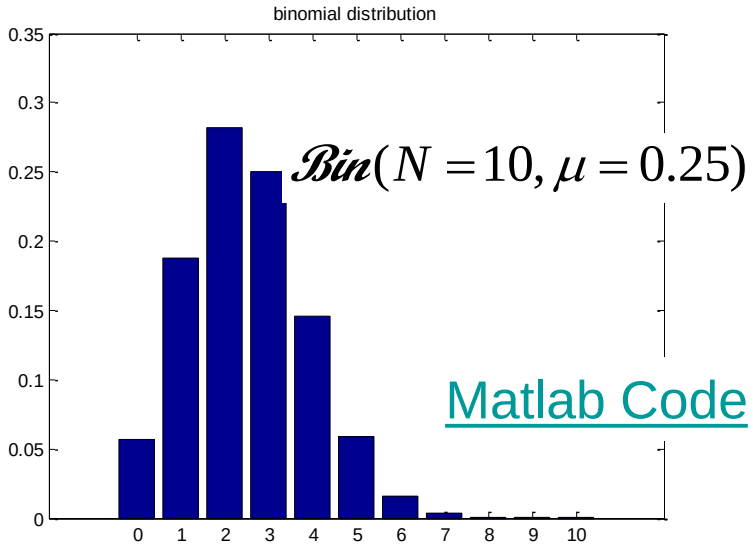
- One consequence of the CLT is that *the binomial distribution*

$$\mathcal{Bin}\ m \mid N, \mu\ = \binom{N}{m} \mu^m (1-\mu)^{N-m}$$

which is a distribution over m defined by the sum of N observations of the random binary variable x, will tend to a Gaussian as $N \to \infty$.

$$\frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{m}{N} \sim \mathcal{N}\left(\mu, \frac{\mu\ 1-\mu}{N}\right) \Rightarrow$$

$$m \sim \mathcal{N}\ N\mu, N\mu\ 1-\mu$$



Poisson Process

Consider that we count the number of photons from a light source. Let $N(t)$ be the number of photons observed in the time interval $[0, t]$. $N(t)$ is an integer-valued random variable. We make the following assumptions:

a. **Stationarity:** Let Δ_1 and Δ_2 be any two time intervals of equal length, n any non-negative integer. Assume that

Prob. of n photons in Δ_1 = Prob. of n photons in Δ_2

b. **Independent increments:** Let $\Delta_1, \Delta_2, \ldots, \Delta_n$ be non-overlapping time intervals and $k_1, k_2, \ldots, k_n$ non-negative integers. Denote by A_j the event defined as

$A_j = k_j$ photons arrive in the time interval Δ_j

Assume that these events are mutually independent, i.e.

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1)P(A_2)\ldots P(A_n)$$

c. **Negligible probability of coincidence:** Assume that the probability of two or more events at the same time is negligible. More precisely, $N(0) = 0$ and

$$\lim_{h \to 0} \frac{P\{N(h) > 1\}}{h} = 0$$

Poisson Process

- If these assumptions hold, then for a given time t, N is a Poisson process:

$$P\{N(t)=n\}=\frac{(\lambda t)^n}{n!}e^{-\lambda t},\ \lambda>0,\ n=0,1,2,\ldots,\infty$$

- Let us *fix $t = T$ = observation time and define a random variable $N = N(T)$.* Let us define the parameter $\theta = \lambda T$. We then denote:

$$N \sim \textit{Poisson}(\theta)=\frac{\theta^n}{n!}e^{-\theta}$$

- *D. Calvetti and E. Somersalo, Introduction to Bayesian Scientific Computing, 2007*
- *S Ghahramani: Fundamentals of Probability,1996.*

Poisson Process

- Consider the Poisson (discrete) distribution $N \in \{0,1,2,...,\infty\}$

$$P(N=n) = \pi_{Poisson}(n \mid \theta) = \frac{\theta^n}{n!} e^{-\theta}$$

- The mean and the variance are both equal to θ.

$$\mathbb{E}[N] = \sum_{n=0}^{\infty} n \pi_{Poisson}(n \mid \theta) = \theta,$$

$$\mathbb{E}\left[(N-\theta)^2\right] = \theta$$

Approximating a Poisson Distribution With a Gaussian

- *Theorem: A random variable $X \sim \mathcal{Poisson}(\theta)$ can be considered as the sum of n independent random variables $X_i \sim \mathcal{Poisson}(\theta/n)$.*[a]

- According to the Central Limit Theorem, when n is large enough,

$$Take\ X_i \sim \mathcal{Poisson}(\frac{\theta}{n}, \frac{\theta}{n}) \Rightarrow \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}(\frac{\theta}{n}, \frac{\theta}{n^2})$$

- Then $X = \sum_{i=1}^{n} X_i$ based on the Theorem is a draw from $\mathcal{Poisson}(\theta)$ and from the CLT also follows a Gaussian for large *n* with:

$$\mathbb{E}[X] = n\frac{\theta}{n} = \theta$$

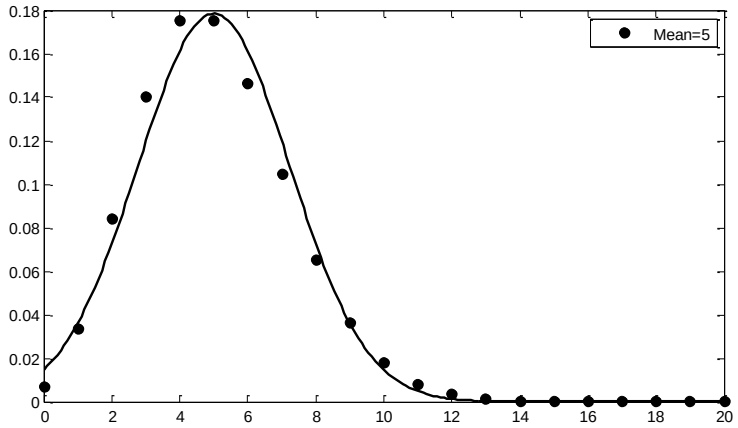
$$var[X] = n^2\frac{\theta}{n^2} = \theta$$

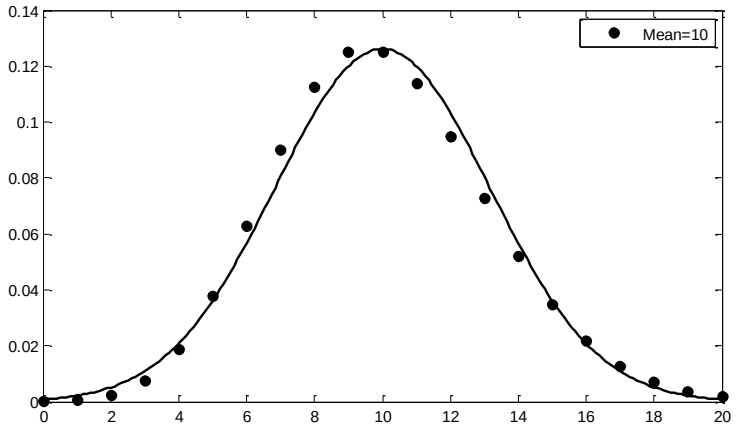
- Thus $X \sim \mathcal{N}(\theta, \theta)$

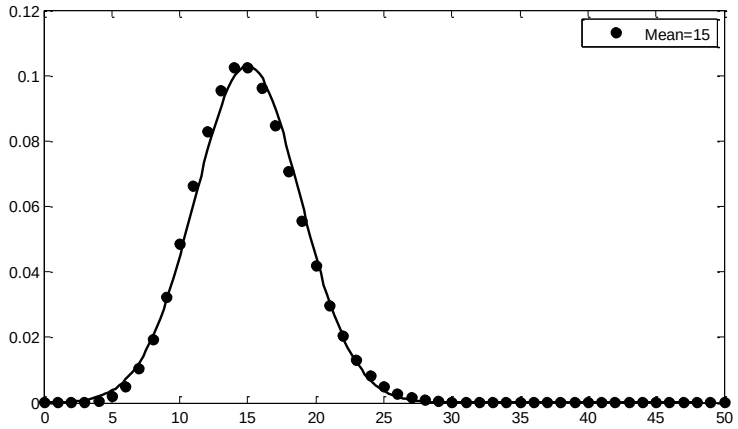
- The approximation of a Poisson distribution with a Gaussian for large n is a result of the CLT!

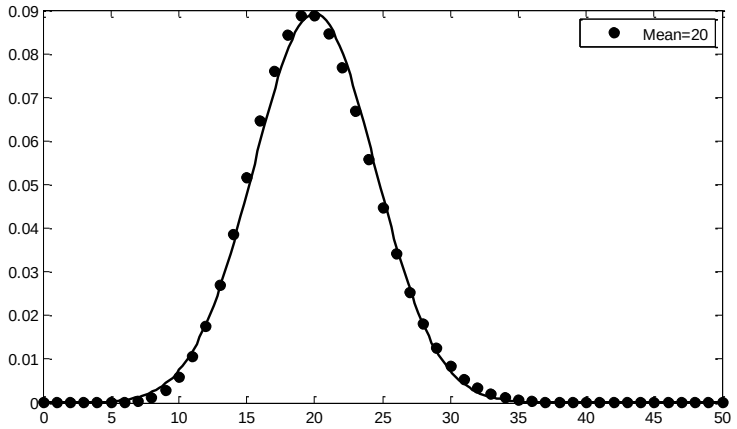
[a] For a proof that the sum of independent Poisson Random Variables is a Poisson distribution see this document.

Approximating a Poisson Distribution with a Gaussian









Poisson distributions (dots) vs their Gaussian approximations (solid line) for various values of the mean θ. The higher the θ, the smaller the distance between the two distributions. *See this MatLab implementation.*

Kullback-Leibler Distance Between Two Densities

- Let us consider the following two distributions:

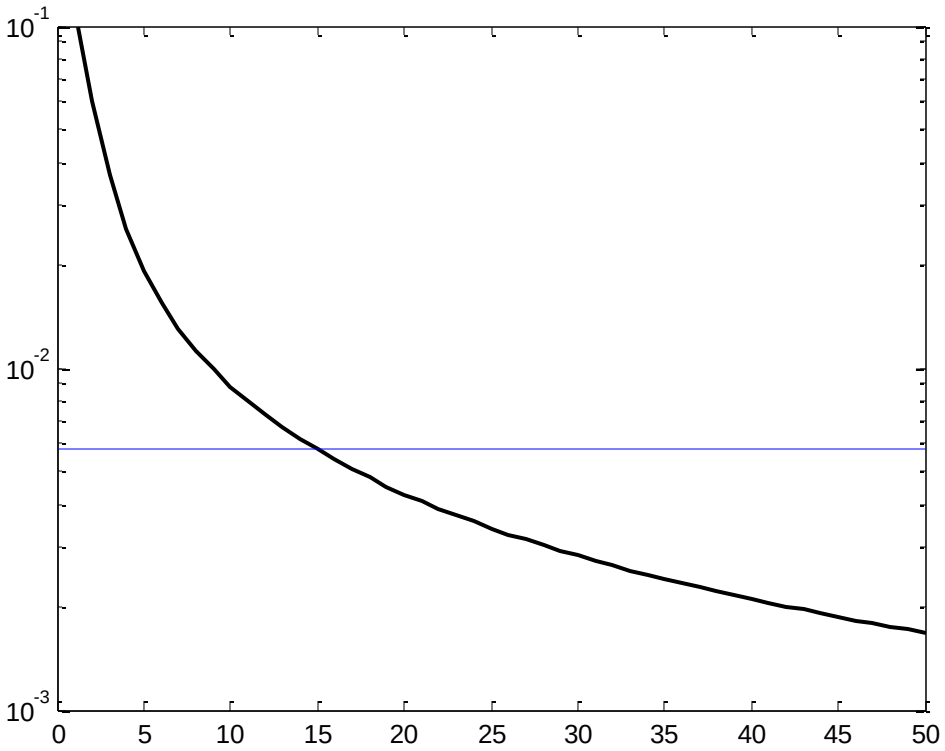
$$n \to \pi_{Poisson}(n \mid \theta) = \frac{\theta^n}{n!} e^{-\theta}$$

$$x \to \pi_{Gaussian}(x \mid \theta, \theta) = \frac{1}{\sqrt{2\pi\theta}} \exp\left(-\frac{1}{2\theta}(x-\theta)^2 \right)$$

- We often use the Kullback-Leibler distance to define the distance between two distributions. In particular, in approximating the Poisson distribution with a Gaussian distribution, we have the following:

$$KL\ distance\left(\pi_{Poisson}(.\mid\theta), \pi_{Gaussian}(.\mid\theta,\theta)\right) = \sum_{n=0}^{\infty} \pi_{Poisson}(n\mid\theta) \log \frac{\pi_{Poisson}(n\mid\theta)}{\pi_{Gaussian}(n\mid\theta,\theta)}$$

Approximating a Poisson Distribution With a Gaussian



The KL distance of the Poisson distribution from its Gaussian approximation as a function of the mean θ *in a logarithmic scale. The horizontal* line indicates where the KL distance has dropped to 1/10 of its value at $\theta = 1$.

See the following MatLab implementation.

Application of the CLT: Noise Signals

- Consider discrete sampling where the output is noise of length n.
- The noise vector $x \in \mathbb{R}^n$ is a realization of $X : \Omega \to \mathbb{R}^n$
- We estimate the mean and the variance of the noise in a single measurement as follows:

$$x_0 = \frac{1}{n}\sum_{j=1}^{n} x_j, \quad \sigma^2 = \frac{1}{n}\sum_{j=1}^{n}\left(x_j - x_0\right)^2$$

- To improve the signal to noise ratio, we repeat the measurement and average the noise vector signals:

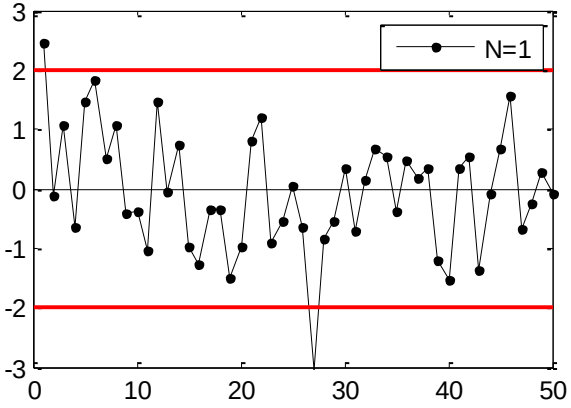
$$x = \frac{1}{N}\sum_{k=1}^{N} x^{(k)} \in \mathbb{R}^n$$

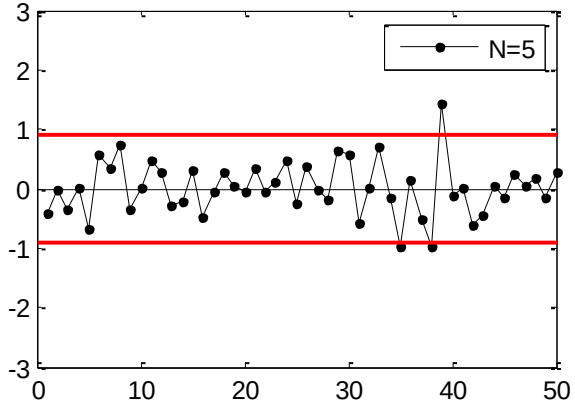
- The average noise is a realization of a random variable: $X = \frac{1}{N}\sum_{k=1}^{N} X^{(k)} \in \mathbb{R}^n$
- If $X^{(1)}, X^{(2)}, \ldots$ are i.i.d., X is asymptotically a Gaussian by the CLT, and its variance is $\mathrm{var}(X) = \sigma^2/N$. We need to repeat the experiment until $\sigma^2/N < \tau^2$ (a given tolerance).

Noise Reduction By Averaging

Gaussian noise vectors of size 50 are used with $\sigma^2 = 1$

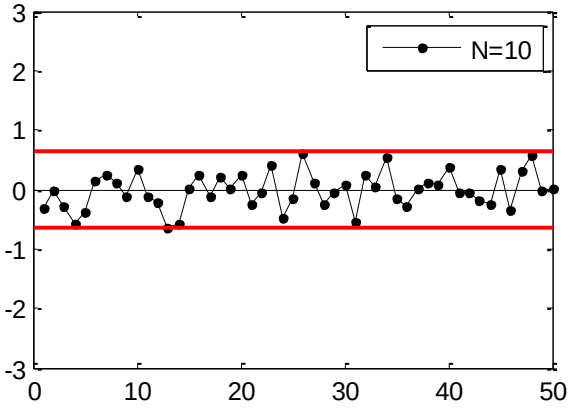
σ is the std of a single noise vector

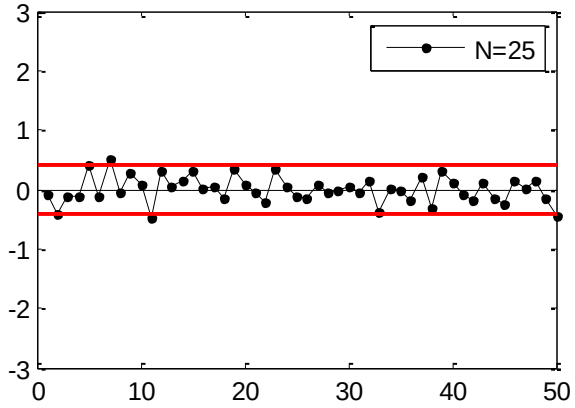




Estimated Noise level :

$$2\frac{\sigma}{\sqrt{N}}$$





See the following MatLab implementation.

- *D. Calvetti and E. Somersalo, Introduction to Bayesian Scientific Computing, 2007*

Introduction to Information Theory

- ***Information theory*** is concerned
  - with representing data in a compact fashion (***data compression*** *or* ***source coding***), and
  - transmitting and storing it in a way that is robust to errors (***error correction or channel coding***).

- To compactly representing data requires *allocating short codewords to highly probable bit strings*, and reserving longer codewords to less probable bit strings.
  - e.g. in natural language, common words (“a”, “the”, “and”) are much shorter than rare words.

- D. MacKay, Information Theory, Inference and Learning Algorithms (Video Lectures)

Introduction to Information Theory

- Decoding messages sent over noisy channels requires having a good probability model of the kinds of messages that people tend to send.

- We need *models that can predict which kinds of data are likely and which unlikely*.

- David MacKay, Information Theory, Inference and Learning Algorithms , 2003 (available on line)
- Thomas M. Cover, Joy A. Thomas , Elements of Information Theory , Wiley, 2006.
- Viterbi, A. J. and J. K. Omura (1979). *Principles of Digital Communication and Coding*. McGraw-Hill.

Introduction to Information Theory

- Consider a discrete random variable x. We ask how much information ('degree of surprise') is received when we observe (learn) a specific value for this variable?

- Observing a highly probable event provides little additional information.

- If we have two events x and y that are unrelated, then the information gain from observing both of them should be $h(x, y) = h(x) + h(y)$.

- Two unrelated events will be statistically independent, so $p(x, y) = p(x)p(y)$.

Entropy

- From $h(x,y) = h(x) + h(y)$ and $p(x,y) = p(x)p(y)$, it is easily shown that $h(x)$ must be given by the logarithm of $p(x)$ and so we have

$$h(x) = -\log_2 p(x) \geq 0$$

the units of *h(x)* are bits ('binary digits')

- Low probability events correspond to high information content.

- When transmitting a random variable**, the average amount of transmitted information is:**

$$\textit{Entropy of } X : \mathbb{H}[X] = -\sum_{k=1}^{K} p(X=k)\log_2 p(X=k)$$

Noiseless Coding Theorem (Shanon)

- **Example 1** (Coding theory): x discrete random variable with 8 possible states; how many bits to transmit the state of x?

  All states equally likely $\quad \mathbb{H}[x] = -8 \times \frac{1}{8} \log_2 \frac{1}{8} = 3\, bits$

- **Example 2**: consider a variable having 8 possible states $\{a, b, c, d, e, f, g, h\}$ for which the respective (non-uniform) probabilities are given by $(1/2, 1/4, 1/8, 1/16, 1/64, 1/64, 1/64, 1/64)$.

  The entropy in this case is smaller than for the uniform distribution.

x	a	b	c	d	e	f	g	h
$p(x)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{64}$	$\frac{1}{64}$	$\frac{1}{64}$	$\frac{1}{64}$
code	0	10	110	1110	111100	111101	111110	111111

Note: shorter codes for the more probable events vs longer codes for the less probable events.

$$\mathbb{H}[x] = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{8}\log_2\frac{1}{8} - \frac{1}{16}\log_2\frac{1}{16} - \frac{4}{64}\log_2\frac{1}{64} = 2\, bits$$

$$average\ code\ length = \frac{1}{2}\times 1 + \frac{1}{4}\times 2 + \frac{1}{8}\times 3 + \frac{1}{16}\times 4 + 4\times\frac{1}{64}\times 6 = 2\, bits$$

Shanon's Noiseless Coding Theorem (1948): The entropy is a lower bound on the number of bits needed to transmit the state of a random variable

Alternative Definition of Entropy

- Considering a set of N identical objects that are to be divided amongst a set of bins, such that there are n_i objects in the i^{th} bin. Consider the number of different ways of allocating the objects to the bins.

- In the i^{th} bin there are $n_i!$ ways of reordering the objects (microstates), and so the total number of ways of allocating the N objects to the bins is given by (multiplicity)

$$W = \frac{N!}{\prod_i n_i!}$$

- The entropy is defined as $\mathbb{H} = \frac{1}{N}\ln W = \frac{1}{N}\ln N! - \frac{1}{N}\sum_i \ln n_i!$

- We now consider the limit $N \to \infty$, $\ln N! \approx N\ln N - N, \ln n_i! \approx n_i \ln n_i - n_i$

$$\mathbb{H} = -\lim_{N\to\infty}\sum_i \frac{n_i}{N}\ln\frac{n_i}{N} = -\sum_i p_i \ln p_i$$

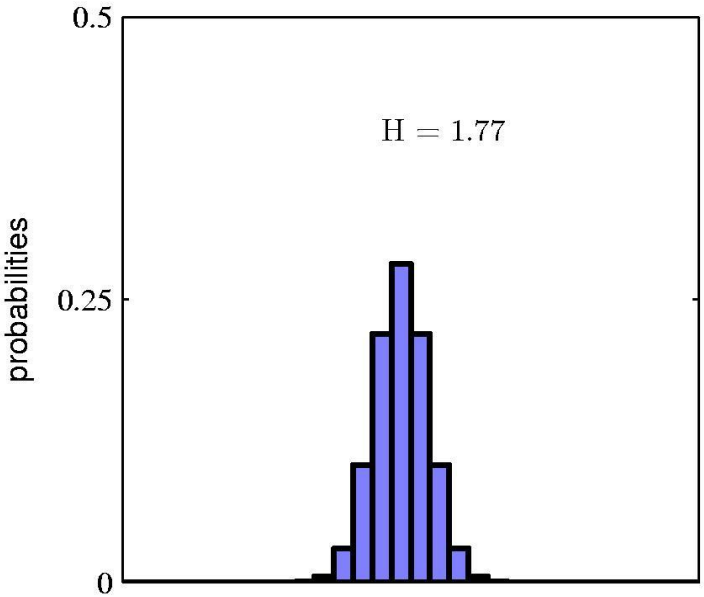
- p_i is the probability of an object assigned to the i^{th} bin.
- The occupation numbers p_i correspond to macrostates.

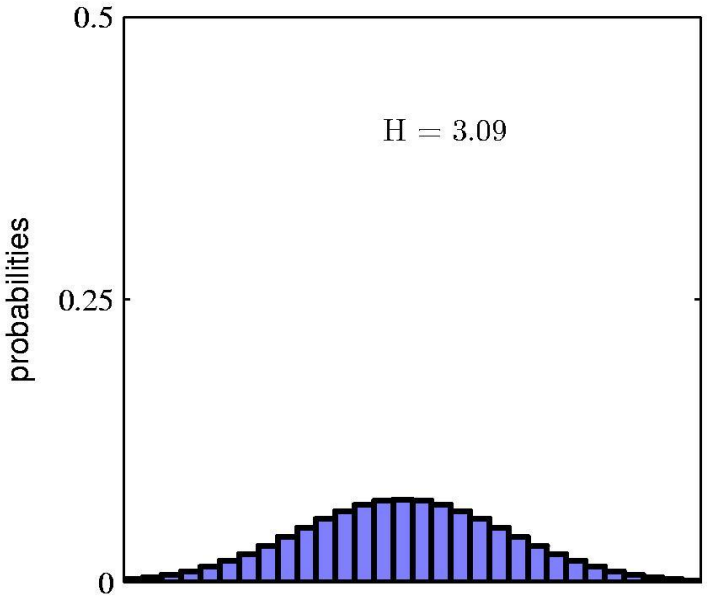
Alternative Definition of Entropy

- Interpret the bins as the states x_i of a discrete random variable X, where $p(X = x_i) = p_i$. The entropy of the random variable X is then

$$\mathbb{H}[p] = -\sum_i p(x_i) \ln p(x_i)$$

- *Distributions $p(x)$ that are sharply peaked around a few values will have a relatively low entropy, whereas those that are spread more evenly across many values will have higher entropy.*





Maximum Entropy: Uniform Distribution

- *The maximum entropy configuration* can be found by maximizing $\mathbb{H}$ using a Lagrange multiplier to enforce the normalization constraint on the probabilities. Thus we maximize

$$\overline{\mathbb{H}} = -\sum_i p(x_i)\ln p(x_i) + \lambda\left(\sum_i p(x_i) - 1\right)$$

- We find $p(x_i) = 1/M$, M is the number of possible states and $\mathbb{H} = \ln_2 M$.

- To verify that the stationary point is indeed a maximum, we can evaluate the 2nd derivative of the entropy, which gives

$$\frac{\partial^2 \overline{\mathbb{H}}}{\partial p(x_i)\partial p(x_j)} = -I_{ij}\frac{1}{p_i}$$

where I_{ij} are the elements of the identity matrix.

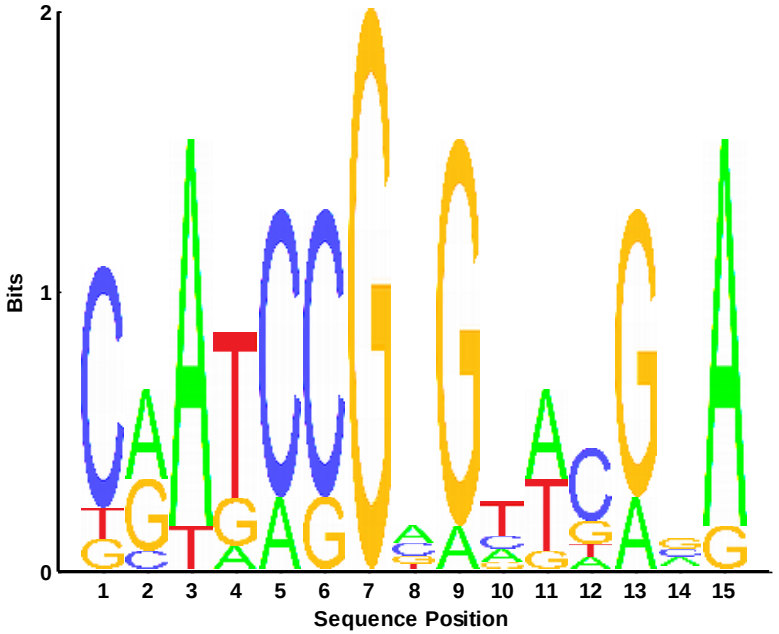
- ***For any discrete distribution with M states***, we have: $\mathbb{H}[x] \leq \ln_2 M$

$$\mathbb{H} = -\sum_i p(x_i)\ln p(x_i) = \sum_i p(x_i)\ln\frac{1}{p(x_i)} \leq \ln\sum_i p(x_i)\frac{1}{p(x_i)} = \ln M$$

- Here we used Jensen's inequality (for the concave function log)

Example: Biosequence Analysis

- Recall the DNA Sequence logo example earlier.
- The height of each bar is defined to be $2 - \mathbb{H}$, where $\mathbb{H}$ is the entropy of that distribution, and $2\ (= \ln_2 4)$ is the maximum possible entropy.
- Thus a bar of height 0 corresponds to a uniform distribution $(\ln_2 4)$, whereas a bar of height 2 corresponds to a deterministic distribution.



seqlogoDemo from PMTK

Binary Variable

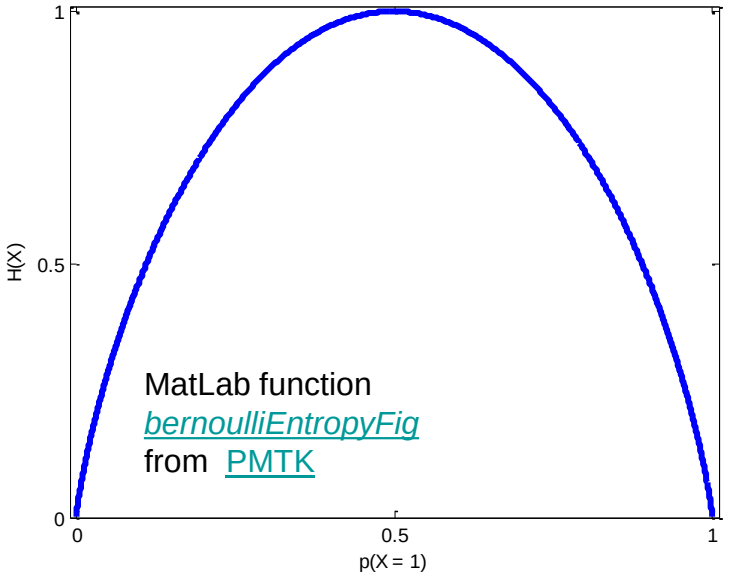
- Consider binary random variables, $X \in \{0, 1\}$, we can write $p(X = 1) = \theta$ and $p(X = 0) = 1 - \theta$.

$$X \in \{0,1\},\ p(X=1) = \theta,\ p(X=0) = 1-\theta$$

- Hence the entropy becomes (binary entropy function)

$$\mathbb{H}[X] = -\left[\theta \log_2 \theta + (1-\theta)\log_2 (1-\theta)\right]$$

- *The maximum value of 1 occurs when the distribution is uniform, $\theta = 0.5$.*



Differential Entropy

- Divide x into bins of width Δ. Assuming $p(x)$ is continuous, for each such bin, there must exist x_i such that

$$\int_{i\Delta}^{(i+1)\Delta} p(x)dx = p(x_i)\Delta = \textit{probability in falling in bin } \Delta$$

$$\mathbb{H}_\Delta = -\sum_i p(x_i)\Delta \ln\left(p(x_i)\Delta\right) = -\sum_i p(x_i)\Delta \ln\left(p(x_i)\right) - \ln\Delta$$

$$\lim_{\Delta\to 0}\left\{\sum_i p(x_i)\Delta \ln p(x_i)\right\} = -\int p(x)\ln p(x)dx \textit{ (can be negative)}$$

- *The $\ln\Delta$ term is omitted since it diverges as $\Delta\to 0$ (indicating that infinite bits are needed to describe a continuous variable)*

Differential Entropy

- For a density defined over multiple continuous variables, denoted collectively by the vector $\boldsymbol{x}$, the differential entropy is given by

$$\mathbb{H}[\boldsymbol{x}] = -\int p(\boldsymbol{x}) \ln p(\boldsymbol{x}) d\boldsymbol{x}$$

- *Differential (unlike the discrete) entropy can be negative*

- When doing variable transformation $\boldsymbol{y}(\boldsymbol{x})$, use $p(\boldsymbol{x})d\boldsymbol{x} = p(\boldsymbol{y})d\boldsymbol{y}$, e.g. if $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ then:

$$\mathbb{H}[\boldsymbol{x}] = -\int p(\boldsymbol{y}) \ln\left(p(\boldsymbol{y})|\boldsymbol{A}|\right) d\boldsymbol{y} = \mathbb{H}[\boldsymbol{y}] - \ln|\boldsymbol{A}| \Rightarrow \mathbb{H}[\boldsymbol{y}] = \mathbb{H}[\boldsymbol{x}] + \ln|\boldsymbol{A}|$$

Differential Entropy and the Gaussian Distribution

- The distribution that maximizes the differential entropy with constraints on the first two moments is a Gaussian:

$$\tilde{\mathbb{H}} = -\int p(x)\ln p(x)dx + \lambda_1\underbrace{\left(\int_{-\infty}^{+\infty} p(x)dx - 1\right)}_{\text{Normalization}} + \lambda_2\underbrace{\left(\int_{-\infty}^{+\infty} xp(x)dx - \mu\right)}_{\text{Given mean}} + \lambda_3\underbrace{\left(\int_{-\infty}^{+\infty} (x-\mu)^2 p(x)dx - \sigma^2\right)}_{\text{Given std}}$$

- Using calculus of variations ,

$$\delta\tilde{\mathbb{H}} = -\int \delta p(x)\ln p(x)dx - \int \delta p(x)dx + \lambda_1\int \delta p(x)dx + \lambda_2\int x\delta p(x)dx + \lambda_3\int (x-\mu)^2\delta p(x)dx = 0$$

$$p(x) = e^{-1+\lambda_1+\lambda_2 x+\lambda_3(x-\mu)^2} \quad \overset{\text{Use the constraints}}{\Rightarrow} \quad p(x) = \frac{1}{\left(2\pi\sigma^2\right)^{1/2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- Evaluating the differential entropy of the Gaussian, we obtain (an expression for a multivariate Gaussian is also given)

$$\mathbb{H}[x] = \frac{1}{2}\left(1+\ln\left(2\pi\sigma^2\right)\right) = \frac{1}{2}\ln\left((2\pi e)^d \det\Sigma\right),\ d=1,\ \det\Sigma = \sigma^2$$

Note $\mathbb{H}[x] < 0$ *for* $\sigma^2 < 1/(2\pi e)$

Kullback-Leibler Divergence and Cross Entropy

- Consider some unknown distribution $p(x)$, and suppose that we have modeled this using an approximating distribution $q(x)$.

- If we use $q(x)$ to construct a coding scheme for the purpose of transmitting values of x to a receiver, then the additional information to specify x is:

$$KL(p\,\|\,q) = -\underbrace{\int p(x)\ln q(x)dx}_{\text{I transmit q(x) but I average it with the exact probability p(x)}} - \left(-\int p(x)\ln p(x)dx\right) = -\int p(x)\ln\left\{\frac{q(x)}{p(x)}\right\}dx$$

- The *cross entropy* is defined as:

$$\mathbb{H}(p,q) = -\int p(x)\ln q(x)dx$$

KL Divergence and Cross Entropy

- The cross entropy $\mathbb{H}(p,q) = -\int p(x) \ln q(x) dx$ 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.
- $\mathbb{H}(p) = \mathbb{H}(p,p)$ is the expected # of bits using the true model.
- *The KL divergence is the average number of extra bits needed to encode the data, because we used distribution q to encode the data instead of the true distribution p.*
- The "extra number of bits" interpretation makes it clear that

$$KL(p \| q) = -\int p(x) \ln q(x) dx - \left(-\int p(x) \ln p(x) dx\right) = -\int p(x) \ln \left\{\frac{q(x)}{p(x)}\right\} dx$$

- The KL distance is not a symmetrical quantity, that is

$$KL(p \| q) \neq KL(q \| p)$$

KL Divergence Between Two Gaussians

- Consider $p(x) = \mathcal{N}(x|\mu, \sigma^2)$ and $q(x) = \mathcal{N}(x|m, s^2)$.

$$KL(p\|q) = \underbrace{-\int p(x)\ln q(x)dx}_{\int \mathcal{N}(x|\mu,\sigma^2)\frac{1}{2}\left(\ln\left(2\pi s^2\right)+\frac{(x-m)^2}{s^2}\right)dx} \underbrace{-\left(-\int p(x)\ln p(x)dx\right)}_{\frac{1}{2}\ln\left(2\pi e\sigma^2\right)}$$

- Note that the first term can be computed using the moments and normalization condition of a Gaussian and the second term from the differential entropy of a Gaussian.

- Finally we obtain:

$$KL(p\|q) = \frac{1}{2}\left(\ln\left(\frac{s^2}{\sigma^2}\right)+\frac{\sigma^2+\mu^2-2\mu m+m^2}{s^2}-1\right)$$

KL Divergence Between Two Gaussians

- Consider now $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $q(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{m}, \boldsymbol{L})$.

$$KL(p\|q) = \underbrace{-\int p(\boldsymbol{x}) \ln q(\boldsymbol{x}) d\boldsymbol{x}}_{\int \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) \frac{1}{2}\left(D\ln(2\pi)+\ln|\boldsymbol{L}|+(\boldsymbol{x}-\boldsymbol{m})^T \boldsymbol{L}^{-1}(\boldsymbol{x}-\boldsymbol{m})\right) d\boldsymbol{x}}$$

$$\frac{1}{2}\left(D\ln(2\pi)+\ln|\boldsymbol{L}|+Tr\left(\boldsymbol{L}^{-1}\left(\boldsymbol{\mu}\boldsymbol{\mu}^T+\boldsymbol{\Sigma}\right)\right)-\boldsymbol{\mu}^T\boldsymbol{L}^{-1}\boldsymbol{m}-\boldsymbol{m}^T\boldsymbol{L}^{-1}\boldsymbol{\mu}+\boldsymbol{m}^T\boldsymbol{L}^{-1}\boldsymbol{m}\right)$$

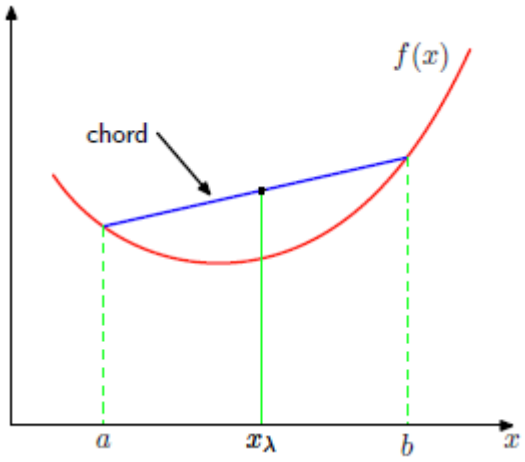
$$-\underbrace{\left(-\int p(\boldsymbol{x}) \ln p(\boldsymbol{x}) d\boldsymbol{x}\right)}_{\frac{1}{2}\ln|\boldsymbol{\Sigma}|+\frac{D}{2}(1+\ln(2\pi))}$$

$$= \frac{1}{2}\left(-\frac{D}{2}+\ln\frac{|\boldsymbol{L}|}{|\boldsymbol{\Sigma}|}+Tr\left(\boldsymbol{L}^{-1}\left(\boldsymbol{\mu}\boldsymbol{\mu}^T+\boldsymbol{\Sigma}\right)\right)-\boldsymbol{\mu}^T\boldsymbol{L}^{-1}\boldsymbol{m}-\boldsymbol{m}^T\boldsymbol{L}^{-1}\boldsymbol{\mu}+\boldsymbol{m}^T\boldsymbol{L}^{-1}\boldsymbol{m}\right)$$

Jensen's Inequality

- *For a convex function f*, Jensen's inequality gives (can be proven easily by induction)

$$f\left(\sum_{i=1}^{M}\lambda_i x_i\right)\le\sum_{i=1}^{M}\lambda_i f(x_i),\ \lambda_i\ge 0\ and\ \sum_i\lambda_i=1$$



- This is equivalent (assume $M=2$) to our requirement for convexity $f''(x)>0$.
  - *Assume $f''(x)>0$ (strict convexity) for any x.*

$$f(x)=f(x_0)+f'(x_0)(x-x_0)+\frac{1}{2}f''(x^*)(x-x_0)^2>f(x_0)+f'(x_0)(x-x_0)$$

$$For\ x=a,b:\ \left.\begin{array}{l}f(a)>f(x_0)+f'(x_0)(a-x_0)\\ f(b)>f(x_0)+f'(x_0)(b-x_0)\end{array}\right\}\Rightarrow\lambda f(a)+(1-\lambda)f(b)>f(x_0)+f'(x_0)(\underbrace{\lambda a+(1-\lambda)b}_{Set:\,x_0}-x_0)$$

Jensen's inequality is thus shown: $\lambda f(a)+(1-\lambda)f(b)>f\left(\lambda a+(1-\lambda)b\right)$

Jensen's Inequality

- Assume Jensen's inequality. We should show that $f''(x) > 0$ (strict convexity) for any x.
- Set the following: $a = b - 2\varepsilon, b = a + 2\varepsilon > a, \varepsilon > 0$. Using Jensen's inequality, we can easily derive the above equation as:

$$\frac{1}{2} f(a) + \frac{1}{2} f(b) > f\left(0.5a + 0.5b\right)$$

$$= \frac{1}{2} f\left(0.5(b - 2\varepsilon) + 0.5b\right) + \frac{1}{2} f\left(0.5a + 0.5(a + 2\varepsilon)\right)$$

$$= \frac{1}{2} f(b - \varepsilon) + \frac{1}{2} f(a + \varepsilon) \Rightarrow f(b) - f(b - \varepsilon) > f(a + \varepsilon) - f(a)$$

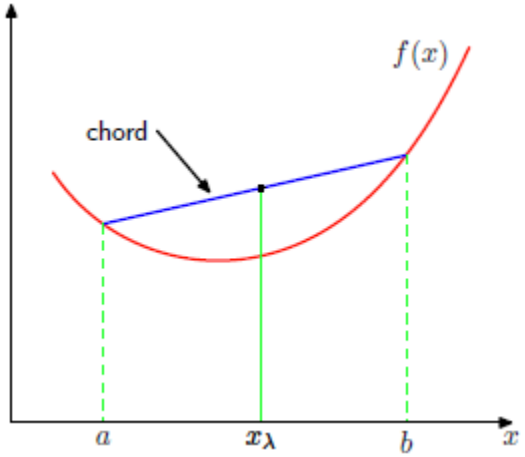
- For ε small, we thus have:

$$\frac{f(b) - f(b - \varepsilon)}{\varepsilon} > \frac{f(a + \varepsilon) - f(a)}{\varepsilon} \; or \; f'(b) > f'(a) \Rightarrow f(.) \; is \; convex$$

Jensen's Inequality

- Using Jensen's inequality $f\left(\sum_{i=1}^{M}\lambda_i x_i\right)\leq\sum_{i=1}^{M}\lambda_i f(x_i),\ \lambda_i\geq 0\ and\ \sum_i\lambda_i=1$ for a discrete random variable results in:

$$Set:\ \lambda_i=p_i\Rightarrow f\left(\mathbb{E}[x]\right)\leq\mathbb{E}[f(x)]$$



- We can generalize this result to continuous random variables:

$$(\ for\ continuous\ rv)\ f\left(\int xp(x)dx\right)\leq\int f(x)p(x)dx$$

- We will use this shortly in the context of the KL distance.
- *We often use Jensen's inequality for concave functions (e.g. $\log x$). In that case, be sure you reverse the inequality!*

Jensen's Inequality: Example

- As another example of Jensen's inequality, consider the arithmetic and geometric means of a set of real variables:

$$\bar{x}_A = \frac{1}{M}\sum_{i=1}^{M} x_i \,,\; \bar{x}_G = \left(\prod_{i=1}^{M} x_i\right)^{1/M}$$

- Using Jensen's inequality for $f(x) = \log(x)$ (concave), i.e.

$\mathbb{E}\left[\ln(x)\right] \leq \ln\left(\mathbb{E}\left[x\right]\right)$, we can show:

$$\ln\bar{x}_G = \frac{1}{M}\ln\left(\prod_{i=1}^{M} x_i\right) = \sum_{i=1}^{M}\frac{1}{M}\ln x_i \leq \ln\left(\sum_{i=1}^{M}\frac{1}{M}x_i\right) = \ln\bar{x}_A \Rightarrow \bar{x}_G \leq \bar{x}_A$$

The Kullback-Leibler Divergence

$$f\left(\mathbb{E}[x]\right) \leq \mathbb{E}[f(x)] \Rightarrow f\left(\int xp(x)dx\right) \leq \int f(x)p(x)dx$$

- Using Jensen's inequality, we can show *(−log is a convex function)* that:

$$KL(p\,||\,q) = -\int p(x)\ln\left\{\frac{q(x)}{p(x)}\right\}dx \geq -\ln\int p(x)\frac{q(x)}{p(x)}dx = -\ln\int q(x)dx = 0$$

- Thus we derive the following Information Inequality:

$$KL(p\,||\,q) \geq 0, \text{ with } KL(p\,||\,q) \geq 0 \text{ if and only if } p(x) = q(x)$$

Principle of Insufficient Reason

- An important consequence of the information inequality is that *the discrete distribution with the maximum entropy is the uniform distribution*.

- More precisely, $\mathbb{H}(X) \leq \log |\mathcal{X}|$, where $|\mathcal{X}|$ is the number of states for X, with equality iff $p(x)$ is uniform. To see this, let $u(x) = 1/|\mathcal{X}|$. Then

$$KL(p \| u) = -\sum_x p(x) \log u(x) + \sum_x p(x) \log p(x) = \log |\mathcal{X}| - \mathbb{H}(x) \geq 0$$

- This ***principle of insufficient reason***, argues in favor of using uniform distributions when there are no other reasons to favor one distribution over another.

The Kullback-Leibler Divergence

- Data compression is in some way related to density estimation.

- The Kullback-Leibler divergence is measuring the distance between two distributions and it is zero when the two densities are identical.

- Suppose the data is generated from an unknown $p(\boldsymbol{x})$ that we try to approximate with a parametric model $q(\boldsymbol{x}|\theta)$. Suppose we have observed training points $\boldsymbol{x}_n \sim p(\boldsymbol{x}), n = 1, \ldots, N$. Then:

$$KL(p \,\|\, q) = -\int p(x) \ln \left\{ \frac{q(x)}{p(x)} \right\} dx \approx \frac{1}{N} \sum_{n=1}^{N} \left\{ -\ln q(\boldsymbol{x}_n \,|\, \theta) + \ln p(\boldsymbol{x}_n) \right\}$$

Sample average approximation of the mean

The KL Divergence Vs. MLE

- Note that only the first term is a function of q.

- Thus minimizing $KL(p\|q)$ is equivalent to maximizing the likelihood function for θ under the distribution q.

$$KL(p\|q) = -\int p(\boldsymbol{x}) \ln \left\{ \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})} \right\} dx \approx \frac{1}{N} \sum_{n=1}^{N} \left\{ -\ln q(\boldsymbol{x}_n \mid \theta) + \ln p(\boldsymbol{x}_n) \right\}$$

- So the MLE estimate minimizes the KL divergence to the empirical distribution

$$p_{emp}(\boldsymbol{x}) = \frac{1}{N} \sum_{n=1}^{N} \delta_{\boldsymbol{x}_n}(\boldsymbol{x})$$

$$\arg\min_{q} KL(p_{emp}(\boldsymbol{x}) \| q) = -\int p_{emp}(\boldsymbol{x}) \ln \left\{ \frac{q(\boldsymbol{x})}{p_{emp}(\boldsymbol{x})} \right\} d\boldsymbol{x} = const - \frac{1}{N} \sum_{n=1}^{N} \ln q(\boldsymbol{x}_n \mid \theta)$$

Conditional Entropy

- For a joint distribution, *the conditional entropy* is

$$\mathbb{H}[y \mid x] = -\iint p(y,x) \ln p(y \mid x) dy dx$$

- This represents the average information to specify y if we already know the value of x
  - It is easily seen, using $p(y,x) = p(y \mid x) p(x)$, and substituting inside the log in $\mathbb{H}[x,y] = -\iint p(x,y) \ln p(x,y) dy dx$ that the conditional entropy satisfies the relation

$$\mathbb{H}[x,y] = \mathbb{H}[y \mid x] + \mathbb{H}[x]$$

where $\mathbb{H}[x,y]$ is the differential entropy of $p(x,y)$ and $\mathbb{H}[x]$ is the differential entropy of $p(x)$.

Conditional Entropy for Discrete Variables

- Consider *the conditional entropy* for discrete variables

$$\mathbb{H}[y \mid x] = -\sum_i \sum_j p(y_i, x_j) \ln p(y_i \mid x_j)$$

- To understand further the meaning of conditional entropy, *let us consider the implications of* $\mathbb{H}[y|x] = 0$.
  - We have:

$$\mathbb{H}[y \mid x] = \sum_i \sum_j \underbrace{\left(-p(y_i \mid x_j) \ln p(y_i \mid x_j)\right)}_{\geq 0} p(x_j) = 0$$

  - From this we can conclude that $For\ each\ x_j\ s.t.\ p(x_j) \neq 0$

$$the\ following\ must\ hold: p(y_i \mid x_j) \ln p(y_i \mid x_j) = 0$$

  - Since $plogp = 0 \leftrightarrow p = 0$ or $p = 1$ and since $p(y_i|x_j)$ is normalized, there is only one y_i s.t. $p(y_i \mid x_j) = 1$ with all other $p(. \mid x_j) = 0$. Thus y is a function of x.

Mutual Information

- If the variables are not independent, we can gain some idea of whether they are 'close' to being independent by considering the KL divergence between the joint distribution and the product of the marginals:

$$Mutual\ Information: \mathbb{I}[x,y] = KL\big(p(x,y) \,\|\, p(x)p(y)\big) =$$

$$= -\iint p(x,y) \ln \frac{p(x)p(y)}{p(x,y)} dxdy \geq 0$$

$$\mathbb{I}[x,y] = 0 \ iff \ \ x, y \ independent$$

- The mutual information is related to the conditional entropy through

$$\mathbb{I}[x,y] = -\iint p(x,y) \ln \frac{p(y)}{p(y \mid x)} dxdy = \mathbb{H}[y] - \mathbb{H}[y \mid x] \Rightarrow$$

$$\mathbb{I}[x,y] = \mathbb{H}[x] - \mathbb{H}[x \mid y] = \mathbb{H}[y] - \mathbb{H}[y \mid x]$$

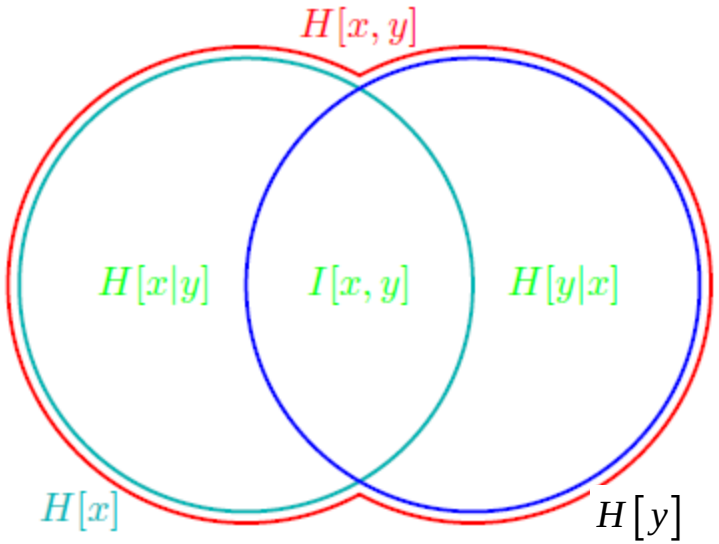
Mutual Information

- The mutual information represents the reduction in the uncertainty about x once we learn the value of y (and reversely).

$$\mathbb{I}[x,y] = \mathbb{H}[x] - \mathbb{H}[x \mid y] = \mathbb{H}[y] - \mathbb{H}[y \mid x]$$

$$\mathbb{H}[x] \geq \mathbb{H}[x \mid y]$$

$$\mathbb{H}[y] \geq \mathbb{H}[y \mid x]$$



- In a Bayesian setting, $p(x)$ =prior, $p(x|y)$ posterior, and $\mathbb{I}[x,y]$ represents the reduction in uncertainty in x once we observe y.

Note that $\mathbb{H}[x,y] \leq \mathbb{H}[x] + \mathbb{H}[y]$

- This is easy to prove noticing that

$$\mathbb{I}[x,y] = \mathbb{H}[y] - \mathbb{H}[y \mid x] \geq 0 \; (KL \; divergence)$$

and

$$\mathbb{H}[x,y] = \mathbb{H}[y \mid x] + \mathbb{H}[x]$$

from which

$$\mathbb{H}[x,y] = \mathbb{H}[x] + \mathbb{H}[y] - \mathbb{I}[x,y] \leq \mathbb{H}[x] + \mathbb{H}[y]$$

- *The equality here is true only if x, y are independent:*

$$\mathbb{H}[x,y] = -\iint p(x,y) \ln p(x,y) dy dx = -\iint p(x,y) \big(\ln p(x) + \ln p(y) \big) dy dx = \mathbb{H}[x] + \mathbb{H}[y]$$

(sufficiency condition)

$$\mathbb{H}[y \mid x] = \mathbb{H}[y] \Rightarrow \mathbb{I}[x,y] = 0 \Rightarrow p(x,y) = p(x)p(y)$$ (necessary condition)

Mutual Information for Correlated Gaussians

- Consider two correlated Gaussians as follows:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{matrix} X \\ Y \end{matrix} \middle| \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix} \right)$$

- For each of these variables we can write:

$$\mathbb{H}[X] = \mathbb{H}[Y] = \frac{1}{2}\ln\left(2\pi e\sigma^2\right)$$

- The joint entropy is also given similarly as

$$\mathbb{H}[X,Y] = \frac{1}{2}\ln\left((2\pi e)^2 \underbrace{\sigma^4(1-\rho^2)}_{\det\Sigma} \right)$$

- *Thus:* $\mathbb{I}[x,y] = \mathbb{H}[x] + \mathbb{H}[y] - \mathbb{H}[x,y] = \frac{1}{2}\log\frac{1}{1-\rho^2}$

- *Note:* $\rho = 0\ (independent\ X,Y) \Rightarrow \mathbb{I}[x,y] = 0$

$\rho = \pm 1\ (linear\ correlated\ X = \pm Y) \Rightarrow \mathbb{I}[x,y] = \infty$

Pointwise Mutual Information

- 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

$$PMI(x,y) =: -\log \frac{p(x)p(y)}{p(x,y)} = \log \frac{p(x|y)}{p(x)} = \log \frac{p(y|x)}{p(y)}$$

- This measures the discrepancy between these events occurring together compared to what would be expected by chance. *Clearly the* $MI, \mathbb{I}[x,y]$*, of* X *and* Y *is just the expected value of the* PMI*.*

- *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 $p(y|x)$.

Mutual Information

- For continuous random variables, it is common to first ***discretize*** *or* ***quantize*** *them into bins*, and computing how many values fall in each histogram bin (Scott 1979).

- The number of bins used, and the location of the bin boundaries, can have a significant effect on the results.

- One can estimate the MI directly, *without performing density estimation* (Learned-Miller, 2004). Another approach is to *try many different bin sizes and locations, and to compute the maximum MI achieved*.

- Scott, D. (1979). On optimal and data-based histograms, *Biometrika 66*(3), 605–610.
- Learned-Miller, E. (2004). Hyperspacings and the estimation of information theoretic quantities. Technical Report 04-104, U. Mass. Amherst Comp. Sci. Dept.
- Reshef, D., Y. Reshef, H. Finucane, S. Grossman, G. McVean, P. Turnbaugh, E. Lander, M. Mitzenmacher, and P. Sabeti (2011, December). Detecting novel associations n large data sets. *Science 334*, 1518–1524.
- Speed, T. (2011, December). A correlation for the 21st century. *Science 334*, 152–1503.

*Use MatLab function *miMixedDemo* from Kevin Murphys' PMTK

Maximal Information Coefficient

- This statistic appropriately normalized is known as the **maximal information coefficient** (MIC).
- We first define: $m(x,y) = \dfrac{\max_{G \in \mathcal{G}(x,y)} \mathbb{I}\big(X(G);Y(G)\big)}{\log \min(x,y)}$
- *Here* $\mathcal{G}(x,y)$ is the set of *2d grids of size* $x \times y$, and $X(G), Y(G)$ *represents a discretization of the variables onto this grid* (The maximization over bin locations is performed efficiently using *dynamic programming*)
- Now define the MIC as

$$MIC = \max_{x,y:\, xy<B} m(x,y)$$

- Reshef, D., Y. Reshef, H. Finucane, S. Grossman, G. McVean, P. Turnbaugh, E. Lander, M. Mitzenmacher, and P. Sabeti (2011, December). Detecting novel associations n large data sets. *Science 334*, 1518–1524.

Maximal Information Coefficient

- The MIC is defined as:

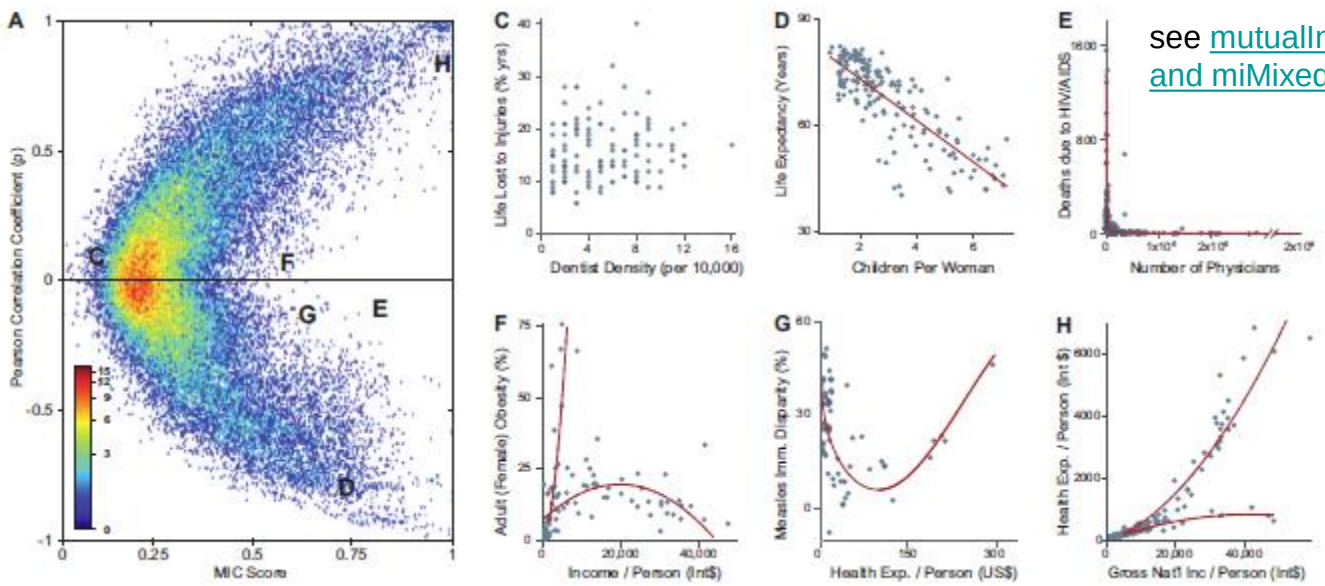
$$m(x,y) = \frac{\max_{G \in \mathcal{G}(x,y)} \mathbb{I}\left(X(G);Y(G)\right)}{\log \min(x,y)} \qquad MIC \equiv \max_{x,y:xy<B} m(x,y)$$

- B is some sample-size dependent bound on the number of bins we can use and still reliably estimate the distribution (Reshef et al. suggest $B \sim N^{0.6}$).

- MIC lies in the range $[0,1]$, where 0 represents no relationship between the variables, and 1 represents a noise-free relationship of any form, not just linear.

- Reshef, D., Y. Reshef, H. Finucane, S. Grossman, G. McVean, P. Turnbaugh, E. Lander, M. Mitzenmacher, and P. Sabeti (2011, December). Detecting novel associations n large data sets. *Science 334*, 1518–1524.

Correlation Coefficient Vs MIC

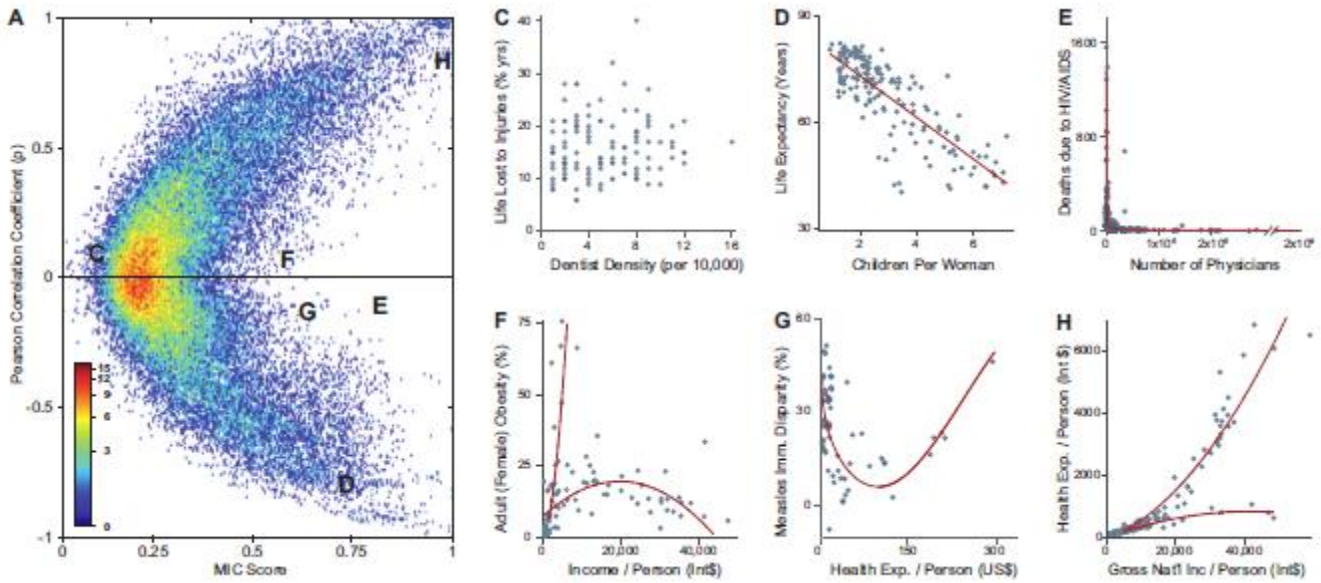


see mutualInfoAllPairsMixed for and miMixedDemo from PMTK3

- Reshef, D., Y. Reshef, H. Finucane, S. Grossman, G. McVean, P. Turnbaugh, E. Lander, M. Mitzenmacher, and P. Sabeti (2011, December). Detecting novel associations n large data sets. *Science 334*, 1518–1524.

- The data consists of 357 variables measuring a variety of social, economic, etc. indicators, collected by WHO.
- On the left, we see the *correlation coefficient (CC) plotted against the MIC for all 63,566 variable pairs.*
- On the right, we see scatter plots for particular pairs of variables.

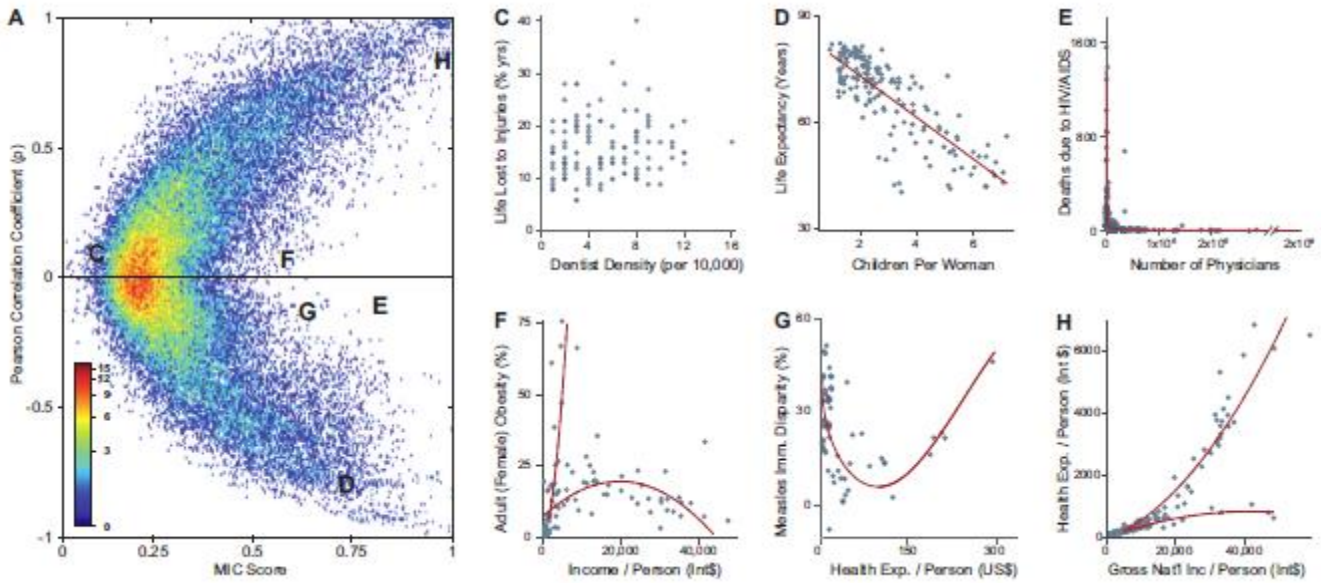
Correlation Coefficient Vs MIC



- Point marked C has a *low CC and a low MIC*. From the corresponding scatter we see that there is *no relationship between these two variables.*

- The points marked D and H have high CC (in absolute value) and high MIC and we see from the scatter plot that they represent nearly linear relationships.

Correlation Coefficient Vs MIC



- The points E, F, and G have low CC but high MIC. They correspond to non-linear (and sometimes, as in E and F, one-to-many) relationships between the variables.

- Statistics (such as MIC) based on mutual information can be used to discover interesting relationships between variables in a way that correlation coefficients cannot.