### **Information Theory**

Sargur N. Srihari

### **Topics**

- 1. Entropy as an Information Measure
  - Discrete variable definition
     Relationship to Code Length
  - 2. Continuous Variable
    Differential Entropy
- 2. Maximum Entropy
- 3. Conditional Entropy
- 4. Kullback-Leibler Divergence (Relative Entropy)
- 5. Mutual Information

### Information Measure

- How much information is received when we observe a specific value for a discrete random variable *x*?
- Amount of information is degree of surprise
  - Certain means no information
  - More information when event is unlikely
- Depends on probability distribution p(x), a quantity h(x)
- If there are two unrelated events x and y we want h(x,y) = h(x) + h(y)
- Thus we choose  $h(x) = -\log_2 p(x)$ 
  - Negative assures that information measure is positive
- Average amount of information transmitted is the expectation wrt p(x) refered to as entropy

$$H(x) = -\sum_{x} p(x) \log_2 p(x)$$

### Usefulness of Entropy

- Uniform Distribution
  - Random variable x has 8 possible states, each equally likely
    - We would need 3 bits to transmit
    - Also,  $H(x) = -8 \times (1/8) \log_2(1/8) = 3 \text{ bits}$
- Non-uniform Distribution
  - If x has 8 states with probabilities
     (1/2,1/4,1/8,1/16,1/64,1/64,1/64,1/64)
     H(x)=2 bits
- Non-uniform distribution has smaller entropy than uniform
- Has an interpretation of in terms of disorder

### Relationship of Entropy to Code Length

 Take advantage of non-uniform distribution to use shorter codes for more probable events

- If x has 8 states (a,b,c,d,e,f,g,h) with probabilities 1/2 (1/2,1/4,1/8,1/16,1/64,1/64,1/64,1/64) 1/4 Can use codes 0,10,110,111100,111110, 1111111 1/8 average code length = <math>(1/2)1+(1/4)2+(1/8)3+(1/16)4+4(1/64)6=2 bits
- Same as entropy of the random variable
- Shorter code string is not possible due to need to disambiguate string into component parts
- 11001110 is uniquely decoded as sequence cad

## Relationship between Entropy and Shortest Coding Length

- Noiseless coding theorem of Shannon
  - Entropy is a lower bound on number of bits needed to transmit a random variable
- Natural logarithms are used in relationship to other topics
  - Nats instead of bits

# History of Entropy: thermodynamics to information theory

- Entropy is average amount of information needed to specify state of a random variable
- Concept had much earlier origin in physics
  - Context of equilibrium thermodynamics
  - Later given deeper interpretation as measure of disorder (developments in statistical mechanics)

### **Entropy Persona**

#### World of Atoms



World of Bits



- Ludwig Eduard Boltzmann (1844-1906)
  - Created Statistical Mechanics
    - First law: conservation of energy
      - Energy not destroyed but converted from one form to other
    - Second law: principle of decay in nature
       – entropy increases
      - Explains why not all energy is available to do useful work
  - Relate macro state to statistical behavior of microstate
- Claude Shannon (1916-2001)
- Stephen Hawking (Gravitational Entropy)

Srihari

### Physics view of Entropy

• N objects into bins so that  $n_i$  are in  $i^{th}$  bin where

$$\sum_{i} n_{i} = N$$

- No of different ways of allocating objects to bins
  - N ways to choose first, N-1 ways for second leads to N.(N-1) .. 2.1 = N!
  - We don't distinguish between rearrangements within each bin
    - In  $i^{th}$  bin there are  $n_i$ ! ways of reordering objects
  - Total no of ways of allocating N objects to bins is  $W = \frac{IV!}{\prod n_i!}$ Called Multiplicity (also weight of macrostate)
- Entropy: scaled log of multiplicity  $H = \frac{1}{N} \ln W = \frac{1}{N} \ln N! \frac{1}{N} \sum_{i} \ln n_{i}!$

$$H = \frac{1}{N} \ln W = \frac{1}{N} \ln N! - \frac{1}{N} \sum_{i} \ln n_{i}!$$

- Sterlings approx as N → ∞ ln N!≈ NlnN N

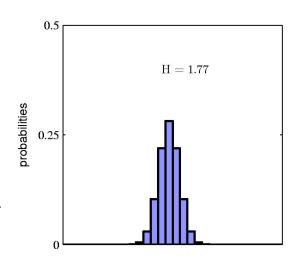
- Which gives
$$H = -\lim_{N \to \infty} \sum_{i} \left(\frac{n_{i}}{N}\right) \ln\left(\frac{n_{i}}{N}\right) = -\sum_{i} p_{i} \ln p_{i}$$

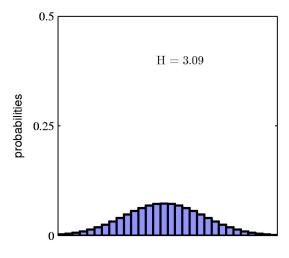
- Overall distribution, as ratios  $n_i/N$ , called *macrostate*
- Specific arrangement of objects in bin is *microstate*

### **Entropy and Histograms**

- If X can take one of M values (bins, states) and  $p(X=x_i)=p_i$  then  $H(p)=-\sum_i p_i \ln p_i$
- Minimum value of entropy is  $\theta$  when one of the  $p_i$ =1 and other  $p_i$  are  $\theta$ 
  - noting that  $\lim_{p\to 0} p \ln p = 0$
- Sharply peaked distribution has low entropy
- Distribution spread more evenly will have higher entropy

30 bins, higher value for broader distribution





### Maximum Entropy Configuration

- Found by maximizing H using Lagrange multiplier to enforce constraint of probabilities
- Maximize  $\tilde{H} = -\sum_{i} p(x_i) \ln p(x_i) + \lambda \left( \sum_{i} p(x_i) 1 \right)$  Solution: all  $p(x_i)$  are equal or  $p(x_i) = 1/M$  M = no of states
- Maximum value of entropy is ln M
- To verify it is a maximum, evaluate second derivative of entropy  $\frac{\partial \tilde{H}}{\partial p(x_{+})\partial p(x_{+})} = -I_{ij}\frac{1}{n_{-}}$ 
  - where  $I_{ii}$  are elements of identity matrix

### Entropy with Continuous Variable

- Divide x into bins of width ∆
- For each bin there must exist a value  $x_i$  such that  $\int_{i\Delta}^{(i+1)\Delta} p(x)d(x) = p(x_i)\Delta$
- Gives a discrete distribution with probabilities p(x<sub>i</sub>)Δ
- Entropy  $H_{\Delta} = -\sum p(x_i) \Delta \ln(p(x_i) \Delta) = -\sum p(x_i) \Delta \ln p(x_i) \ln \Delta$
- Omit the second term and consider the limit △→0

$$H_{\Delta} = -\int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x}$$

- Known as Differential Entropy
- Discrete and Continuous forms of entropy differ by quantity  $\ln \Delta$  which diverges
  - Reflects to specify continuous variable very precisely requires a large no of bits

### Entropy with Multiple Continuous Variables

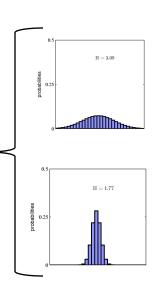
Differential Entropy for multiple continuous variables

$$H(x) = -\int p(x) \ln p(x) dx$$

- For what distribution is differential entropy maximized?
  - For discrete distribution, it is uniform
  - For continuous, it is Gaussian
    - as we shall see

### Entropy as a Functional

- Ordinary calculus deals with functions
- A functional is an operator that takes a function as input and returns a scalar
- A widely used functional in machine learning is entropy H[p(x)] which is a scalar quantity
- We are interested in the maxima and minima of functionals analogous to those for functions
  - Called calculus of variations



### Maximizing a Functional

- Functional: mapping from set of functions to real value
- For what function is it maximized?
- Finding shortest curve length between two points on a sphere (geodesic)
  - With no constraints it is a straight line
  - When constrained to lie on a surface solution is less obvious— may be several
- Constraints incorporated using Lagrangian

### Maximizing Differential Entropy

 Assuming constraints on first and second moments of p(x) as well as normalization

$$\int p(x)dx = 1 \qquad \int xp(x)dx = \mu \qquad \int (x-\mu)^2 p(x)dx = \sigma^2$$

• Constrained maximization is performed using Lagrangian multipliers. Maximize following functional wrt p(x):  $-\int p(x) \ln p(x) dx + \lambda_1 (\int p(x) dx - 1)$ 

+ 
$$\lambda_2 \left( \int x p(x) dx - \mu \right) + \lambda_3 \left( \int (x - \mu)^2 p(x) dx - \sigma^2 \right)$$

- Using the calculus of variations derivative of functional is set to zero giving  $p(x) = \exp\{-1 + \lambda_1 + \lambda_2 x + \lambda_3 (x \mu)^2\}$
- Backsubstituting into three constraint equations leads to the result that distribution that maximizes differential entropy is Gaussian!

### Differential Entropy of Gaussian

Distribution that maximizes Differential Entropy is Gaussian

$$p(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{\frac{-(x-\mu)^2}{\sigma^2}\right\}$$

Value of maximum entropy is

$$H(x) = \frac{1}{2} \left\{ 1 + \ln(2\pi\sigma^2) \right\}$$

- Entropy increases as variance increases
- Differential entropy, unlike discrete entropy, can be negative for  $\sigma^2 < 1/2\pi e$

### **Conditional Entropy**

- If we have joint distribution p(x,y)
  - We draw pairs of values of x and y
  - If value of x is already known, additional information to specify corresponding value of y is -ln p(y|x)
- Average additional information needed to specify y is the conditional entropy

$$H[y | x] = -\iint p(y | x) \ln p(y | x) dy dx$$

- By product rule H[x,y] = H[y|x] + H[x]
  - where H[x,y] is differential entropy of p(x,y)
  - H[x] is differential entropy of p(x)
  - Information needed to describe x and y is given by information needed to describe x plus additional information needed to specify y given x

### Relative Entropy

- If we have modeled unknown distribution p(x) by approximating distribution q(x)
  - i.e., q(x) is used to construct a coding scheme of transmitting values of x to a receiver
  - Average additional amount of information required to specify value of x as a result of using q(x) instead of true distribution p(x) is given by relative entropy or K-L divergence
- Important concept in Bayesian analysis
  - Entropy comes from Information Theory
  - K-L Divergence, or relative entropy, comes from Pattern Recognition, since it is a distance (dissimilarity) measure

### Relative Entropy or K-L Divergence

 Additional information required as a result of using q(x) in place of p(x)

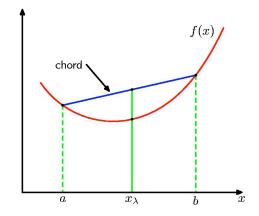
$$KL(p || q) = -\int p(x) \ln q(x) dx - \left(\int p(x) \ln p(x) dx\right)$$
$$= -\int p(x) \ln \left\{\frac{p(x)}{q(x)}\right\} dx$$

- Not a symmetrical quantity:  $KL(p||q) \neq KL(q||p)$
- K-L divergence satisfies KL(p||q)>0 with equality iff p(x)=q(x)
  - Proof involves convex functions

### **Convex Function**

- A function f(x) is convex if every chord lies on or above function
  - Any value of x in interval from x=a to x=b can be written as  $\lambda a + (1-\lambda)b$  where  $0 \le \lambda \le 1$
  - Corresponding point on chord is  $\lambda f(a) + (1-\lambda)f(b)$
  - Convexity implies

$$f(\lambda a + (1-\lambda)b) \le \lambda f(a) + (1-\lambda)f(b)$$
  
Point on curve  $\le$  Point on chord



By induction, we get Jensen's inequality

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

### Proof of positivity of K-L Divergence

Applying to KL definition yields desired result

$$KL(p \parallel q) = -\int p(x) \ln \left\{ \frac{p(x)}{q(x)} \right\} dx \ge -\ln \int q(x) dx = 0$$

$$\ge 0$$

### **Mutual Information**

- Given joint distribution of two sets of variables
   p(x,y)
  - If independent, will factorize as p(x,y)=p(x)p(y)
  - If not independent, whether close to independent is given by
    - KL divergence between joint and product of marginals

$$I[x, y] = KL(p(x, y) || p(x)p(y))$$

$$= \iint p(x, y) \ln \left( \frac{p(x)p(y)}{p(x, y)} \right) dx dy$$

Called Mutual Information between variables x and y

### **Mutual Information**

Using Sum and Product Rules

$$I[x,y] = H[x] - H[x|y] = H[y] - H[y|x]$$

- Mutual Information is reduction in uncertainty about x given value of y (or vice versa)
- Bayesian perspective:
  - if p(x) is prior and p(x|y) is posterior, mutual information is reduction in uncertainty after y is observed