

# Bayes Networks and Probabilistic Programming

Adam Massmann

Water Center NN Meetings

Week 5: Oct. 17th, 2017

# Introduction - Information Theory (see section 1.6 in Bishop)

- ▶ Consider the amount of information gained/learned by an event or observation (we'll call  $x$ ). We would receive more new information from a very surprising event than an event we expect (because we already know something about the expected event).
- ▶ So say we want to quantify this “amount of information” contained in an event. This should then be a function of the probability of the event ( $p(x)$ ). So the amount of information we'll call a function  $h(\cdot)$ , which will be a function of  $p(x)$ .

## Guidance for the functional form of $h(\cdot)$

- ▶ If two events  $x$  and  $y$  are independent, then the amount of information gained by both events should be  $h(x, y) = h(x) + h(y)$ .
- ▶ We also know that the joint probability of  $x$  and  $y$ 's occurrence would be:  $p(x, y) = p(x) p(y)$ .
- ▶ So the question is, what function  $\hat{h}$  satisfies:  
 $h(x, y) = \hat{h}(p(x) p(y)) = \hat{h}(p(x)) + \hat{h}(p(y))$ ?

# Information Entropy

- ▶  $\hat{h}(\cdot) = \log(\cdot)$  satisfies  
 $h(x, y) = \hat{h}(p(x) p(y)) = \hat{h}(p(x)) + \hat{h}(p(y))$ , so  
 $h(\cdot) = \log(p(\cdot))$ .
- ▶ It's desirable for  $h$  to be positive, so because  $0 \leq p \leq 1$ , let's make it  $h(\cdot) = -\log p(\cdot)$ .
- ▶ Now say we have a bunch of random variables  $x$  for which we want to know the average amount of information (i.e. expectation of  $h(x)$ ). This would be given by:

$$H[x] = - \sum_x p(x) \log p(x)$$

- ▶ This is known as the *entropy* of  $x$ .
- ▶ Extending this to continuous variables gives the *differential entropy*:

$$H[x] = - \int p(x) \log p(x) dx$$

[Bishop 2006, Shannon 1948]

# So what does information entropy look like?

- From thermodynamics and statistical mechanics we have some idea of entropy as a measure of the disorder or randomness in a system. For information theory it is similar.<sup>1</sup>

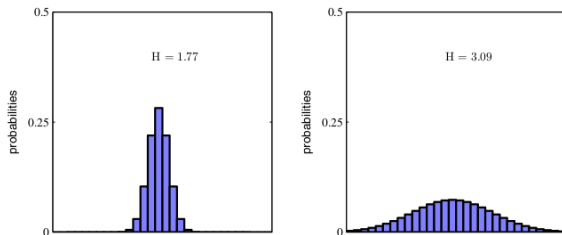


Figure 1: From Bishop 2006: Histograms of two probability distributions over thirty bins illustrating the higher value of entropy  $H$  for the broader distribution. The largest entropy would arise from a uniform distribution that would give  $H = -\ln 1/30 = 3.40$



---

<sup>1</sup>von Neumann told Shannon he should also call it entropy because “nobody knows what entropy really is, so in any discussion you will always have an advantage.”

# Kullback-Leibler divergence

- ▶ Why should we even care about information theory or entropy?
  - ▶ Because we can use entropy ideas to perform inference on a probabilistic model (e.g. a Bayes network), given data.

# References

-  Bishop, Christopher M (2006). *Pattern recognition and machine learning*. springer.
-  Shannon, Claude E (1948). “A mathematical theory of communication, Part I, Part II”. In: *Bell Syst. Tech. J.* 27, pp. 623–656.