NEU 560: Statistical Modeling and Analysis of Neural Data **Spring 2018**

Lecture 8: Information Theory and Maximum Entropy

Lecturer: Mike Morais *Scribes:*

8.1 Fundamentals of Information theory

Information theory started with Claude Shannon's *A mathematical theory of communication.* The first building block was **entropy**, which he sought as a functional $H(\cdot)$ of probability densities with two desired properties:

1. Decreasing in $P(X)$, such that if $P(X_1) < P(X_2)$, then $h(P(X_1)) > h(P(X_2))$.
2. Independent variables add, such that if X and Y are independent, then $H(P(X,Y)) = H(P(X)) + H(P(Y))$.

These are only satisfied for $-\log(\cdot)$. Think of it as a "surprise" function.

Definition 8.1 (Entropy) *The entropy of a random variable is the amount of information needed to fully describe it; alternate interpretations: average number of yes/no questions needed to identify X, how uncertain you are about X?*

$$H(X) = -\sum_X P(X)\log P(X) = -\mathbb{E}_X[\log P(X)] \tag{8.1}$$

Average information, surprise, or uncertainty are all somewhat parsimonious plain English analogies for entropy. There are a few ways to measure entropy for multiple variables; we'll use two, X and Y.

Definition 8.2 (Conditional entropy) *The conditional entropy of a random variable is the entropy of one random variable conditioned on knowledge of another random variable, on average.*

Alternative interpretations: the average number of yes/no questions needed to identify X given knowledge of Y, on average; or How uncertain you are about X if you know Y, on average?

$$\begin{aligned} H(X \mid Y) &= \sum_Y P(Y)[H(P(X \mid Y))] = \sum_Y P(Y)\Big[-\sum_X P(X \mid Y)\log P(X \mid Y)\Big] \\ &= = -\sum_{X,Y} P(X,Y)\log P(X \mid Y) \\ &= -\mathbb{E}_{X,Y}[\log P(X \mid Y)] \end{aligned} \tag{8.2}$$

Definition 8.3 (Joint entropy)

$$H(X,Y) = -\sum_{X,Y} P(X,Y)\log P(X,Y) = -\mathbb{E}_{X,Y}[\log P(X,Y)] \tag{8.3}$$

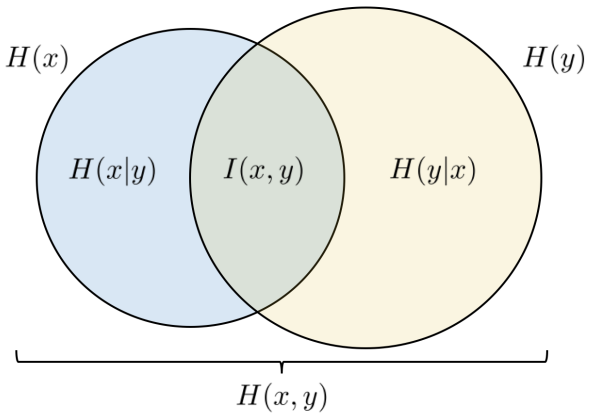
- **Bayes' rule for entropy**

$$H(X_1 \mid X_2) = H(X_2 \mid X_1) + H(X_1) - H(X_2) \tag{8.4}$$

- **Chain rule of entropies**

$$H(X_n, X_{n-1}, ...X_1) = \sum_{i=1}^{n} H(X_n \mid X_{n-1}, ...X_1) \tag{8.5}$$

It can be useful to think about these interrelated concepts with a so-called information diagram. These aid intuition, but are somewhat of a disservice to the mathematics behind them. Think of the area of each circle as the information needed to describe it, and any overlap would imply the "same information" (sorry.) describes both processes.



The entropy of X is the entire blue circle. Knowledge of Y removes the green slice. The joint entropy is the union of both circles. How do we describe their intersection, the green slice?

Definition 8.4 (Mutual information) *The mutual information between two random variables is the "amount of information" describing one random variable obtained through the other (mutual dependence); alternate interpretations: how much is your uncertainty about X reduced from knowing Y, how much does X inform Y?*

$$\begin{aligned} I(X,Y) &= \sum_{X,Y} P(X,Y) \log \frac{P(X,Y)}{P(X)P(Y)} \\ &= H(X) - H(X \mid Y) \\ &= H(Y) - H(Y \mid X) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned} \tag{8.6}$$

Note that $I(X,Y) = I(Y,X) \geq 0$, with equality if and only if X and Y are independent.

8.1.1 KL Divergence

From Bayes' rule, we can rewrite the joint distribution $P(X,Y) = P(X \mid Y)P(Y)$ and rewrite the mutual information as

$$I(X,Y) = \sum_Y P(Y) \sum_X P(X \mid Y) \log \frac{P(X \mid Y)}{P(X)} = \mathbb{E}_Y \Big[D_{KL}\big(P(X \mid Y) \,\|\, P(X)\big) \Big] \tag{8.7}$$

which we introduce as the Kullback-Leibler, or KL, divergence from $P(X)$ to $P(X \mid Y)$. Definition first, then intuition.

Definition 8.5 (Relative entropy, KL divergence) *The KL divergence $D_{KL}(p\|q)$ from q to p, or the relative entropy of p with respect to q, is the information lost when approximating p with q, or conversely the information gained when updating q with p.*

In terms of p and q: (8.8)

$$D_{KL}\big(p(X) \,\|\, q(X)\big) = \sum_X p(X) \log \frac{p(X)}{q(X)}$$

In terms of a prior and a posterior:

$$D_{KL}\big(p(X \mid Y) \,\|\, p(X)\big) = \sum_X p(X \mid Y) \log \frac{p(X \mid Y)}{p(X)}$$

We can think of it as the amount of extra information needed to describe $p(X \mid Y)$ (the posterior) if we used $p(X)$ (the prior) instead. Conversely, in Bayesian syntax, we can think of it as the **information gain** when updating belief from a prior $p(X)$ to the posterior $p(X \mid Y)$; it is the information gained about X by observing Y.

Claim 8.6 *Maximizing log-likelihood of observing data X with respect to model parameters θ is equivalent to minimizing KL divergence between the likelihood and the true source distribution of the data.*

Proof: The KL divergence from $p_{\text{true}}(X)$, the true source of the data (unknown), to $p(X \mid \theta)$, the model likelihood fit to the data, is given by

$$\begin{aligned} D_{KL}\big(p_{\text{true}}(X) \,\|\, p(X \mid \theta)\big) &= \sum_X p_{\text{true}}(X) \log \frac{p_{\text{true}}(X)}{p(X \mid \theta)} \\ &= -\sum_X p_{\text{true}}(X) \log p(X \mid \theta) + \sum_X p_{\text{true}}(X) \log p_{\text{true}}(X) \\ &= \left(\lim_{N \to \infty} -\frac{1}{N} \sum_{i=1}^{N} \log p(x_i \mid \theta) \right) + H\big[\, p_{\text{true}}(X)\big] \end{aligned} \tag{8.9}$$

For an observed dataset $\{x_1, x_2, ..., x_N\}$, we approximate the first sum with a Monte Carlo integral that is equal in the infinite limit. Other names for these sums are the cross entropy and the log-likelihood (you'll see this ML/cross-entropy equivalence leveraged when optimizing parameters in deep learning). Since the entropy of the data source is fixed with respect to our model parameters, it follows that

$$\arg\min_\theta D_{KL}\big(p_{\text{true}}(X) \,\|\, p(X \mid \theta)\big) = \arg\max_\theta \left(\lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} \log p(x_i \mid \theta) \right) \triangleq \hat{\theta}_{ML} \tag{8.10}$$

■

8.1.2 Data processing inequality

Garbage in, garbage out.
Suppose three random variables form a Markov chain $X \to Y_1 \to Y_2$; this is a sequence of events where each depends only on the former, such that

$$p(X, Y_1, Y_2) = p(Y_2 \mid Y_1)p(Y_1 \mid X)p(X) \tag{8.11}$$

The data processing inequality tells us that processing (e.g. from Y_1 to Y_2) cannot possibly increase information, so

$$I(X, Y_1) \geq I(X, Y_2) \tag{8.12}$$

8.2 Principle of maximum entropy

Entropy underlies a core theory for selecting probability distributions. Thomas Jaynes argues that the maxent distribution is "uniquely determined as the one which is maximally noncommittal with regard to missing information, in that it agrees with what is known, but expresses maximum uncertainty with respect to all other matters". Therefore, this is the most principled choice.

Many common probability distributions naturally arise as maximum entropy distributions under moment constraints. The basic problem looks like this:

$$\begin{aligned} \underset{p(X)}{\text{maximize}} \;\; -\sum_X p(X)\log p(X) \;\; \text{subject to} \;\; \sum_X p(X)f_i(X) = c_i \;\; \text{for all constraints } f_i \\ \text{with solution: } p(X) = \exp\Big(-1 + \lambda_0 + \sum_i \lambda_i f_i(X)\Big) \end{aligned} \tag{8.13}$$

One constraint is always $f_0(X) = 1$ and $c_0 = 1$; that is, we constrain that it must be a proper probability distribution and integrate (sum) to 1.

8.2.1 Optimization with Langrange multipliers

We solve the constrained optimization problem by forming a Langrangian and introducing Lagrange multipliers λ_i (recall when we derived PCA!).

$$\mathcal{L}\big(p(X), \lambda_0, \{\lambda_i\}\big) = -\sum_X p(X)\log p(X) + \lambda_0\left(\sum_X p(X) - 1\right) + \sum_i \lambda_i\left(\sum_X p(X)f_i(X) - c_i\right) \tag{8.14}$$

A solution, if it exists, will do so at a critical point of this Lagrangian, i.e. when it's gradient $\nabla\mathcal{L}\big(p(X), \lambda_0, \{\lambda_i\}\big) \equiv 0$. Recall that the gradient is the vector of all partial derivatives of $\mathcal{L}$ with respect to $p(X)$ *and* all of the Lagrange multipliers, identically zero when each partial derivative is zero. So

$$\begin{aligned} \frac{\partial\mathcal{L}}{\partial p(X)} &= 0 = -\log p(X) - 1 + \lambda_0 + \sum_i \lambda_i f_i(X) \\ \frac{\partial\mathcal{L}}{\partial \lambda_0} &= 0 = \sum_X p(X) - 1 \\ \frac{\partial\mathcal{L}}{\partial \lambda_i} &= 0 = \sum_X p(X)f_i(X) - c_i \end{aligned} \tag{8.15}$$

For λ_0 and λ_i such that the constraints are satisfied (*note*: the methods for doing so are unfortunately very problem-specific and algebraically laborious), we can solve the first equation for $p(X)$ and retrieve the solution from above:

$$p(X) = \exp\Big(-1 + \lambda_0 + \sum_i \lambda_i f_i(X)\Big) \tag{8.16}$$

Intuition for Langrange multipliers: Lagrange multipliers are additional variables we introduce to make our optimization well-defined. Geometrically, they represent the rate of change of the maximum entropy with respect to the constraint constants c_i, e.g. the marginal effect of the constant upon the optimal attainable value of entropy. Note that

$$\frac{\partial \mathcal{L}}{\partial c_i} = \lambda_i \tag{8.17}$$

While the maximum we care about is the maximum *entropy*, we can show that, at that maximum, the maximum of the Langrangian is equal to the maximum of the entropy:

$$\mathcal{L}\big(p^*(X), \lambda_0^*, \{\lambda_i^*\}\big) = -\sum_X p^*(X)\log p^*(X) + \lambda_0(0) + \sum_i \lambda_i(0) = H\big[p^*(X)\big] \tag{8.18}$$

You could instead count equations and unknowns. Even if we only have the constraint where $\sum_X p(X) = 1$, that's still two equations and only one unknown: $p(X)$. That's overdetermined. If we introduce λ_0, now we have two equations and two unknowns. We can solve that. Indeed, that rate of change of the constraints has an optimum *in addition to* the maxent distribution we seek.

8.2.2 Examples

- **What is the maximum entropy distribution over a finite (discrete) range $\{0, 1, ..., N\}$?** This is the unconstrained problem; the Lagrangian is simply

$$\begin{aligned}
\mathcal{L}\big(p(X), \lambda_0\big) &= -\sum_X p(X)\log p(X) + \lambda_0\left(\sum_X p(X) - 1\right) \\
\text{and} \quad \frac{\partial \mathcal{L}}{\partial p(X)} &= 0 = -\log p(X) - 1 + \lambda_0 \\
\frac{\partial \mathcal{L}}{\partial \lambda_0} &= 0 = \sum_X p(X) - 1 \\
\implies p(X) &= \exp\big(-1 + \lambda_0\big) \text{ with } \sum_X p(X) = 1 \\
\implies p(X) &= \frac{1}{N+1}
\end{aligned} \tag{8.19}$$

  This is the uniform distribution. In Bayesian estimation, this means that we select a uniform prior when we know nothing about the source of our data, yielding maximum likelihood estimation. Recall that the MLE is unbiased; maxent echoes this property.

- **What is the maximum entropy distribution for a continuous random variable X with**

mean μ? This problem has one constraint on its mean; the Lagrangian now becomes

$$\begin{aligned}
\mathcal{L}\big(p(X), \lambda_0, \lambda_1\big) &= -\int_X dX p(X) \log p(X) + \lambda_0 \left(\int_X dX p(X) - 1 \right) + \lambda_1 \left(\int_X dX X \cdot p(X) - \mu \right) \\
\text{and} \quad \frac{\partial \mathcal{L}}{\partial p(X)} &= 0 = -\log p(X) - 1 + \lambda_0 + \lambda_1 X \\
\frac{\partial \mathcal{L}}{\partial \lambda_0} &= 0 = \int_X dX p(X) - 1 \\
\frac{\partial \mathcal{L}}{\partial \lambda_1} &= 0 = \int_X dX X \cdot p(X) - \mu \\
\implies p(X) &= \exp\big(-1 + \lambda_0 + \lambda_1 X\big) \text{ with } \int_X dX p(X) = 1 \\
\implies p(X) &= \frac{1}{\mu} \exp\Big(-\frac{X}{\mu}\Big) \text{ only if } X \in [0, \infty)
\end{aligned} \tag{8.20}$$

This is the exponential distribution. The integral isn't defined on the full real line, so we note that it must be defined on the positive reals.

- **What is the maximum entropy distribution with a variance σ^2?** This problem has a variance constraint, and implicitly a mean constraint which we'll call μ, such that

$$\begin{aligned}
\mathcal{L}\big(p(X), \lambda_0, \lambda_1\big) &= -\int_X dX p(X) \log p(X) + \lambda_0 \left(\int_X dX p(X) - 1 \right) + \lambda_1 \left(\int_X dX (X-\mu)^2 \cdot p(X) - \sigma^2 \right) \\
\implies p(X) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2}(X-\mu)^2\Big)
\end{aligned} \tag{8.21}$$

abbreviated for brevity. The normal distribution is therefore the maximum entropy distribution for a distribution with known mean and variance. Yet another reason that Gaussians are so ubiquitous.

8.3 Barlow's efficient coding hypothesis: Infomax

See slides.

8.4 Entropy and reundancy in the English language

A bonus example! Try to read these quotations:

```
``Th_ onl_ wa_ to ge_ ri_ of a tempta____ is to yie__ to it.  Resi__ it, an_ you_
soul gro__ sic_ wi__ longi__ fo_ th_ thin__ it ha_ forbi____ to itse__.''
```

```
``Y cn prbbly gss wht ths sntnc sys, vn wth ll f th vwls mssng.  Tht ndcts tht th
nfrmtn cntnt cn b xtrctd frm th rmnng smbls.''
```

```
``It deosn't mttaer in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng
is taht the frist and lsat ltteer be at the rghit pclae.  The rset can be a ttoal
mses and you can sitll raed it wouthit porbelm.  Tihs is bcuseae the huamn mnid deos
not raed ervey lteter by istlef, but the wrod as a wlohe.''
```

Language is a code, and any code trades off entropy for redundancy. And these examples should have made clear that English (or any language for that matter) has a lot of redundancy. A language (or a neural code), is efficient if it minimizes redundancy. Quantified how?

Definition 8.7 (Efficiency) *The entropy of any (nonuniform) code will be less than the maximum entropy if it were uniform, measured by the efficiency or normalized entropy*

$$\eta(X) = \frac{-\sum_{i=1}^{n} p(x_i)\ \log p(x_i)}{\log n} \tag{8.22}$$

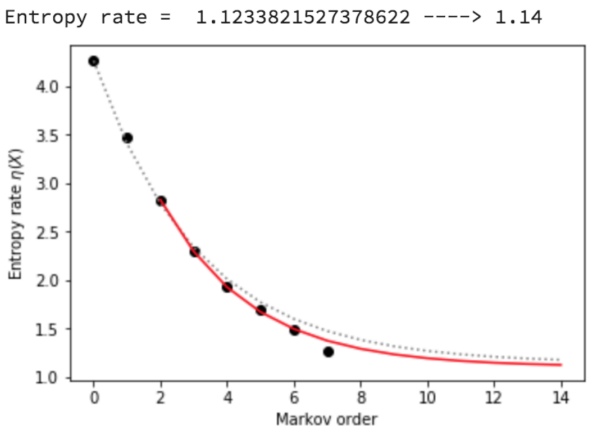
Shannon, before modern computing, constructed an experiment involving humans: subjects were asked to guess the letters of a text one by one, and incorrect guesses per character were recorded to estimate the source entropy of sentences; in this experimental setting Shannon obtained an estimate of the information rate of English between 0.6 and 1.3 bits per letter.

Definition 8.8 (Information rate) *The average entropy per symbol in a code is called the information rate.*

$$r = \lim_{n\to\infty} \frac{1}{n} H(X_n, X_{n-1}, ...X_1) = \lim_{n\to\infty} H(X_n \mid X_{n-1}, ...X_1) \ \textit{if stationary process} \tag{8.23}$$

Take a compilation of English text with $n = 14,700,000$ and assume that it is stationary, formally that $\forall l, \forall n, H(X_{n+1+l} \mid X_{n+l}, ...X_{1+l}) = H(X_{n+1} \mid X_n, ...X_1)$. We take all character sequences of length k (as a surrogate for the infinity limit) and compute the frequency, or empirical distribution, of each k-tuple in the text. We can then take the expected conditional entropy of all of these k-tuples to get an estimate of the information rate, using a hybrid of the definitions above.

It can also generate fake English text by sampling from this empirical distribution, which is increasingly convincing as we increase k:




```
k =  1  --------------------

agaca owhe ba mo fouremas phers r the ayoi hthinde—omme tthuppthe h t
asthandithearepseeronces, wnexach pat, fumia biceend be benth o olope,
by ou cathearexisincut wan, am coge isk akirm memerer aresciofthi
bldu'r atove, d saciomokrerrl wire see ischar t wearais tr din k
twod't pasu amait aingoncofo


k =  5  --------------------

ies of its basket. escape them, or poor; tell it necessibly. reg. good
escalus. i had been download timon's swelling me not suspect taste of
pain volant, i will, see! for it was too strip out. a selectronical
existed in confusion so longer, unique l'argencountess. henry. i might
1990-1993 by world w


k =  9  --------------------

fenderly public play about; and if a bawd! page  enter? ha! well keep
present my prospecting an at first, forms that is often means who once
and the saying the middle and thee, had led across in those imagined
out the towers to the you a good angels. for hose. i remembered upon a
wombs as attendants
```

If we let $k \to \infty$, we could randomly sample Shakespeare. Anyways, the entropy rate of the English language is approximately 1.14 bits per character.