Information Theory

Cosma Shalizi

15 June 2010
Complex Systems Summer School

Entropy and Information Measuring randomness and dependence in bits

Entropy and Ergodicity Dynamical systems as information sources, long-run randomness

Information and Inference The connection to statistics

Cover and Thomas (1991) is the best single book on information theory.

Entropy

The most fundamental notion in information theory
X = a discrete random variable, values from $\mathcal{X}$
The **entropy of** X is

$$H[X] \equiv -\sum_{x \in \mathcal{X}} \Pr(X = x) \log_2 \Pr(X = x)$$

Proposition

$H[X] \geq 0$, *and* $= 0$ *only when* $\Pr(X = x) = 1$ *for some* x

Proposition

$H[X] \geq 0$, and $= 0$ only when $\Pr(X = x) = 1$ for some x (EXERCISE)

Proposition

$H[X] \geq 0$, and $= 0$ only when $\Pr(X = x) = 1$ for some x (EXERCISE)

Proposition

$H[X]$ is maximal when all X are equally probable, and then $H[X] = \log_2 \#\mathcal{X}$

Proposition

$H[X] \geq 0$, and $= 0$ only when $\Pr(X = x) = 1$ for some x (EXERCISE)

Proposition

$H[X]$ is maximal when all X are equally probable, and then $H[X] = \log_2 \#\mathcal{X}$ (EXERCISE)

Proposition

$H[X] \geq 0$, and $= 0$ only when $\Pr(X = x) = 1$ for some x (Exercise)

Proposition

$H[X]$ is maximal when all X are equally probable, and then $H[X] = \log_2 \#\mathcal{X}$ (Exercise)

Proposition

$H[f(X)] \leq H[X]$, equality if and only if f is 1-1

Interpretations

$H[X]$ measures

Interpretations

$H[X]$ measures

- how *random* X is

Interpretations

$H[X]$ measures

- how *random* X is
- How *variable* X is

Interpretations

$H[X]$ measures

- how *random* X is
- How *variable* X is
- How *uncertain* we should be about X

Interpretations

$H[X]$ measures

- how *random* X is
- How *variable* X is
- How *uncertain* we should be about X
  "paleface" problem

Interpretations

$H[X]$ measures

- how *random* X is
- How *variable* X is
- How *uncertain* we should be about X

  "paleface" problem

  consistent resolution leads to a completely subjective probability theory

Interpretations

$H[X]$ measures

- how *random* X is
- How *variable* X is
- How *uncertain* we should be about X

  "paleface" problem

  consistent resolution leads to a completely subjective probability theory

but the more fundamental interpretation is **description length**

Description Length

$H[X] =$ how concisely can we describe X?
Imagine X as text message:

Description Length

$H[X]$ = how concisely can we describe X?
Imagine X as text message:

in Reno

Description Length

$H[X] =$ how concisely can we describe X?
Imagine X as text message:

in Reno
in Reno send money

Description Length

$H[X]$ = how concisely can we describe X?
Imagine X as text message:

in Reno
in Reno send money
in Reno divorce final

Description Length

$H[X]$ = how concisely can we describe X?
Imagine X as text message:

in Reno
in Reno send money
in Reno divorce final
marry me?

Description Length

$H[X]$ = how concisely can we describe X?
Imagine X as text message:

in Reno
in Reno send money
in Reno divorce final
marry me?
in Reno send lawyers guns and money kthxbai

Known and finite number of possible messages ($\#\mathcal{X}$)
I know what X is but won't show it to you
You can guess it by asking yes/no (binary) questions

First goal: ask as few questions as possible

First goal: ask as few questions as possible
Starting with “is it y?” is optimal iff $X = y$

First goal: ask as few questions as possible
Starting with “is it y?” is optimal iff $X = y$
Can always achieve no worse than $\approx \log_2 \#\mathcal{X}$ questions

First goal: ask as few questions as possible
Starting with "is it y?" is optimal iff $X = y$
Can always achieve no worse than $\approx \log_2 \#\mathcal{X}$ questions
New goal: minimize the *mean* number of questions

First goal: ask as few questions as possible
Starting with “is it y?” is optimal iff $X = y$
Can always achieve no worse than $\approx \log_2 \#\mathcal{X}$ questions
New goal: minimize the *mean* number of questions
Ask about more probable messages first

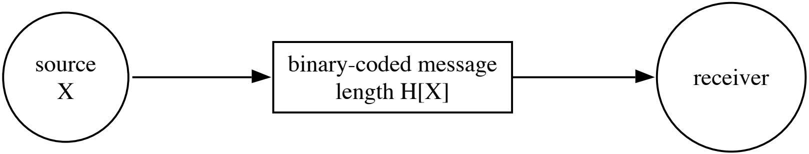
First goal: ask as few questions as possible
Starting with “is it y?” is optimal iff $X = y$
Can always achieve no worse than $\approx \log_2 \# \mathcal{X}$ questions
New goal: minimize the *mean* number of questions
Ask about more probable messages first
Still takes $\approx -\log_2 \Pr(X = x)$ questions to reach x

First goal: ask as few questions as possible
Starting with "is it y?" is optimal iff $X = y$
Can always achieve no worse than $\approx \log_2 \#\mathcal{X}$ questions
New goal: minimize the *mean* number of questions
Ask about more probable messages first
Still takes $\approx -\log_2 \Pr(X = x)$ questions to reach x
Mean is then $H[X]$

Theorem

$H[X]$ is the minimum mean number of binary distinctions needed to describe X

Units of $H[X]$ are **bits**



Multiple Variables — Joint Entropy

Joint entropy of two variables X and Y:

$$H[X,Y] \equiv -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \Pr(X=x, Y=y) \log_2 \Pr(X=x, Y=y)$$

Multiple Variables — Joint Entropy

Joint entropy of two variables X and Y:

$$H[X, Y] \equiv -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \Pr(X = x, Y = y) \log_2 \Pr(X = x, Y = y)$$

Entropy of joint distribution

Multiple Variables — Joint Entropy

Joint entropy of two variables X and Y:

$$H[X, Y] \equiv -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \Pr(X = x, Y = y) \log_2 \Pr(X = x, Y = y)$$

Entropy of joint distribution
This is the minimum mean length to describe both X and Y

Multiple Variables — Joint Entropy

Joint entropy of two variables X and Y:

$$H[X,Y] \equiv -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \Pr(X=x, Y=y) \log_2 \Pr(X=x, Y=y)$$

Entropy of joint distribution
This is the minimum mean length to describe both X and Y

$$\begin{aligned} H[X,Y] &\geq H[X] \\ H[X,Y] &\geq H[Y] \\ H[X,Y] &\leq H[X] + H[Y] \\ H[f(X),X] &= H[X] \end{aligned}$$

Conditional Entropy

Entropy of conditional distribution:

$$H[X|Y=y] \equiv -\sum_{x \in \mathcal{X}} \Pr(X=x|Y=y) \log_2 \Pr(X=x|Y=y)$$

Average over y:

$$H[X|Y] \equiv \sum_{y \in \mathcal{Y}} \Pr(Y=y) H[X|Y=y]$$

On average, how many bits are needed to describe X, *after* Y is given?

$$H[X|Y] = H[X,Y] - H[Y]$$

"text completion" principle
Note: $H[X|Y] \neq H[Y|X]$, in general

$$H[X|Y] = H[X,Y] - H[Y]$$

"text completion" principle
Note: $H[X|Y] \neq H[Y|X]$, in general
Chain rule:

$$H[X_1^n] = H[X_1] + \sum_{t=1}^{n-1} H[X_{t+1}|X_1^t]$$

Describe one variable, then describe 2nd with 1st, 3rd with first two, etc.

Mutual Information

Mutual information between X and Y

$$I[X;Y] \equiv H[X] + H[Y] - H[X,Y]$$

How much shorter is the *actual* joint description than the sum of the individual descriptions?

Mutual Information

Mutual information between X and Y

$$I[X;Y] \equiv H[X] + H[Y] - H[X,Y]$$

How much shorter is the *actual* joint description than the sum of the individual descriptions?
Equivalent:

$$I[X;Y] = H[X] - H[X|Y] = H[Y] - H[Y|X]$$

How much can I shorten my description of either variable by using the other?

Mutual Information

Mutual information between X and Y

$$I[X;Y] \equiv H[X] + H[Y] - H[X,Y]$$

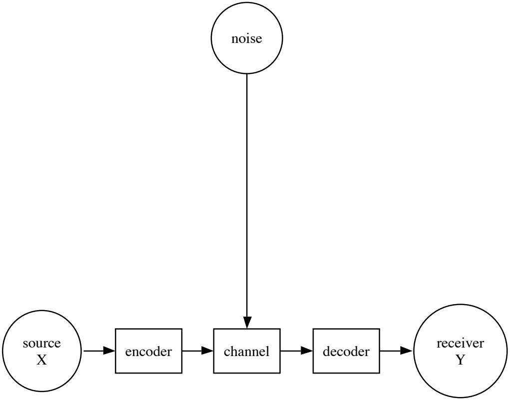
How much shorter is the *actual* joint description than the sum of the individual descriptions?
Equivalent:

$$I[X;Y] = H[X] - H[X|Y] = H[Y] - H[Y|X]$$

How much can I shorten my description of either variable by using the other?

$$0 \leq I[X;Y] \leq \min H[X], H[Y]$$

$I[X;Y] = 0$ if and only if X and Y are statistically independent



How much can we learn about what was sent from what we receive? $I[X;Y]$

Historically, this is the origin of information theory: sending coded messages efficiently (Shannon, 1948)

Historically, this is the origin of information theory: sending coded messages efficiently (Shannon, 1948)

Stephenson (1999) is a historical dramatization with silly late-1990s story tacked on

Historically, this is the origin of information theory: sending coded messages efficiently (Shannon, 1948)

Stephenson (1999) is a historical dramatization with silly late-1990s story tacked on

channel capacity $C = \max I[X; Y]$ as we change distribution of X

Historically, this is the origin of information theory: sending coded messages efficiently (Shannon, 1948)

Stephenson (1999) is a historical dramatization with silly late-1990s story tacked on

channel capacity $C = \max I[X; Y]$ as we change distribution of X

Any rate of information transfer $< C$ can be achieved with arbitrarily small error rate, *no matter what the noise*

No rate $> C$ can be achieved without error

Historically, this is the origin of information theory: sending coded messages efficiently (Shannon, 1948)

Stephenson (1999) is a historical dramatization with silly late-1990s story tacked on

channel capacity $C = \max I[X; Y]$ as we change distribution of X

Any rate of information transfer $< C$ can be achieved with arbitrarily small error rate, *no matter what the noise*

No rate $> C$ can be achieved without error

C is also related to the value of information in gambling (Poundstone, 2005)

Historically, this is the origin of information theory: sending coded messages efficiently (Shannon, 1948)

Stephenson (1999) is a historical dramatization with silly late-1990s story tacked on

channel capacity $C = \max I[X; Y]$ as we change distribution of X

Any rate of information transfer $< C$ can be achieved with arbitrarily small error rate, *no matter what the noise*

No rate $> C$ can be achieved without error

C is also related to the value of information in gambling (Poundstone, 2005)

This is *not* the only model of communication! (Sperber and Wilson, 1995, 1990)

Conditional Mutual Information

$$I[X;Y|Z] = H[X|Z] + H[Y|Z] - H[X,Y|Z]$$

How much extra information do X and Y give, over and above what's in Z?

Conditional Mutual Information

$$I[X;Y|Z] = H[X|Z] + H[Y|Z] - H[X,Y|Z]$$

How much extra information do X and Y give, over and above what's in Z?

$X \perp\!\!\!\perp Y|Z$ if and only if $I[X;Y|Z] = 0$

Conditional Mutual Information

$$I[X; Y|Z] = H[X|Z] + H[Y|Z] - H[X, Y|Z]$$

How much extra information do X and Y give, over and above what's in Z?

$X \perp\!\!\!\perp Y|Z$ if and only if $I[X; Y|Z] = 0$

Markov property is completely equivalent to

$$I[X_{t+1}^{\infty}; X_{-\infty}^{t-1}|X_t] = 0$$

Markov property is really about information flow

What About Continuous Variables?

Differential entropy:

$$H(X) \equiv -\int dx p(x) \log_2 p(x)$$

where p has to be the probability density function

What About Continuous Variables?

Differential entropy:

$$H(X) \equiv -\int dx p(x) \log_2 p(x)$$

where p has to be the probability density function
$H(X) < 0$ entirely possible

What About Continuous Variables?

Differential entropy:

$$H(X) \equiv -\int dx p(x) \log_2 p(x)$$

where p has to be the probability density function
$H(X) < 0$ entirely possible
Differential entropy *varies* under 1-1 maps (e.g. coordinate changes)

What About Continuous Variables?

Differential entropy:

$$H(X) \equiv -\int dx p(x) \log_2 p(x)$$

where p has to be the probability density function
$H(X) < 0$ entirely possible
Differential entropy *varies* under 1-1 maps (e.g. coordinate changes)
Joint and conditional entropy definitions carry over

What About Continuous Variables?

Differential entropy:

$$H(X) \equiv -\int dx p(x) \log_2 p(x)$$

where p has to be the probability density function

$H(X) < 0$ entirely possible

Differential entropy *varies* under 1-1 maps (e.g. coordinate changes)

Joint and conditional entropy definitions carry over

Mutual information definition carries over

What About Continuous Variables?

Differential entropy:

$$H(X) \equiv -\int dx p(x) \log_2 p(x)$$

where p has to be the probability density function
$H(X) < 0$ entirely possible
Differential entropy *varies* under 1-1 maps (e.g. coordinate changes)
Joint and conditional entropy definitions carry over
Mutual information definition carries over
MI *is* non-negative and invariant under 1-1 maps

Relative Entropy

$P, Q =$ two distributions on the same space $\mathcal{X}$

$$D(P\|Q) \equiv \sum_{x \in \mathcal{X}} P(x) \log_2 \frac{P(x)}{Q(x)}$$

Or, if $\mathcal{X}$ is continuous,

$$D(P\|Q) \equiv \int_{\mathcal{X}} dx\, p(x) \log_2 \frac{p(x)}{q(x)}$$

Or, if you like measure theory,

$$D(P\|Q) \equiv \int dP(\omega) \log_2 \frac{dP}{dQ}(\omega)$$

a.k.a. **Kullback-Leibler divergence**

Relative Entropy Properties

Relative Entropy Properties

$D(P\|Q) \geq 0$, with equality if and only if $P = Q$

Relative Entropy Properties

$D(P \| Q) \geq 0$, with equality if and only if $P = Q$
$D(P \| Q) \neq D(Q \| P)$, in general

Relative Entropy Properties

$D(P\|Q) \geq 0$, with equality if and only if $P = Q$
$D(P\|Q) \neq D(Q\|P)$, in general
$D(P\|Q) = \infty$ if Q gives probability zero to something with positive P probability (P not dominated by Q)

Relative Entropy Properties

$D(P\|Q) \geq 0$, with equality if and only if $P = Q$
$D(P\|Q) \neq D(Q\|P)$, in general
$D(P\|Q) = \infty$ if Q gives probability zero to something with positive P probability (P not dominated by Q)
Invariant under 1-1 maps

Joint and Conditional Relative Entropies

P, Q now distributions on $\mathcal{X}, \mathcal{Y}$

$$D(P\|Q) = D(P(X)\|Q(X)) + D(P(Y|X)\|Q(Y|X))$$

where

$$\begin{aligned} D(P(Y|X)\|Q(Y|X)) &= \sum_x P(x) D(P(Y|X=x)\|Q(Y|X=x)) \\ &= \sum_x P(x) \sum_y P(y|x) \log_2 \frac{P(y|x)}{Q(y|x)} \end{aligned}$$

and so on for more than two variables

Relative entropy can be the basic concept

$$H[X] = \log_2 m - D(P \| U)$$

where $m = \#\mathcal{X}$, $U =$ uniform dist on $\mathcal{X}$, $P =$ dist of X

Relative entropy can be the basic concept

$$H[X] = \log_2 m - D(P\|U)$$

where $m = \#\mathcal{X}$, $U =$ uniform dist on $\mathcal{X}$, $P =$ dist of X

$$I[X;Y] = D(J\|P \otimes Q)$$

where $P =$ dist of X, $Q =$ dist of Y, $J =$ joint dist

Relative Entropy and Miscoding

Suppose real distribution is P but we think it's Q and we use that for coding

Our average code length (**cross-entropy**) is

$$-\sum_{x} P(x) \log_2 Q(x)$$

But the optimum code length is

$$-\sum_{x} P(x) \log_2 P(x)$$

Difference is relative entropy

Relative entropy is the extra description length from getting the distribution wrong

Basics: Summary

Entropy $=$ minimum mean description length; variability of the random quantity
Mutual information $=$ reduction in description length from using dependencies
Relative entropy $=$ excess description length from guessing the wrong distribution

Information Sources

$X_1, X_2, \ldots X_n, \ldots$ a sequence of random variables
$X_s^t = (X_s, X_{s+1}, \ldots X_{t-1}, X_t)$
Any sort of random process process will do

Information Sources

$X_1, X_2, \ldots X_n, \ldots$ a sequence of random variables

$X_s^t = (X_s, X_{s+1}, \ldots X_{t-1}, X_t)$

Any sort of random process process will do

Sequence of messages

Successive outputs of a stochastic system

Information Sources

$X_1, X_2, \ldots X_n, \ldots$ a sequence of random variables
$X_s^t = (X_s, X_{s+1}, \ldots X_{t-1}, X_t)$
Any sort of random process process will do
Sequence of messages
Successive outputs of a stochastic system
Need not be from a communication channel
e.g., successive states of a dynamical system

Information Sources

$X_1, X_2, \ldots X_n, \ldots$ a sequence of random variables
$X_s^t = (X_s, X_{s+1}, \ldots X_{t-1}, X_t)$
Any sort of random process process will do
Sequence of messages
Successive outputs of a stochastic system
Need not be from a communication channel
e.g., successive states of a dynamical system
or *coarse-grained* observations of the dynamics

Definition (Strict or Strong Stationarity)

for any $k > 0$, $T > 0$, for all $w \in \mathcal{X}^k$

$$\Pr\left(X_1^k = w\right) = \Pr\left(X_{1+T}^{k+T} = w\right)$$

i.e., the distribution is invariant over time

Law of large numbers for stationary sequences

Theorem (Ergodic Theorem)

If X is stationary, then the empirical distribution converges

$$\hat{P}_n \rightarrow \rho$$

for some limit ρ, and for all nice functions f

$$\frac{1}{n}\sum_{t=1}^{n} f(X_t) \rightarrow \mathbf{E}_{\rho}\left[f(X)\right]$$

but ρ may be random and depend on initial conditions
one ρ per attractor

Entropy Rate

Entropy rate, a.k.a. **Shannon entropy rate**, a.k.a. **metric entropy rate**

$$h_1 \equiv \lim_{n \to \infty} H[X_n | X_1^{n-1}]$$

How many extra bits to we need to describe the next observation (in the limit)?

Theorem

h_1 exists for any stationary process (and some others)

Examples of entropy rates

IID $H[X_n|X_1^{n-1}] = H[X_1] = h_1$

Markov $H[X_n|X_1^{n-1}] = H[X_n|X_{n-1}] = H[X_2|X_1] = h_1$

k^{th}-order Markov $h_1 = H[X_{k+1}|X_1^k]$

Using chain rule, can re-write h_1 as

$$h_1 = \lim_{n \to \infty} \frac{1}{n} H[X_1^n]$$

description length per unit time

Topological Entropy Rate

$W_n \equiv$ number of allowed words of length n
$\equiv \# \left\{ w \in \mathcal{X}^n : \Pr\left(X_1^n = w\right) > 0 \right\}$
$\log_2 W_n \equiv$ **topological entropy**
topological entropy rate

$$h_0 = \lim_{n \to \infty} \frac{1}{n} \log_2 W_n$$

$H[X_1^n] = \log_2 W_n$ if and only if each word is equally probable
Otherwise $H[X_1^n] < \log_2 W_n$

Metric vs. Topological Entropy Rates

$$\begin{aligned} h_0 &= \text{growth rate in \# allowed words, counting all equally} \\ h_1 &= \text{growth rate, counting more probable words more heavily} \end{aligned}$$

Metric vs. Topological Entropy Rates

h_0 = growth rate in # allowed words, counting all equally
h_1 = growth rate, counting more probable words more heavily
effective number of words

Metric vs. Topological Entropy Rates

$h_0 =$ growth rate in # allowed words, counting all equally
$h_1 =$ growth rate, counting more probable words more heavily
effective number of words
So:

$$h_0 \geq h_1$$

Metric vs. Topological Entropy Rates

$h_0 =$ growth rate in # allowed words, counting all equally
$h_1 =$ growth rate, counting more probable words more heavily
effective number of words
So:

$$h_0 \geq h_1$$

2^{h_1} = *effective* # of choices of how to go on

KS Entropy Rate

h_1 = growth rate of mean description length of *trajectories*

Chaos needs $h_1 > 0$

Coarse-graining deterministic dynamics, each partition $\mathcal{B}$ has its own $h_1(\mathcal{B})$

Kolmogorov-Sinai (KS) entropy rate:

$$h_{KS} = \sup_{\mathcal{B}} h_1(\mathcal{B})$$

Theorem

If $\mathcal{G}$ is a generating partition, then $h_{KS} = h_1(\mathcal{G})$

h_{KS} is the *asymptotic randomness* of the dynamical system

or, the rate at which the symbol sequence provides *new information* about the initial condition

Entropy Rate and Lyapunov Exponents

In general (Ruelle's inequality),

$$h_{KS} \leq \sum_{i=1}^{d} \lambda_i \mathbf{1}_{x>0}(\lambda_i)$$

If the invariant measure is smooth, this is equality (Pesin's identity)

Asymptotic Equipartition Property

When n is large, for any word x_1^n, either

$$\Pr\left(X_1^n = x_1^n\right) \approx 2^{-nh_1}$$

or

$$\Pr\left(X_1^n = x_1^n\right) \approx 0$$

More exactly, it's almost certain that

$$-\frac{1}{n}\log \Pr\left(X_1^n\right) \rightarrow h_1$$

This is the **entropy ergodic theorem** or **Shannon-MacMillan-Breiman theorem**

Relative entropy version:

$$-\frac{1}{n}\log Q_\theta(X_1^n) \rightarrow h_1 + d(P\|Q_\theta)$$

where

$$d(P\|Q_\theta) = \lim_{n\rightarrow\infty}\frac{1}{n}D(P(X_1^n)\|Q_\theta(X_1^n))$$

Relative entropy AEP implies entropy AEP

Entropy and Ergodicity: Summary

h_1 is the growth rate of the entropy, or number of choices made in continuing the trajectory

Measures instability in dynamical systems

Typical sequences have probabilities shrinking at the entropy rate

Relative Entropy and Sampling; Large Deviations

$X_1, X_2, \ldots X_n$ all IID with distribution P
Empirical distribution $\equiv \hat{P}_n$
Law of large numbers (LLN): $\hat{P}_n \rightarrow P$

Relative Entropy and Sampling; Large Deviations

$X_1, X_2, \ldots X_n$ all IID with distribution P
Empirical distribution $\equiv \hat{P}_n$
Law of large numbers (LLN): $\hat{P}_n \rightarrow P$

Theorem (Sanov)

$$-\frac{1}{n}\log_2 \Pr\left(\hat{P}_n \in A\right) \rightarrow \underset{Q \in A}{\operatorname{argmin}}\, D(Q\|P)$$

or, for non-mathematicians,

$$\Pr\left(\hat{P}_n \approx Q\right) \approx 2^{-nD(Q\|P)}$$

Sanov's theorem is part of the general theory of **large deviations**:

$$\Pr(\text{fluctuations away from law of large numbers}) \rightarrow 0$$

exponentially in n

rate functon generally a relative entropy

More on large devations: Bucklew (1990); den Hollander (2000)
LDP explains statistical mechanics; see Touchette (2008), or talk to Eric Smith

Relative Entropy and Hypothesis Testing

Testing P vs. Q
Optimal error rate (chance of guessing Q when really P) goes like

$$\Pr\left(\mathrm{error}\right) \approx 2^{-nD(Q\|P)}$$

For dependent data, substitute sum of conditional relative entropies for nD

Relative Entropy and Hypothesis Testing

Testing P vs. Q
Optimal error rate (chance of guessing Q when really P) goes like

$$\Pr(\text{error}) \approx 2^{-nD(Q\|P)}$$

For dependent data, substitute sum of conditional relative entropies for nD
More exact statement:

$$\frac{1}{n}\log_2 \Pr(\text{error}) \rightarrow -D(Q\|P)$$

For dependent data, substitute sum conditional relative entropy rate for D

Relative Entropy and Hypothesis Testing

Testing P vs. Q
Optimal error rate (chance of guessing Q when really P) goes like

$$\Pr\left(\text{error}\right) \approx 2^{-nD(Q\|P)}$$

For dependent data, substitute sum of conditional relative entropies for nD
More exact statement:

$$\frac{1}{n}\log_2 \Pr\left(\text{error}\right) \rightarrow -D(Q\|P)$$

For dependent data, substitute sum conditional relative entropy rate for D
The bigger $D(Q\|P)$, the easier is to test which is right

Method of Maximum Likelihood

Fisher (1922)
Data $= X$ with true distribution $= P$
Model distributions $= Q_\theta$, $\theta =$ parameter
Look for the Q_θ which best describes the data
Likelihood at θ is probability of generating the data
$Q_\theta(x) \equiv \mathcal{L}(\theta)$
Estimate θ by maximizing likelihood, equivalently log-likelihood
$\mathcal{L}(\theta) \equiv \log Q_\theta(x)$

$$\widehat{\theta} \equiv \operatorname*{argmax}_{\theta} \mathcal{L}(\theta) = \operatorname*{argmax}_{\theta} \sum_{t=1}^{n} \log Q_\theta(x_t | x_1^{t-1})$$

Maximum likelihood and relative entropy

Suppose we want the Q_θ which will best describe *new* data
Optimal parameter value is

$$\theta^* = \operatorname*{argmin}_{\theta} D(P \| Q_\theta)$$

Maximum likelihood and relative entropy

Suppose we want the Q_θ which will best describe *new* data
Optimal parameter value is

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\, D(P \| Q_\theta)$$

If $P = Q_{\theta_0}$ for some θ_0, then $\theta^* = \theta_0$ (true parameter value)

Maximum likelihood and relative entropy

Suppose we want the Q_θ which will best describe *new* data
Optimal parameter value is

$$\theta^* = \underset{\theta}{\operatorname{argmin}} D(P \| Q_\theta)$$

If $P = Q_{\theta_0}$ for some θ_0, then $\theta^* = \theta_0$ (true parameter value)
Otherwise θ^* is the **pseudo-true** parameter value

$$\begin{aligned}
\theta^* &= \underset{\theta}{\operatorname{argmin}} \sum_x P(x) \log_2 \frac{P(x)}{Q_\theta(x)} \\
&= \underset{\theta}{\operatorname{argmin}} \sum_x P(x) \log_2 P(x) - P(x) \log_2 Q_\theta(x) \\
&= \underset{\theta}{\operatorname{argmin}} -H_P[X] - \sum_x P(x) \log_2 Q_\theta(x) \\
&= \underset{\theta}{\operatorname{argmin}} -\sum_x P(x) \log_2 Q_\theta(x) \\
&= \underset{\theta}{\operatorname{argmax}} \sum_x P(x) \log_2 Q_\theta(x)
\end{aligned}$$

This is the *expected log-likelihood*

We don't know P but we do have $\hat{P}_n$
For IID case

$$\begin{aligned}\hat{\theta} &= \underset{\theta}{\operatorname{argmax}} \sum_{t=1}^{n} \log Q_\theta(x_t) \\ &= \underset{\theta}{\operatorname{argmax}} \frac{1}{n} \sum_{t=1}^{n} \log_2 Q_\theta(x_t) \\ &= \underset{\theta}{\operatorname{argmax}} \sum_{x} \hat{P}_n(x) \log_2 Q_\theta(x)\end{aligned}$$

So $\hat{\theta}$ comes from approximating P by $\hat{P}_n$
$\hat{\theta} \rightarrow \theta^*$ because $\hat{P}_n \rightarrow P$

Non-IID case (e.g. Markov) similar, more notation

Relative Entropy and Log Likelihood

In general:

$$-H[X] - D(P\|Q) = \text{expected log-likelihood of } Q$$
$$-H[X] = \text{optimal expected log-likelihood (ideal model)}$$

Why Maximum Likelihood?

1. The inherent compelling rightness of the optimization principle

Why Maximum Likelihood?

1. The inherent compelling rightness of the optimization principle (a bad answer)

Why Maximum Likelihood?

1. The inherent compelling rightness of the optimization principle (a bad answer)
2. Generally **consistent**: $\widehat{\theta}$ converges on the optimal value

Why Maximum Likelihood?

1. The inherent compelling rightness of the optimization principle (a bad answer)
2. Generally **consistent**: $\widehat{\theta}$ converges on the optimal value (as we just saw)

Why Maximum Likelihood?

1. The inherent compelling rightness of the optimization principle (a bad answer)
2. Generally **consistent**: $\widehat{\theta}$ converges on the optimal value (as we just saw)
3. Generally **efficient**: converges faster than other consistent estimators

(2) and (3) are really theorems of probability theory
let's look a bit more at efficiency

Fisher Information

Fisher: Taylor-expand $\mathcal{L}$ to second order around maximum

Fisher Information

Fisher: Taylor-expand $\mathcal{L}$ to second order around maximum
Fisher information matrix

$$F_{uv}(\theta_0) \equiv -\mathbf{E}_{\theta_0}\left[\left. \frac{\partial^2 \log Q_{\theta_0}(X)}{\partial\theta_u \partial\theta_v} \right|_{\theta=\theta_0} \right]$$

Fisher Information

Fisher: Taylor-expand $\mathcal{L}$ to second order around maximum
Fisher information matrix

$$F_{uv}(\theta_0) \equiv -\mathbf{E}_{\theta_0}\left[\left.\frac{\partial^2 \log Q_{\theta_0}(X)}{\partial\theta_u \partial\theta_v}\right|_{\theta=\theta_0} \right]$$

$F \propto n$ (for IID, Markov, etc.)

Fisher Information

Fisher: Taylor-expand $\mathcal{L}$ to second order around maximum
Fisher information matrix

$$F_{uv}(\theta_0) \equiv -\mathbf{E}_{\theta_0}\left[\left. \frac{\partial^2 \log Q_{\theta_0}(X)}{\partial\theta_u \partial\theta_v} \right|_{\theta=\theta_0} \right]$$

$F \propto n$ (for IID, Markov, etc.)
Variance of $\hat{\theta} = F^{-1}$ (under some regularity conditions)

The Information Bound

Theorem (Cramér-Rao)

F^{-1} is the minimum variance for any unbiased estimator

The Information Bound

Theorem (Cramér-Rao)

F^{-1} is the minimum variance for any unbiased estimator

because uncertainty in $\hat{\theta}$ depends on curvature at maximum

The Information Bound

Theorem (Cramér-Rao)

F^{-1} is the minimum variance for any unbiased estimator

because uncertainty in $\hat{\theta}$ depends on curvature at maximum leads to a whole **information geometry**, with F as the metric tensor (Amari *et al.*, 1987; Kass and Vos, 1997; Kulhavý, 1996; Amari and Nagaoka, 1993/2000)

Relative Entropy and Fisher Information

$$
\begin{aligned}
F_{uv}(\theta_0) &\equiv -\mathbf{E}_{\theta_0}\left[\left.\frac{\partial^2 \log Q_{\theta_0}(X)}{\partial\theta_u \partial\theta_v}\right|_{\theta=\theta_0}\right] \\
&= \left.\frac{\partial^2}{\partial\theta_u \partial\theta_v} D(Q_{\theta_0} \| Q_\theta)\right|_{\theta=\theta_0}
\end{aligned}
$$

Relative Entropy and Fisher Information

$$
\begin{aligned}
F_{uv}(\theta_0) &\equiv -\mathbf{E}_{\theta_0}\left[\left.\frac{\partial^2 \log Q_{\theta_0}(X)}{\partial\theta_u\partial\theta_v}\right|_{\theta=\theta_0}\right] \\
&= \left.\frac{\partial^2}{\partial\theta_u\partial\theta_v} D(Q_{\theta_0}\|Q_\theta)\right|_{\theta=\theta_0}
\end{aligned}
$$

Fisher information is how quickly the relative entropy grows with small changes in parameters

$$
D(\theta_0\|\theta_0+\epsilon) \approx \epsilon^T F \epsilon + O(\|\epsilon\|^3)
$$

Relative Entropy and Fisher Information

$$F_{uv}(\theta_0) \equiv -\mathbf{E}_{\theta_0}\left[\left.\frac{\partial^2 \log Q_{\theta_0}(X)}{\partial\theta_u \partial\theta_v}\right|_{\theta=\theta_0}\right]$$

$$= \left.\frac{\partial^2}{\partial\theta_u \partial\theta_v} D(Q_{\theta_0} \| Q_\theta)\right|_{\theta=\theta_0}$$

Fisher information is how quickly the relative entropy grows with small changes in parameters

$$D(\theta_0 \| \theta_0 + \epsilon) \approx \epsilon^T F \epsilon + O(\|\epsilon\|^3)$$

Intuition: “easy to estimate” = “easy to reject sub-optimal values”

Maximum Entropy: A Dead End

Given *constraints* on expectation values of functions
$\mathbf{E}[g_1(X)] = c_1, \mathbf{E}[g_2(X)] = c_2, \ldots \mathbf{E}[g_q(X)] = c_q$

Maximum Entropy: A Dead End

Given *constraints* on expectation values of functions
$\mathbf{E}[g_1(X)] = c_1, \mathbf{E}[g_2(X)] = c_2, \ldots \mathbf{E}[g_q(X)] = c_q$

$$\tilde{P}_{ME} \equiv \underset{P}{\operatorname{argmax}}\, H[P] : \forall i,\ \mathbf{E}_P[g_i(X)] = c_i$$

$$= \underset{P}{\operatorname{argmax}}\, H[P] - \sum_{i=1}^{q} \lambda_i (\mathbf{E}_P[g_i(X)] - c_i)$$

with **Lagrange multipliers** λ_i chosen to enforce the constraints

Solution: Exponential Families

Generic solution:

$$P(x) = \frac{e^{-\sum_{i=1}^{q} \beta_i g_i(x)}}{\int dx e^{-\sum_{i=1}^{q} \beta_i g_i(x)}} = \frac{e^{-\sum_{i=1}^{q} \beta_i g_i(x)}}{Z(\beta_1, \beta_2, \ldots \beta_q)}$$

again β enforces constraints
Physics: **canonical ensemble** with extensive variables g_i and intensive variables β_i
Statistics: **exponential family** with sufficient statistics g_i and natural parameters β_i
If we take this family of distributions as basic, MLE is β such that $\mathbf{E}[g_i(X)] = g_i(x)$, i.e., mean = observed

Best discussion of the connection is still Mandelbrot (1962)

The Method of Maximum Entropy

Calculate sample statistics $g_i(x)$

The Method of Maximum Entropy

Calculate sample statistics $g_i(x)$
Assume that the distribution of X is the one which maximizes the entropy under those constraints

The Method of Maximum Entropy

Calculate sample statistics $g_i(x)$
Assume that the distribution of X is the one which maximizes the entropy under those constraints
i.e., the MLE in the exponential family with those sufficient statistics

The Method of Maximum Entropy

Calculate sample statistics $g_i(x)$
Assume that the distribution of X is the one which maximizes the entropy under those constraints
i.e., the MLE in the exponential family with those sufficient statistics
Refinement: **Minimum relative entropy** , minimize divergence from a reference distribution — also leads to an exponential family but with a prefactor of the base density

The Method of Maximum Entropy

Calculate sample statistics $g_i(x)$
Assume that the distribution of X is the one which maximizes the entropy under those constraints
i.e., the MLE in the exponential family with those sufficient statistics
Refinement: **Minimum relative entropy** , minimize divergence from a reference distribution — also leads to an exponential family but with a prefactor of the base density
Update distributions under new data by minimizing relative entropy

The Method of Maximum Entropy

Calculate sample statistics $g_i(x)$
Assume that the distribution of X is the one which maximizes the entropy under those constraints
i.e., the MLE in the exponential family with those sufficient statistics
Refinement: **Minimum relative entropy** , minimize divergence from a reference distribution — also leads to an exponential family but with a prefactor of the base density
Update distributions under new data by minimizing relative entropy
Often said to be the “least biased” estimate of P, or the one which makes “fewest assumptions”

About MaxEnt

MaxEnt has lots of devotees who basically think it's the answer to everything

About MaxEnt

MaxEnt has lots of devotees who basically think it's the answer to everything

And it sometimes works, because

About MaxEnt

MaxEnt has lots of devotees who basically think it's the answer to everything

And it sometimes works, because

1. Exponential families often decent approximations, MLE is cool

About MaxEnt

MaxEnt has lots of devotees who basically think it's the answer to everything

And it sometimes works, because

1. Exponential families often decent approximations, MLE is cool but not everything is an exponential family

About MaxEnt

MaxEnt has lots of devotees who basically think it's the answer to everything
And it sometimes works, because

1. Exponential families often decent approximations, MLE is cool but not everything is an exponential family
2. Conditional large deviations principle (Csiszár, 1995): if $\hat{P}$ is constrained to lie in a convex set A, then

$$-\frac{1}{n}\log \Pr\left(\hat{P} \in B | \hat{P} \in A\right) \rightarrow \inf_{Q \in B \cap A} D(Q\|P) - D(Q\|A)$$

so $\hat{P}$ is exponentially close to $\operatorname{argmin}_{Q \in A} D(Q\|P)$

About MaxEnt

MaxEnt has lots of devotees who basically think it's the answer to everything
And it sometimes works, because

1. Exponential families often decent approximations, MLE is cool but not everything is an exponential family
2. Conditional large deviations principle (Csiszár, 1995): if $\hat{P}$ is constrained to lie in a convex set A, then

$$-\frac{1}{n}\log \Pr\left(\hat{P} \in B | \hat{P} \in A\right) \rightarrow \inf_{Q \in B \cap A} D(Q \| P) - D(Q \| A)$$

so $\hat{P}$ is exponentially close to $\mathrm{argmin}_{Q \in A} D(Q \| P)$
but the conditional LDP doesn't always hold

Updating by minimizing relative entropy can disagree with Bayes's rule (Seidenfeld, 1979, 1987; Grünwald and Halpern, 2003)

Updating by minimizing relative entropy can disagree with Bayes's rule (Seidenfeld, 1979, 1987; Grünwald and Halpern, 2003) , *contra* claims by physicists

Updating by minimizing relative entropy can disagree with Bayes's rule (Seidenfeld, 1979, 1987; Grünwald and Halpern, 2003) , *contra* claims by physicists

The "constraint rule" is certainly not required by logic or probability (Seidenfeld, 1979, 1987; Uffink, 1995, 1996)

- Updating by minimizing relative entropy can disagree with Bayes's rule (Seidenfeld, 1979, 1987; Grünwald and Halpern, 2003) , *contra* claims by physicists
- The "constraint rule" is certainly not required by logic or probability (Seidenfeld, 1979, 1987; Uffink, 1995, 1996)
- MaxEnt (or MinRelEnt) is not the best rule for coming up with a prior distribution to use with Bayesian updating; all such rules suck (Kass and Wasserman, 1996)

Minimum Description Length Inference

Rissanen (1978, 1989)

Chose a model to concisely describe the data
maximum likelihood minimizes description length of the *data*
...but you need to describe the model as well!
Two-part MDL:

$$\begin{aligned}
\mathcal{D}_2(x, \theta, \Theta) &= -\log_2 Q_\theta(x) + C(\theta, \Theta) \\
\widehat{\theta}_{MDL} &= \operatorname*{argmin}_{\theta \in \Theta} \mathcal{D}_2(x, \theta, \Theta) \\
\mathcal{D}_2(x, \Theta) &= \mathcal{D}_2(x, \widehat{\theta}_{MDL}, \Theta)
\end{aligned}$$

where C is a **coding scheme** for the parameters

Must fix coding scheme before seeing the data (EXERCISE: why?)
By AEP

$$n^{-1}\mathcal{D}_2 \rightarrow h_1 + \operatorname*{argmin}_{\theta \in \Theta} d(P \| Q_\theta)$$

still for finite *n* the coding scheme matters
(One-part MDL exists but would take too long)

Why Use MDL?

1. The inherent compelling rightness of the optimization principle
2. Good properties: for reasonable sources, if the **parametric complexity**

$$\mathrm{COMP}(\Theta) = \log \sum_{w \in \mathcal{X}^n} \operatorname*{argmax}_{\theta \in \Theta} Q_\theta(w)$$

   is small — if there aren't all that many words which get high likelihoods — then if MDL did well in-sample, it will generalize well to new data from the same source

See Grünwald (2005, 2007) for much more

Information and Statistics: Summary

Relative entropy controls large deviations
Relative entropy $=$ ease of discriminating distributions
Easy discrimination $\Rightarrow$ good estimation
Large deviations explains why MaxEnt works when it does

Amari, Shun-ichi, O. E. Barndorff-Nielsen, Robert E. Kass, Steffe L. Lauritzen and C. R. Rao (1987). *Differential Geometry in Statistical Inference*, vol. 10 of *Institute of Mathematical Statistics Lecture Notes-Monographs Series*. Hayward, California: Institute of Mathematical Statistics. URL `http://projecteuclid.org/euclid.lnms/1215467056`.

Amari, Shun-ichi and Hiroshi Nagaoka (1993/2000). *Methods of Information Geometry*. Providence, Rhode Island: American Mathematical Society. Translated by Daishi Harada. As *Joho Kika no Hoho*, Tokyo: Iwanami Shoten Publishers.

Bucklew, James A. (1990). *Large Deviation Techniques in Decision, Simulation, and Estimation*. New York: Wiley-Interscience.

Cover, Thomas M. and Joy A. Thomas (1991). *Elements of Information Theory*. New York: Wiley.

Csiszár, Imre (1995). "Maxent, Mathematics, and Information Theory." In *Maximum Entropy and Bayesian Methods: Proceedings of the Fifteenth International Workshop on Maximum Entropy and Bayesian Methods* (Kenneth M. Hanson and Richard N. Silver, eds.), pp. 35–50. Dordrecht: Kluwer Academic.

den Hollander, Frank (2000). *Large Deviations*. Providence, Rhode Island: American Mathematical Society.

Fisher, R. A. (1922). "On the Mathematical Foundations of Theoretical Statistics." *Philosophical Transactions of the Royal Society A*, **222**: 309–368. URL http://digital.library.adelaide.edu.au/coll/special/fisher/stat_math.html.

Grünwald, Peter (2005). "A Tutorial Introduction to the Minimum Description Length Principle." In *Advances in Minimum Description Length: Theory and Applications* (P. Grünwald and I. J. Myung and M. Pitt, eds.). Cambridge, Massachusetts: MIT Press. URL `http://arxiv.org/abs/math.ST/0406077`.

Grünwald, Peter D. (2007). *The Minimum Description Length Principle*. Cambridge, Massachusetts: MIT Press.

Grünwald, Peter D. and Joseph Y. Halpern (2003). "Updating Probabilities." *Journal of Artificial Intelligence Research*, **19**: 243–278. doi:10.1613/jair.1164.

Kass, Robert E. and Paul W. Vos (1997). *Geometrical Foundations of Asymptotic Inference*. New York: Wiley.

Kass, Robert E. and Larry Wasserman (1996). "The Selection of Prior Distributions by Formal Rules." *Journal of the*

American Statistical Association, **91**: 1343–1370. URL http://www.stat.cmu.edu/~kass/papers/rules.pdf.

Kulhavý, Rudolf (1996). *Recursive Nonlinear Estimation: A Geometric Approach*, vol. 216 of *Lecture Notes in Control and Information Sciences*. Berlin: Springer-Verlag.

Mandelbrot, Benoit (1962). "The Role of Sufficiency and of Estimation in Thermodynamics." *Annals of Mathematical Statistics*, **33**: 1021–1038. URL http://projecteuclid.org/euclid.aoms/1177704470.

Poundstone, William (2005). *Fortune's Formula: The Untold Story of the Scientific Betting Systems That Beat the Casinos and Wall Street*. New York: Hill and Wang.

Rissanen, Jorma (1978). "Modeling by Shortest Data Description." *Automatica*, **14**: 465–471.

— (1989). *Stochastic Complexity in Statistical Inquiry*. Singapore: World Scientific.

Seidenfeld, Teddy (1979). "Why I Am Not an Objective Bayesian: Some Reflections Prompted by Rosenkrantz." *Theory and Decision*, **11**: 413–440. URL `http://www.hss.cmu.edu/philosophy/seidenfeld/relating%20to%20other%20probability%20and%20statistical%20issues/Why%20I%20Am%20Not%20an%20Objective%20B.pdf`.

— (1987). "Entropy and Uncertainty." In *Foundations of Statistical Inference* (I. B. MacNeill and G. J. Umphrey, eds.), pp. 259–287. Dordrecht: D. Reidel. URL `http://www.hss.cmu.edu/philosophy/seidenfeld/relating%20to%20other%20probability%20and%`

20statistical%20issues/Entropy%20and%20Uncertainty%20(revised).pdf.

Shannon, Claude E. (1948). "A Mathematical Theory of Communication." *Bell System Technical Journal*, **27**: 379–423. URL http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html. Reprinted in Shannon and Weaver (1963).

Shannon, Claude E. and Warren Weaver (1963). *The Mathematical Theory of Communication*. Urbana, Illinois: University of Illinois Press.

Sperber, Dan and Deirdre Wilson (1990). "Rhetoric and Relevance." In *The Ends of Rhetoric: History, Theory, Practice* (David Wellbery and John Bender, eds.), pp. 140–155. Stanford: Stanford University Press. URL http://dan.sperber.com/rhetoric.htm.

— (1995). *Relevance: Cognition and Communication*. Oxford: Basil Blackwell, 2nd edn.

Stephenson, Neal (1999). *Cryptonomicon*. New York: Avon Books.

Touchette, Hugo (2008). "The Large Deviations Approach to Statistical Mechanics." E-print, arxiv.org. URL `http://arxiv.org/abs/0804.0327`.

Uffink, Jos (1995). "Can the Maximum Entropy Principle be Explained as a Consistency Requirement?" *Studies in History and Philosophy of Modern Physics*, **26B**: 223–261. URL `http://www.phys.uu.nl/~wwwgrnsl/jos/mepabst/mepabst.html`.

— (1996). "The Constraint Rule of the Maximum Entropy Principle." *Studies in History and Philosophy of Modern*

Physics, **27**: 47–79. URL http://www.phys.uu.nl/~wwwgrnsl/jos/mep2def/mep2def.html.