

Entropic Inference

Ariel Caticha

Department of Physics

University at Albany – SUNY

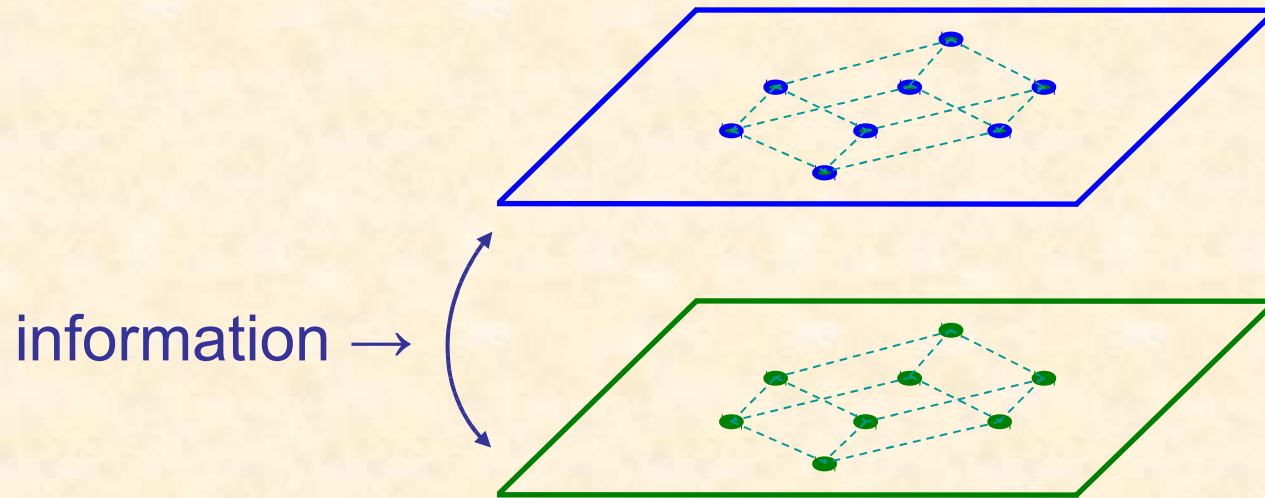
ariel@albany.edu

MaxEnt 2010

Chamonix

The goal:

To update probabilities when new information becomes available.



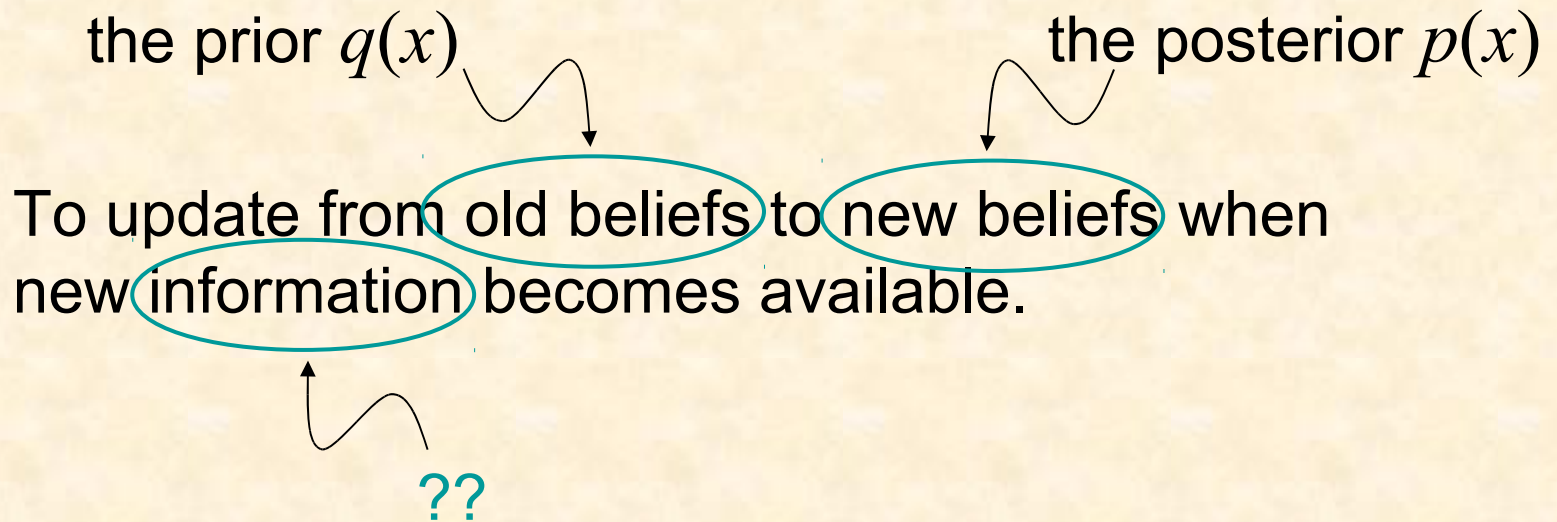
Questions:

What is information?

What is entropy? Why an entropy? Which entropy?

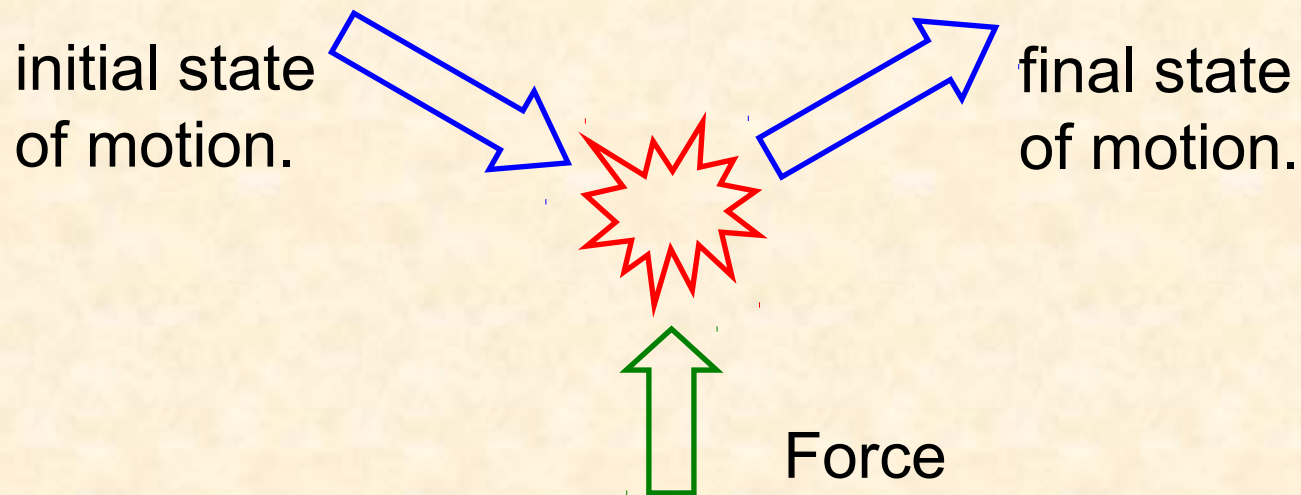
Are Bayesian and Entropic methods compatible?

The goal of entropic inference:



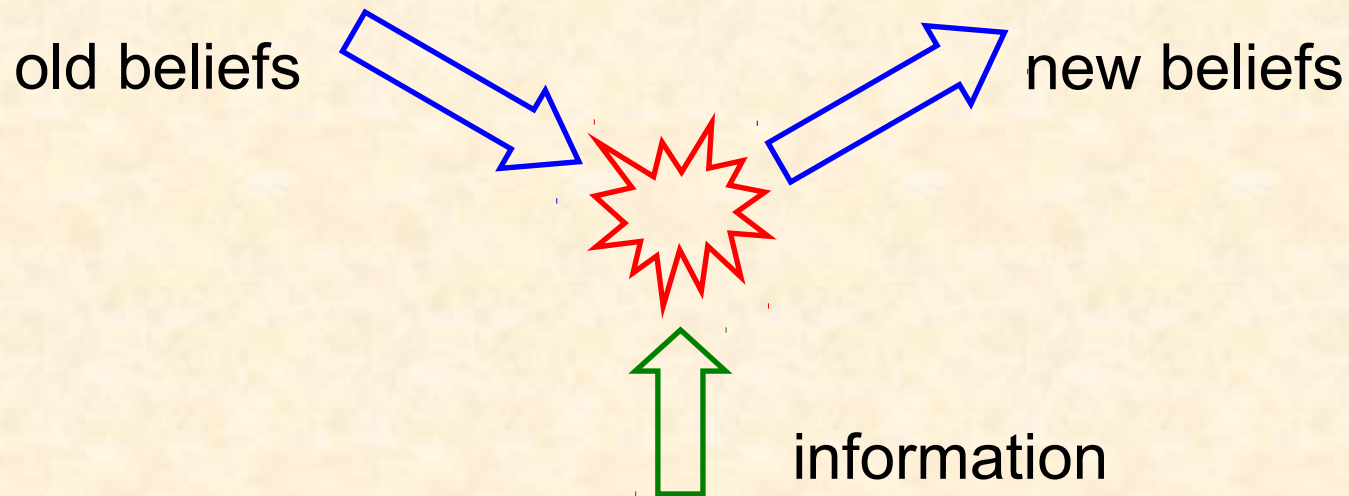
We seek a concept of information directly in terms of how it affects the beliefs of rational agents.

An analogy from physics:



Force is whatever induces a **change** of motion: $\vec{F} = \frac{dp}{dt}$

Inference is dynamics too!



Information is what induces the **change** in rational beliefs.

What is information?

Information is what induces the **change** in rational beliefs.

Information is what **constrains** rational beliefs.

Mathematical expression:

information = constraints on probabilities

Entropic Inference

Question: How do we select a distribution from among all those that satisfy the constraints?

Answer: Rank the distributions according to preference.

(Skilling)

Transitivity: if p_1 is better than p_2 ,
and p_2 is better than p_3 ,
then p_1 is better than p_3 .

To each p assign a real number $S[p, q]$ such that

$$S[p_1, q] > S[p_2, q] > S[p_3, q]$$

Remarks:

This answers the question “**Why an entropy?**”

Entropies are real numbers **designed** to be maximized.

The Method of Maximum Entropy:

Select the posterior that maximizes the entropy $S[p, q]$ subject to the available constraints.

Question: Which entropy functional $S[p, q]$?

Answer: **Eliminative Induction**

- We want an $S[p, q]$ of **universal** applicability.
- Select a **sufficiently broad** family of functionals.
- Identify **criteria/principles** that must be satisfied.
- **Eliminate** the functionals that violate the criteria.

Caution: too many criteria the universal theory might not exist.

Question: What **criteria/principles** govern the choice of the functional $S[p, q]$?

Answer: Marx

These are my principles; if you don't like them...
... I have others.

Groucho Marx

Question: What **criteria/principles** govern the choice of the functional $S[p, q]$?

Answer: **Principle of Minimal Updating**

Prior information is valuable: do not ignore it.

Beliefs ought to be revised... but only to the extent required by new information.

Rather than prescribing what and how to update we prescribe what not to update.

This is designed to maximize objectivity.

Criterion 1: Locality

Local information has local effects.

If the information does not refer to a domain D , then $p(x|D)$ is not updated,

$$p(x | D) = q(x | D).$$

Criterion 2: Coordinate invariance

Coordinates carry no information.

Criterion 3: Consistency for all independent systems

When systems are **known** to be independent it should not matter whether they are treated jointly or separately.

Remark: this applies to **all** independent systems, whether identical, similar or very different, whether few or many.

Conclusion:

The only ranking of universal applicability consistent with Minimal Updating is given by relative entropy,

$$S[p, q] = - \int dx p(x) \log \frac{p(x)}{q(x)}.$$

Other entropies may be useful for other purposes. For the purpose of updating the only candidate of general applicability is the logarithmic relative entropy.

Bayes' rule as Entropic Inference

Maximize the appropriate entropy

$$S[p, q] = - \int dx d\theta \underbrace{p(x, \theta)} \log \frac{p(x, \theta)}{\underbrace{q(x, \theta)}},$$

constrained by the data

$q(\theta)q(x | \theta)$


$$\int d\theta p(x, \theta) = p(x) = \delta(x - x')$$

observed data

Note: this is an ∞ number of constraints.

$$\delta \left\{ S + \int dx \lambda(x) [p(x) - \delta(x - x')] + \alpha [\text{norm.}] \right\} = 0$$

The joint posterior is

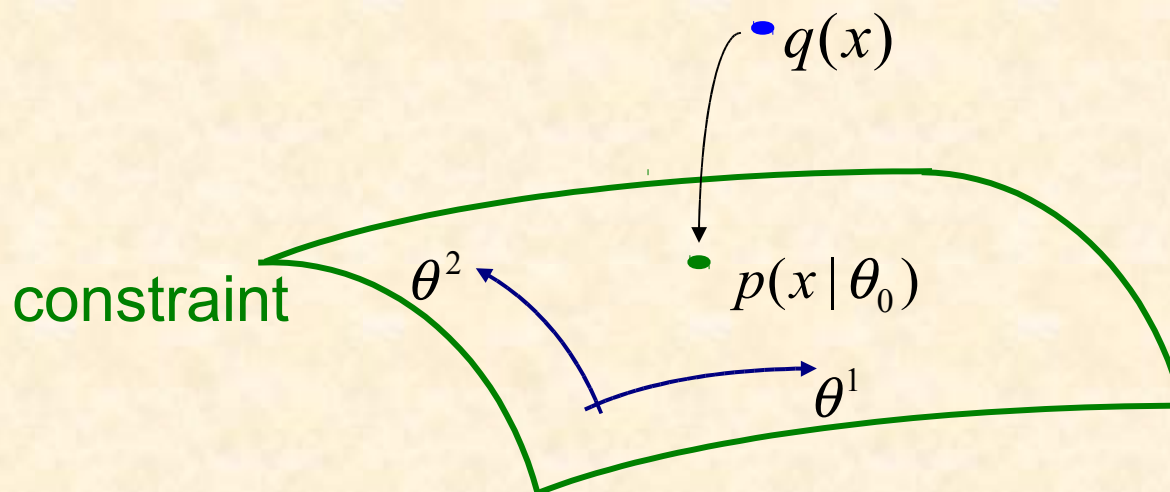
$$p(x, \theta) = p(x) p(\theta | x) = \delta(x - x') q(\theta | x)$$


and the new marginal for θ is

$$p(\theta) = \int dx p(x, \theta) = q(\theta | x') = q(\theta) \frac{q(x' | \theta)}{q(x')}$$

which is Bayes' rule !!

More on Entropic Inference



Maximum entropy selects θ_0 .

Question: To what extent is $\theta \neq \theta_0$ ruled out?

We are asking for the joint distribution $P(x, \theta)$

Maximize

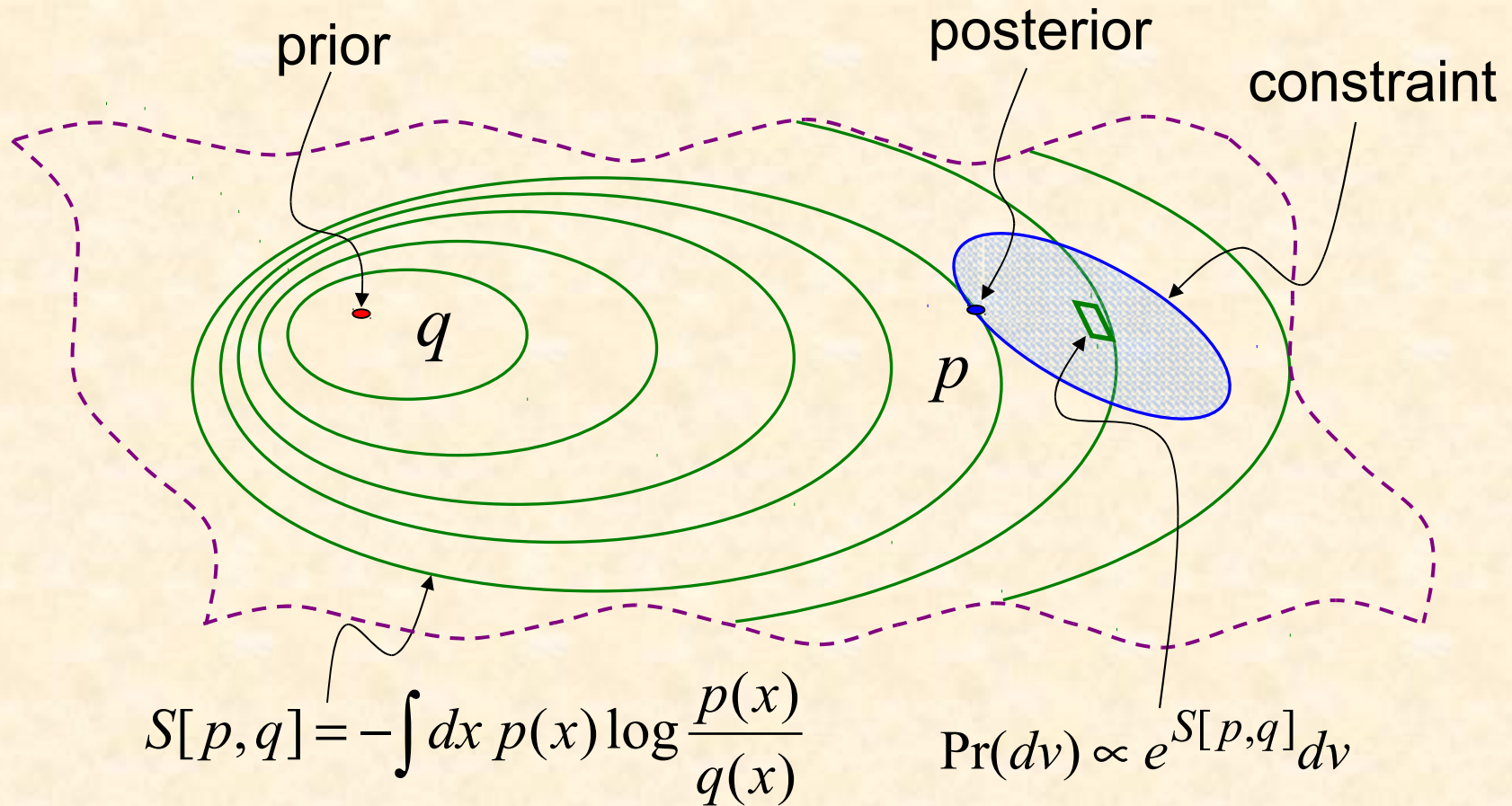
$$S[P, Q] = \int dx d\theta \underbrace{P(x, \theta)}_{P(\theta)p(x|\theta)} \log \frac{P(x, \theta)}{\underbrace{Q(x, \theta)}_{q(x)q(\theta)}}$$

Answer:

$$P(\theta) \propto e^{S(\theta)} \underbrace{q(\theta) d^n \theta}_{d \text{ Vol.}}$$

$$S(\theta) = - \int dx p(x|\theta) \log \frac{p(x|\theta)}{q(x)}$$

Entropic Inference: Summary



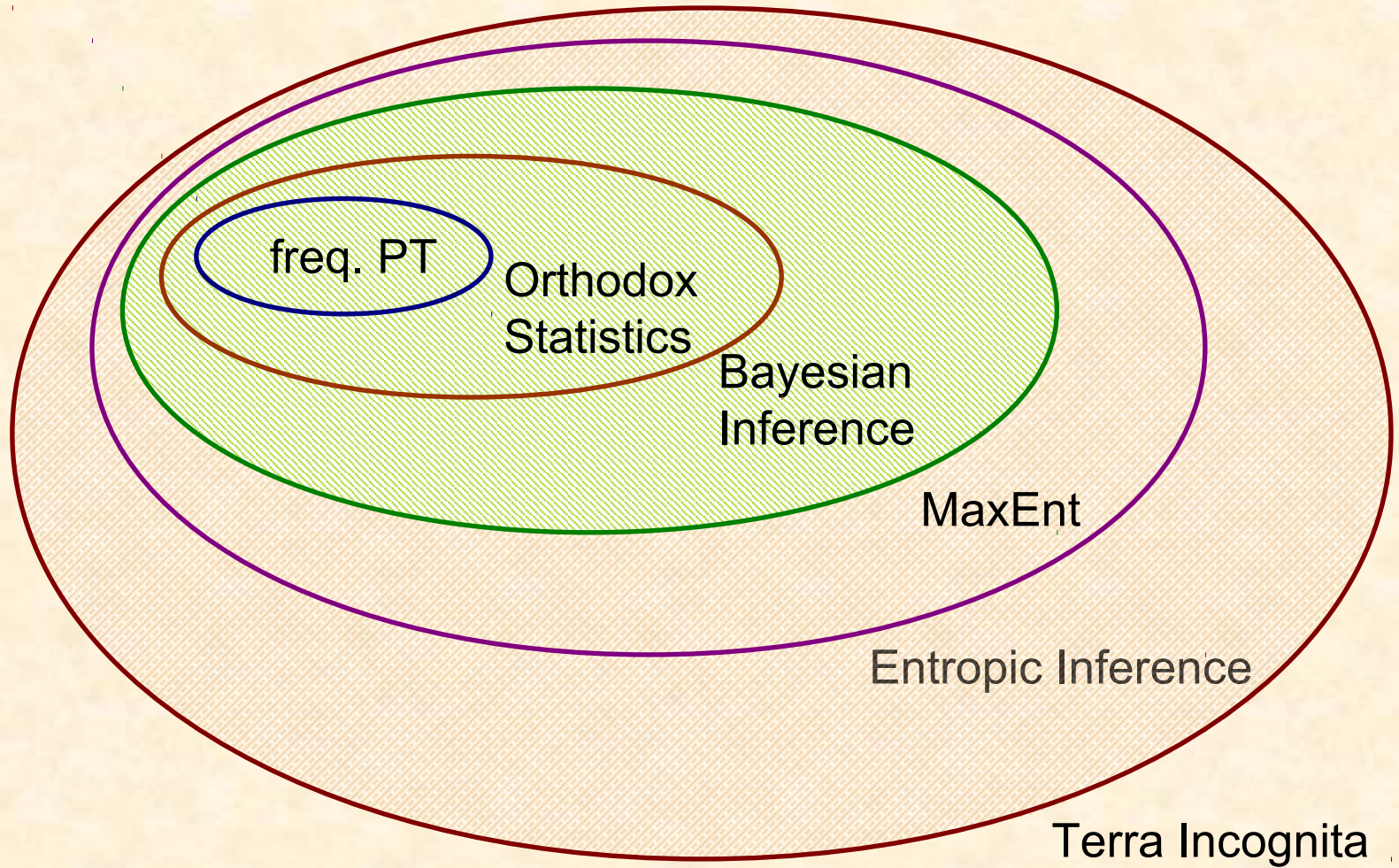
Maximize $S[p, q]$ subject to the appropriate constraints.

(MaxEnt, Bayes' rule and Large Deviations are special cases.)

Conclusions and remarks

- Information is the constraints.
- Minimal updating: prior information is valuable.
- The tool for updating is (relative) Entropy.
- Entropy needs no interpretation.
- MaxEnt, Bayes and Large Deviations are special cases.

Probability Theory— Theory of Inference



What is information?

a) Epistemic: What is conveyed by an informative answer.

Everyday usage.

Concerned with meaning.

b) Probabilistic: Shannon information.

Communication theory, Physics, Econometrics...

Concerned with amount of information, not with meaning.

c) Algorithmic: Kolmogorov complexity.

Computer science, complexity...

Concerned with amount not meaning (arguable...).

Bayes' rule for repeatable experiments

$$S[p, q] = - \int dx d\theta p(x, \theta) \log \frac{p(x, \theta)}{q(x, \theta)},$$

$$x = \{x_1 \dots x_n\} \quad \text{and} \quad q(x, \theta) = q(\theta) q(x_1 | \theta) \dots q(x_n | \theta)$$

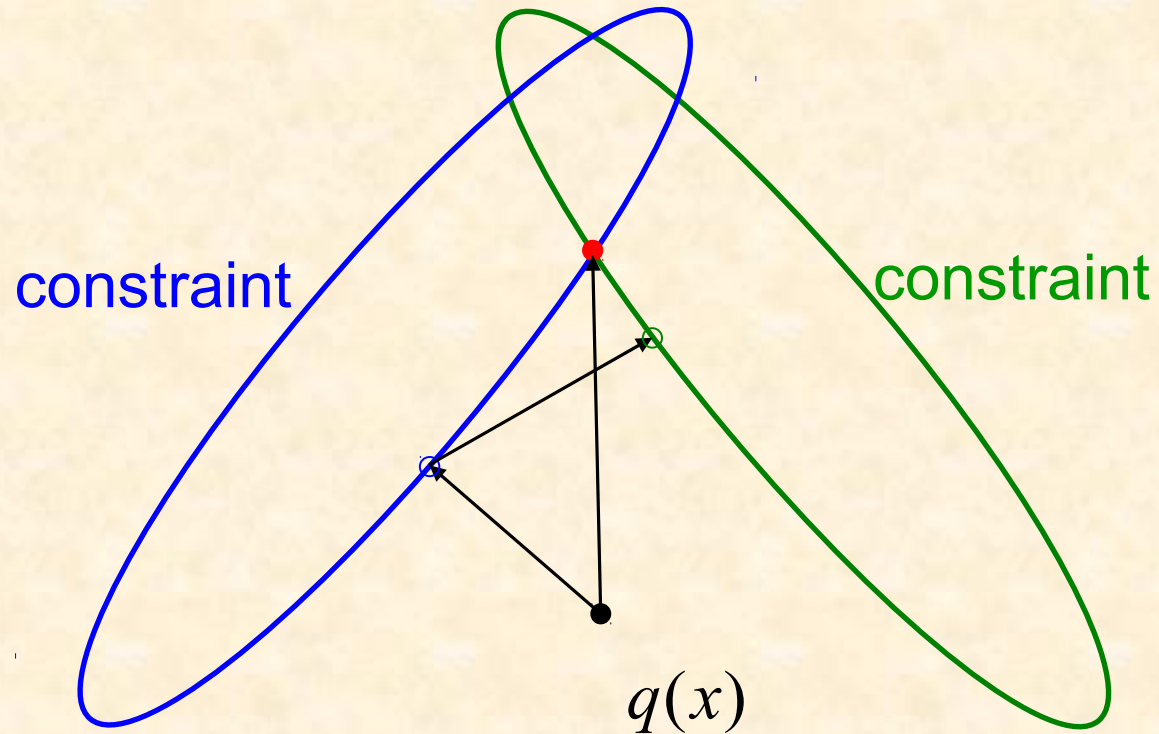
data constraints:

$$\int d\theta dx_2 \dots dx_n p(x, \theta) = p(x_1) = \delta(x_1 - x'_1)$$

posterior:

$$p(\theta, x_2 \dots x_n) = q(\theta | x'_1) q(x_2 | \theta) \dots q(x_n | \theta)$$

Constraints do not commute in general



Criterion 1: Locality

Local information has local effects.

If the information does not refer to a domain D , then $p(x|D)$ is not updated,

$$p(x | D) = q(x | D).$$

Consequence: $S[p, q] = \int dx F(x, p(x), q(x))$

$$S[p, q] = \int dx F(x, p(x), q(x))$$

Criterion 2: Coordinate invariance

Coordinates carry no information.

Consequence:
$$S[p, q] = \int \underbrace{dx m(x)} \Phi\left(\underbrace{\frac{p(x)}{m(x)}}, \underbrace{\frac{q(x)}{m(x)}}\right)$$

invariants

$$S[p, q] = \int dx m(x) \Phi\left(\frac{p(x)}{m(x)}, \frac{q(x)}{m(x)}\right)$$

To determine $m(x)$ use **Criterion 1 (Locality)** again:

If there is no new information there is no update.

Consequence: $m(x) \propto q(x)$ is the prior.

$$S[p, q] = \int dx q(x) \Phi\left(\frac{p(x)}{q(x)}\right)$$

Criterion 3: Consistency for independent systems

When systems are known to be independent it should not matter whether they are treated jointly or separately.

Consequence:

$$S_{\eta}[p, q] = \int dx p(x) \left(\frac{p(x)}{q(x)} \right)^{\eta} \quad \text{for } \eta \neq -1, 0$$

$$S_0[p, q] = - \int dx p(x) \log \frac{p(x)}{q(x)} \quad \text{for } \eta = 0$$

$$S_{-1}[p, q] = S_0[q, p] \quad \text{for } \eta = -1$$

But this applies for two systems with the same \square .

For systems with different \square s use **Criterion 3** again.

Single system 1: use $S_{\eta_1}[p_1, q_1]$

Single system 2: use $S_{\eta_2}[p_2, q_2]$

Combined system 1+2: use $S_{\eta}[p_1 p_2, q_1 q_2]$

But this is equivalent to using $S_{\eta}[p_1, q_1]$ and $S_{\eta}[p_2, q_2]$.

Therefore $\eta = \eta_1$ and $\eta = \eta_2 \Rightarrow \eta_1 = \eta_2$

\square is a universal constant!

For $N \rightarrow \infty$ systems use **Criterion 3** again.

Multinomial distribution:

$$P_N(n_1 \dots n_m | q) = \frac{N!}{n_1! \dots n_m!} q_1^{n_1} \dots q_m^{n_m}$$

For large N :

$$P_N(f_1 \dots f_m | q) \propto \exp NS_0(f, q) \quad \text{where} \quad f_i = \frac{n_i}{N}$$

\nwarrow
 $\eta = 0$

and $f_i \approx p_i$ and $\sum_i f_i a_i \approx \sum_i p_i a_i$

(in probability)