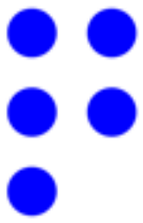
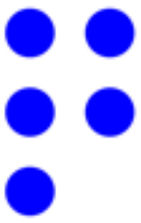
Lecture 19
How do we select priors?
Maximum Entropy



PREDICTIVE
SCIENCE LABORATORY

Objectives

- Assign prior probabilities to propositions with discrete outcomes using:
  - The Principle of Insufficient Reason.
  - The Maximum Entropy Principle.
- Derive all distributions of equilibrium statistical mechanics.
- Derive all equilibrium thermodynamic relations.
- Continuum case.



PREDICTIVE
SCIENCE LABORATORY

Principle of Insufficient Reason

PREDICTIVE
SCIENCE LABORATORY

Principle of Insufficient Reason

- You have a n discrete possible states:

$$x_1,\ldots,x_n$$

- You have no other information.

- What are the probabilities we should assign on each state?

- Principle of insufficient reason:

$$p(x_i) = \frac{1}{n}$$

Jacob Bernoulli and Pierre-Simon Laplace did not even bother giving it a name…

"The theory of chance consists in reducing all the events of the same kind to a certain number of cases equally possible, that is to say, to such as we may be equally undecided about in regard to their existence, and in determining the number of cases favorable to the event whose probability is sought. The ratio of this number to that of all the cases possible is the measure of this probability, which is thus simply a fraction whose numerator is the number of favorable cases and whose denominator is the number of all the cases possible…"

–Pierre-Simon Laplace

John-Maynard Keynes was the first who called it “principle of indifference” in 1921.

PREDICTIVE
SCIENCE LABORATORY

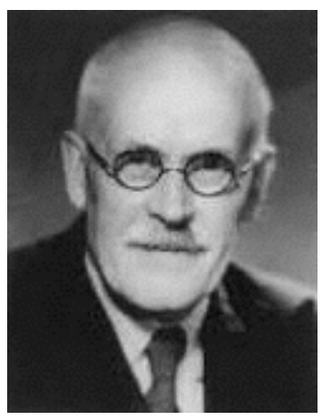
The principle of indifference was generalized by

E. T. Jaynes

principle of

transformation groups

generalizing



H. Jeffreys

principle of maximum

entropy

generalizing

C. Shannon

PREDICTIVE
SCIENCE LABORATORY

The Maximum Entropy Principle

Principle of Insufficient Reason

- You have a n discrete possible states:

$$x_1,\ldots,x_n$$

- We have some **testable** prior information.

- What are the probabilities we should assign on each state:

$$p_i = p(x_i) = ?$$

PREDICTIVE
SCIENCE LABORATORY

Testable Prior Information

- I1 = “It is certain that tanh(x) < 0.7”
- I2 = “There is at least a 90% probability that tanh(x) < 0.7”
- I3 = “The mean value of tanh(x) is 0.675”
- I4 = “The mean value of tanh(x) is probably less than 0.7”
- I5 = “There is some reason to believe that tanh(x) = 0.675”

PREDICTIVE
SCIENCE LABORATORY

Expectation Constraints

- A very common prior information is the expectation of certain functions:

$$\langle f(x) \rangle = \sum_{i=1}^{n} p_i f_k(x_i) = F_k, \; i = k = 1,\ldots,m$$

- It will lead to analytic progress.

Example 1: Broken Windows

- Statistics were collected in a recent earthquake.
- Out of 100 windows broken, there were 976 pieces found.
- But we are not given the numbers 100 and 976.
- We are told that the “average window is broken is broken into 9.76 pieces.”
- What is the probability that a window would be broken in exactly m pieces?

Example 2: Dice on black dots

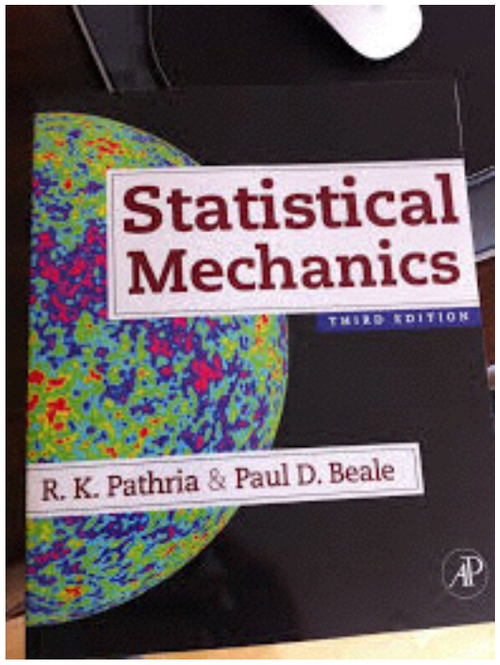
Throw the dice a few thousand times without changing the film

The average number of spots is 4.5.

What is the probability that a given face came up?

Example 3: Statistical Mechanics

- A quantum mechanical system can occupy n different states.



- The energies of the states are:

$$E_1,\ldots,E_n$$

- The average energy of the system is known to be:

$$\langle E \rangle = \bar{E}$$

- What is the probability that the system is at sate i?

Generic Problem

- You have a N discrete possible states:

$$x_1,\ldots,x_n$$

- You are told what the average of some functions of of these states is:

$$\langle f(x)\rangle = \sum_{i=1}^{n} p_i f_k(x_i) = F_k,\ i = k = 1,\ldots,m$$

- What are the probabilities we should assign on each state?

PREDICTIVE
SCIENCE LABORATORY

Idea

E. T. Jaynes:

- "The knowledge of average values does give [us] a reason for preferring some possibilities to others, but we would like […] to assign a probability distribution *which is as uniform as it can be while agreeing with the available information*."

- "The most conservative, noncommittal distribution is the one which is as spread-out as possible."

Idea

- Sounds like a variational principle…
- Maximize a “measure of uncertainty”:

$$H(p_1,\ldots,p_n)$$

- Subject to the constraints:

$$\langle f(x)\rangle = \sum_{i=1}^{n} p_i f_k(x_i) = F_k,\ i = k = 1,\ldots,m$$

Idea

- But, is there a reasonable numerical measure of how much uncertainty there is in a probability distribution?

$$H(p_1,\ldots,p_n) = ?$$

- What do you think?

Could it be the variance?

$$H(p_1,\ldots,p_n) = \sum_{i=1}^{n} (x_i - \bar{x})^2 p_i$$

Example 2: Dice on black dots

Throw the dice a few thousand times without changing the film

The average number of spots is 4.5.

What is the probability that a given face came up?

Could it be the variance?

The solution it gives to the dice problem is:

$$p_1 = 0.3, p_6 = 0.7.$$

Which, of course, does not make sense…

Shannon's Theorem

(1) We assume that some numerical measure $H_n(p_1, p_2, \ldots, p_n)$ exists; *i.e.*, that it is possible to set up some kind of association between "amount of uncertainty" and real numbers.

(2) We assume a continuity property: H_n is a continuous function of the p_i. For otherwise an arbitrarily small change in the probability distribution would still lead to the same big change in the amount of uncertainty.

(3) We require that this measure should correspond qualitatively to common sense in that when there are many possibilities, we are more uncertain than when there are few. This condition takes the form that in case the p_i are all equal, the quantity

$$h(n) = H_n\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$$

is a monotonic increasing function of n. This establishes the "sense of direction."

(4) We require that the measure H_n be consistent in the same sense as before; *i.e.* if there is more than one way of working out its value, we must get the same answer for every possible way.

Shannon's Theorem

(1) We assume that some numerical measure $H_n(p_1, p_2, \ldots, p_n)$ exists; *i.e.*, that it is possible to set up some kind of association between "amount of uncertainty" and real numbers.

(2) We assume a continuity property: H_n is a continuous function of the p_i. For otherwise an arbitrarily small change in the probability distribution would still lead to the same big change in the amount of uncertainty.

(3) We require that this measure should correspond qualitatively to common sense in that when there are many possibilities, we are more uncertain than when there are few. This condition takes the form that in case the p_i are all equal, the quantity

$$h(n) = H_n\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$$

is a monotonic increasing function of n. This establishes the "sense of direction."

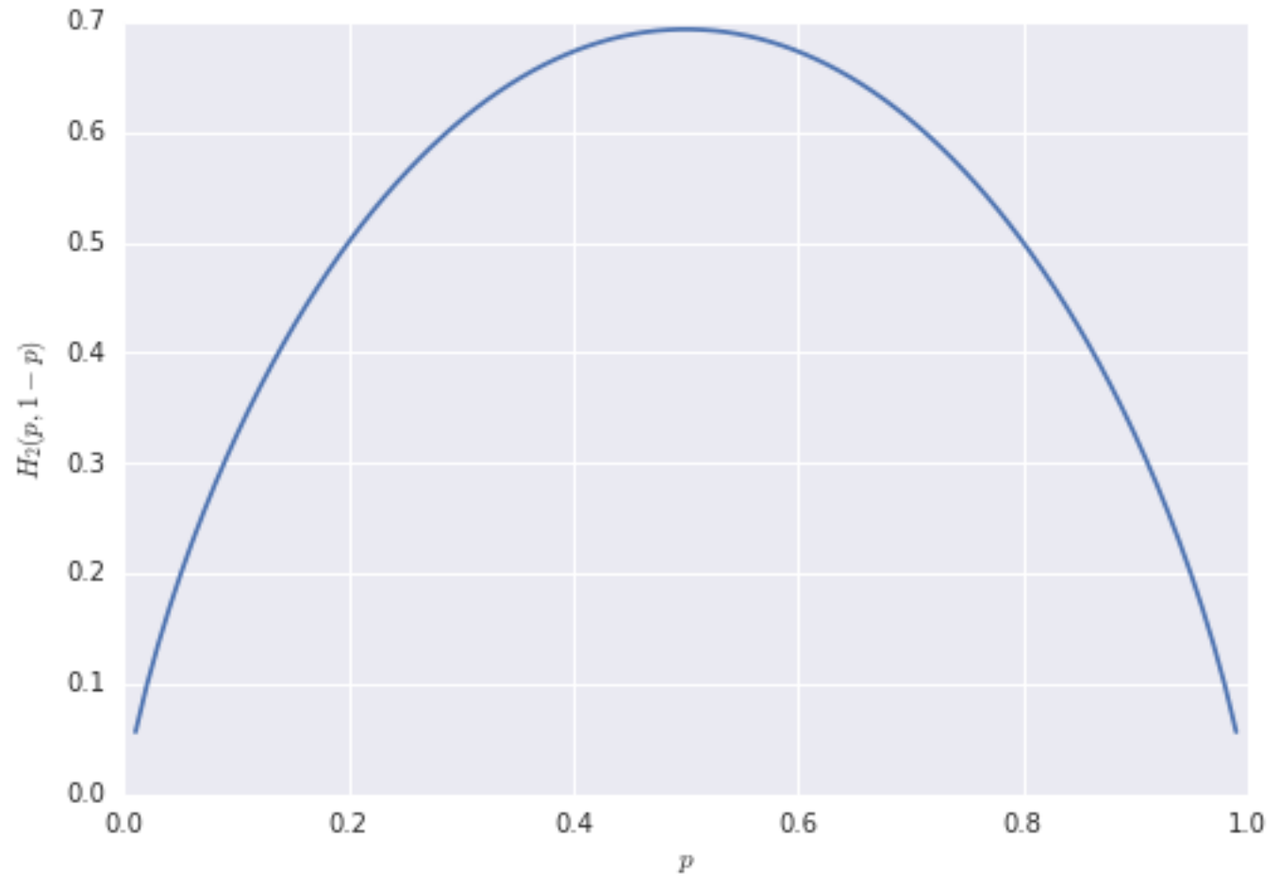
(4) We require that the measure H_n be consistent in the same sense as before; *i.e.* if there is more than one way of working out its value, we must get the same answer for every possible way.

Pose and solve some functional equations

$$H(p_1, \ldots, p_n) = -\cancel{K}^{1} \sum_{i=1}^{n} p_i \log p_i .$$

PREDICTIVE SCIENCE LABORATORY

Two State Probability Distribution



$$H(p) := H_2(p, 1-p) = -p \log p - (1-p) \log (1-p)$$

PREDICTIVE
SCIENCE LABORATORY

The Wallis Derivation

- Make a probability assignment among m mutually exclusive propositions.
- We have testable information, but we do not know how to include.
- We devise a random experiment to do so.

The Wallis Derivation

- Split our probability (total of 1) into n “quanta” of probability each worth 1/n.
- Distribute these randomly among the m propositions.
- If the probability assignment is not consistent with our info, repeat.
- If it is, then stop.

The Wallis Derivation

- Say that the number of quanta assigned to proposition 1 is n1, in 2 n2, etc.

- Then the probability we assign is actually:

$$p_i = \frac{n_i}{n}$$

- The probability that this will actually happen is:

$$m^{-n} \frac{n!}{n_1! \cdots n_m!}$$

- What is the most likely probability that will result from this game?

The Wallis Derivation

It is the one that maximizes:

$$W = \frac{n!}{n_1! \cdots n_m!}$$

$$\log n! = n \log n - n + \sqrt{2\pi n} + \frac{1}{12n} + O(\frac{1}{n^2})$$

$$\frac{1}{n} \log W \to -\sum_{i=1}^{m} p_i \log p_i = H(p_1, \ldots, p_m)$$

PREDICTIVE
SCIENCE LABORATORY

Maximum Entropy Principle

- Sounds like a variational principle…
- Maximize a “measure of uncertainty”:

$$H(p_1,\ldots,p_n) = -\sum_{i=1}^{n} p_i \log p_i.$$

- Subject to the constraints:

$$\langle f(x) \rangle = \sum_{i=1}^{n} p_i f_k(x_i) = F_k,\ i = k = 1,\ldots,m$$

No-Constraints

What is the solution with no expectation constraints?

$$\Lambda(\lambda_0) = H(p_1,\ldots,p_n) - \lambda_0\left(\sum_{i=1}^{n} p_i - 1\right)$$

$$\frac{\partial \Lambda}{\partial p_i} = 0$$

$$\Rightarrow -\log p_i - 1 - \lambda_0 p_i = 0$$

$$\Rightarrow p_i = e^{-1-\lambda_0} = \text{const}$$

$$\Rightarrow p_i = \frac{1}{n}$$

$$S = \max H(p_1,\ldots,p_n) = \log n$$

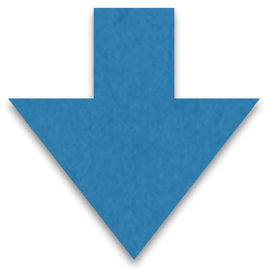
Formal Solution

- Consider the Lagrangian:

$$\Lambda(p_1,\ldots,p_n,\lambda_0,\lambda_1,\ldots,\lambda_m) = H(p_1,\ldots,p_n) - \lambda_0\left(\sum_{i=1}^{n} p_i - 1\right) - \sum_{j=1}^{m}\lambda_j\left(\sum_{i=1}^{n} f_j(x_i)p_i - F_j\right)$$

- We have to find its stationary points.

$$0 = \frac{\Lambda}{\partial p_i} = -\log p_i - 1 - \lambda_0 - \sum_{j=1}^{m}\lambda_j f_j(x_i)$$



$$p_i = \exp\left\{-1 - \lambda_0 - \sum_{j=1}^{m}\lambda_j f_j(x_i)\right\}$$

The Partition Function

$$p_i = \exp\left\{-1-\lambda_0-\sum_{j=1}^{m}\lambda_j f_j(x_i)\right\}$$

The partition function is defined to be:

$$Z(\lambda_1,\ldots,\lambda_m)=\sum_{i=1}^{n}\exp\left\{-\sum_{j=1}^{m}\lambda_j f_j(x_i)\right\}$$

$$1=\sum_{i=1}^{n}p_i=\sum_{i=1}^{n}\exp\left\{-1-\lambda_0-\sum_{j=1}^{m}\lambda_j f_j(x_i)\right\}=\exp\{-1-\lambda_0\}Z(\lambda_1,\ldots,\lambda_m)$$

$$1+\lambda_0=\log Z(\lambda_1,\ldots,\lambda_m)$$

The Partition Function

$$p_i = \frac{1}{Z(\lambda_1,\ldots,\lambda_m)} \exp\left\{-\sum_{j=1}^{m} \lambda_j f_j(x_i)\right\}$$

The partition function is defined to be:

$$Z(\lambda_1,\ldots,\lambda_m) = \sum_{i=1}^{n} \exp\left\{-\sum_{j=1}^{m} \lambda_j f_j(x_i)\right\}$$

Finding the lambdas

$$p_i = \frac{\exp\left\{-\sum_{j=1}^{m} \lambda_j f_j(x_i)\right\}}{Z(\lambda_1,\ldots,\lambda_m)}$$

$$Z(\lambda_1,\ldots,\lambda_m) = \sum_{i=1}^{n} \exp\left\{-\sum_{j=1}^{m} \lambda_j f_j(x_i)\right\}$$

$$F_k = -\frac{\partial \log Z}{\partial \lambda_k}$$

These are m non-linear equations that should be solved to determine what the lambdas are!

Properties of MENT

The attained maximum of the entropy is:

$$S(F_1,\ldots,F_m) := \max H(p_1,\ldots,p_n) = -\sum_{i=1}^{n} p_i \log p_i$$

$$p_i = \frac{1}{Z(\lambda_1,\ldots,\lambda_m)} \exp\left\{-\sum_{j=1}^{m} \lambda_j f_j(x_i)\right\}$$

$$S(F_1,\ldots,F_m) = \log Z + \sum_k \lambda_k F$$

This is the “thermodynamic entropy”.

Properties of MENT

If the thermodynamic entropy was known, then:

$$S(F_1,\ldots,F_m) = \log Z + \sum_k \lambda_k F_k$$

$$\Rightarrow \lambda_k = \frac{\partial S}{\partial F_k}.$$

The Brandeis Dice Problem

See notebook.

Properties of MENT

The best prediction we can make for any quantity q(x):

$$\langle q(x) \rangle = \sum_{i=1}^{n} p_i q(x_i).$$

and we can get various other equations in this way…

Properties of MENT

The functions may depend on some parameters:

$$f_k = f_k(\boldsymbol{x};\alpha_1,\ldots,\alpha_s).$$

(e.g., volume, magnetic field, angular velocity, etc.)

Make an arbitrarily small change in the data of the problem:

$$\{F_k,\alpha_r\} \to \{F_k + \delta F_k, \alpha_r + \delta\alpha_r\}$$

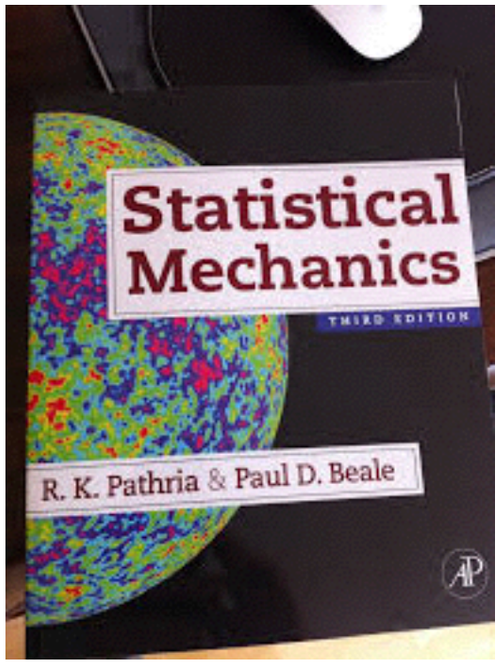
Then, we can show that the change in entropy is:

$$\delta S = \sum_k \lambda_k \delta Q_k$$

$$\delta Q_k = \delta\langle f_k \rangle - \langle \delta f_k \rangle$$

Example 3: Canonical Ensemble

- A quantum mechanical system can occupy n different states.



- The energies of the states are:

$$E_1,\ldots,E_n$$

- The average energy of the system is known to be:

$$\langle E \rangle = U$$

- What is the probability that the system is at sate i?

Example 3: Canonical Ensemble

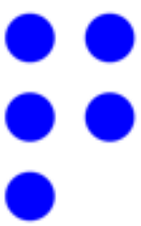
$$p_i = \frac{e^{-\lambda E_i}}{Z(\lambda)} \quad Z(\lambda) = \sum_{i=1}^{n} e^{-\lambda E_i}$$

You should use this to find lambda:

$$-\frac{\partial \log Z(\lambda)}{\partial \lambda} = U$$

Where is the temperature?

$$\lambda = (k_B T)^{-1} \Rightarrow T = \frac{k_B}{\lambda}.$$



Example 3: Canonical Ensemble

Where are the classic thermodynamics?

$$S(F_1,\ldots,F_m) = \log Z + \sum_k \lambda_k F_k \quad \Rightarrow S = \log Z + \left(k_B T\right)^{-1} U$$

$$\Rightarrow U = T(k_B S) + k_B T \log Z$$

Thermodynamic entropy:

$$S \leftarrow k_B S$$

Free energy:

$$F = k_B T \log Z$$

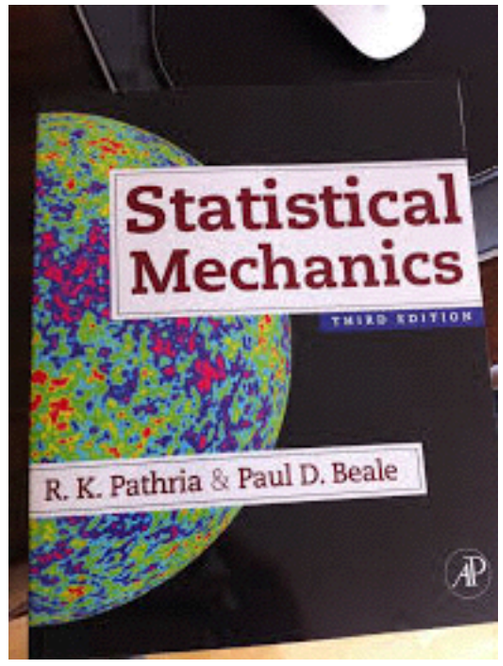
$$U = F + ST$$

Example 3: Grand Canonical Ensemble

- A quantum mechanical system comprises of many different particles of type 1, type 2, …, type s.

- The state of the system is characterized by the number of particles of each type:

$$N_1,\ldots,N_s$$



- Assume that we know, the average number of particles in each state:

$$\langle N_i \rangle = \overline{N_i}$$

- And the average energy:

$$\langle E(N_1,\ldots,N_s) \rangle = U.$$

- What is the probability that the system is at sate i?

Example 3: Canonical Ensemble

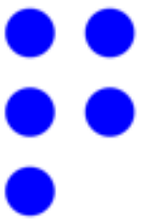
$$p_i = \frac{e^{-\sum_{i=1}^{s} \lambda_s N_s - \lambda_{s+1} E(N_1,\ldots,N_s)}}{Z(\lambda_1,\ldots,\lambda_{s+1})}$$

You should use this to find lambda:

$$-\frac{\partial \log Z}{\partial \lambda_i} = \bar{N}_i, i = 1,\ldots,s \qquad -\frac{\partial \log Z}{\partial \lambda_{s+1}} = U$$

Where is the connection?

$$\lambda_i = \mu_i (k_B T)^{-1} \qquad \lambda_{s+1} = (k_B T)^{-1}$$



Continuous Variables

$$H(p(\cdot)) = -\int p(x)\log p(x)dx.$$

There are problems with this definition…

We could equally well work with any of the following parameters:

$$y = x^3$$

$$y = \tan^{-1}(x)$$

…

Continuous Variables

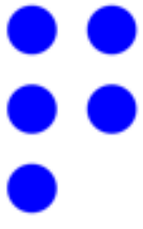
To fix the problem you need to work with:

$$H(p(\cdot)) = -\int p(x)\frac{\log p(x)}{\log m(x)}dx.$$

Where m(x) is a probability measure invariant with respect to all transformations that should not change the probabilities.

$$\max_{p(\cdot)} H(p(\cdot)) \Rightarrow p(x) = \frac{m(x)}{\int m(x)dx}$$

So, it should actually correspond to complete ignorance for x.



Continuous Variables

Maximize:

$$H(p(\cdot)) = -\int p(x)\frac{\log p(x)}{\log m(x)}dx.$$

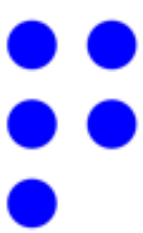
Subject to:

$$\langle f_k(x)\rangle = F_k, k = 1,\ldots,m$$

Gives:

$$p(x) = Z^{-1}m(x)\exp\left\{\sum_{k=1}^{m}\lambda_k f_k(x)\right\}$$

$$Z = \int m(x)\exp\left\{\sum_{k=1}^{K}\lambda_k f_k(x)\right\}$$



PREDICTIVE
SCIENCE LABORATORY

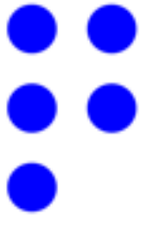
Location Parameters

x = location of something

$$m(x) \propto 1.$$

$$p(x) = Z^{-1} \exp\left\{ \sum_{k=1}^{m} \lambda_k f_k(x) \right\}$$

$$Z = \int \exp\left\{ \sum_{k=1}^{K} \lambda_k f_k(x) \right\}$$



PREDICTIVE
SCIENCE LABORATORY

Location Parameters

x = location of something

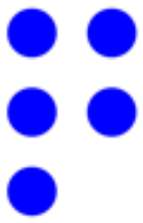
$$m(x) \propto 1. \qquad p(x) = Z^{-1} \exp\left\{\sum_{k=1}^{m} \lambda_k f_k(x)\right\} \qquad Z = \int \exp\left\{\sum_{k=1}^{K} \lambda_k f_k(x)\right\}$$

Assume that we know the mean is:

$$p(x) \propto \exp\{\lambda x\}$$

Assume that we know the first two moments:

$$p(x) \propto \exp\left\{\lambda_1 x + \lambda_2 x^2\right\}$$



PREDICTIVE
SCIENCE LABORATORY

Finding m(x)

In general, finding m(x) requires the method of:
Transformation Groups!