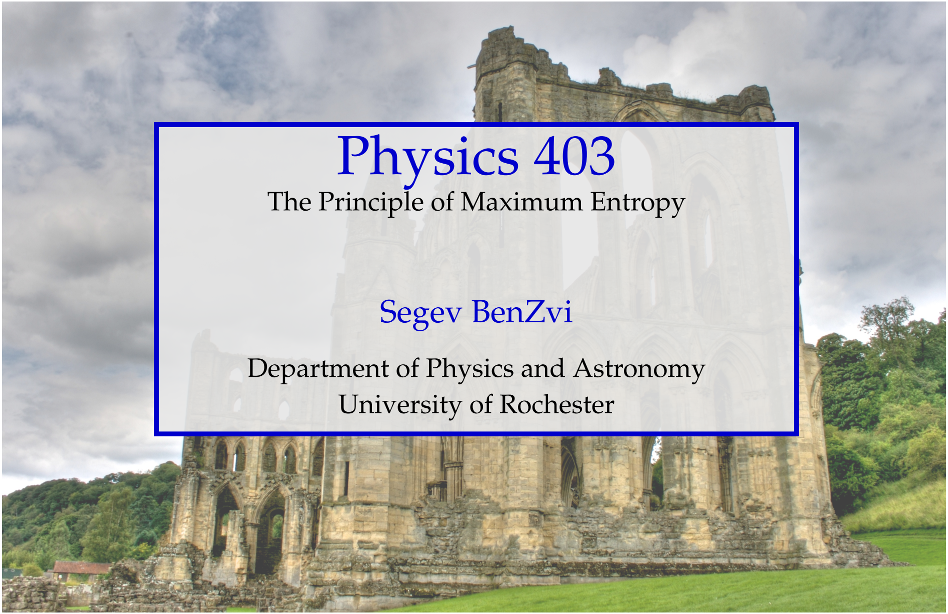


Physics 403

The Principle of Maximum Entropy

Segev BenZvi

Department of Physics and Astronomy
University of Rochester

Reading

- Sivia [1], Ch. 5, 6.2
- P. Gregory, *Bayesian Logical Data Analysis for the Physical Sciences* [2], Ch. 8

The "Kangaroo Problem"

Example

From Gull, S.F. and Skilling, J. [3]: Suppose we are given the following information. By observation,

- 1/3 of kangaroos are left-handed;
- 1/3 of kangaroos drink Foster's.

How can we estimate the proportion of kangaroos that are both left-handed (L) and drink Foster's (F) using *only* this information?

Note that the problem is trickier than you may think at first glance.

The "Kangaroo Problem"

Contingency Table

Basically, we want to construct the 2×2 table of proportions p_{ij}

	L	R	
F	p_{11}	p_{12}	$1/3$
no F	p_{21}	p_{22}	$2/3$
	$1/3$	$2/3$	

Known constraints:

- $p_{11} + p_{12} = 1/3$, since $1/3$ of kangaroos drink Foster's
- $p_{11} + p_{21} = 1/3$, since $1/3$ of kangaroos are left-handed
- $p_{11} + p_{12} + p_{21} + p_{22} = 1$ (Sum Rule)

Under the constraints, what are feasible values of the p_{ij}?

The "Kangaroo Problem"

Contingency Table

The feasible solutions to the problem have one degree of freedom, which we'll call z. From the constraints listed on the last slide, define $0 \leq z \leq \frac{1}{3}$ such that

	L	R	
F	$0 \leq z \leq \frac{1}{3}$	$\frac{1}{3} - z$	$1/3$
no F	$\frac{1}{3} - z$	$\frac{1}{3} + z$	$2/3$
	$1/3$	$2/3$	

Now consider three extreme possibilities:

- Left-handedness and Foster's drinking are independent.
- Left-handedness and Foster's drinking are fully correlated.
- Left-handedness and Foster's drinking are fully anti-correlated.

The "Kangaroo Problem"

No Correlation

If L (left-handedness) and F (Foster's drinking) are independent then

$$p_{11} = p(L, F) = p(L) \cdot p(F) = \frac{1}{3} \cdot \frac{1}{3} = \frac{1}{9},$$

i.e., whether or not a kangaroo drinks Foster's does not affect its handedness.

In this case the contingency table becomes

	L	R	
F	$\frac{1}{9}$	$\frac{2}{9}$	$1/3$
no F	$\frac{2}{9}$	$\frac{4}{9}$	$2/3$
	$1/3$	$2/3$	

The “Kangaroo Problem”

Maximum Correlation

If L and F are maximally correlated then

$$p_{11} = p(L, F) = p(L|F) \cdot p(F) = 1 \cdot \frac{1}{3} = \frac{1}{3},$$

i.e., if a kangaroo drinks Foster’s it must be left-handed.

So the contingency table becomes

	L	R	
F	$\frac{1}{3}$	0	1/3
no F	0	$\frac{2}{3}$	2/3
	1/3	2/3	

The "Kangaroo Problem"

Maximum Anticorrelation

If L and F are maximally anticorrelated then

$$p_{11} = p(L, F) = p(L|F) \cdot p(F) = 0 \cdot \frac{1}{3} = 0,$$

i.e., if a kangaroo drinks Foster's it must be right-handed.

In this case the contingency table becomes

	L	R	
F	0	$\frac{1}{3}$	$1/3$
no F	$\frac{1}{3}$	$\frac{1}{3}$	$2/3$
	$1/3$	$2/3$	

Which is the Best Choice?

- Given these three possibilities (no correlation, maximum correlation, maximum anticorrelation) which satisfies the constraints, which is the best choice?
- Q: does it make sense to select either positive or negative correlations between left-handedness and Foster's-drinking?

Which is the Best Choice?

- Given these three possibilities (no correlation, maximum correlation, maximum anticorrelation) which satisfies the constraints, which is the best choice?
- Q: does it make sense to select either positive or negative correlations between left-handedness and Foster's-drinking?
- No, because there is no prior information favoring correlation or anti-correlation
- In the absence of prior information, the uncorrelated/independent choice

$$p_{11} = p(L, F) = p(L) \cdot p(F) = \frac{1}{9}$$

is the best one.

How to Choose a Probability Distribution?

- The choice we made seems logical: there is no evidence in the data that L and F are correlated. So treating L and F as independent is correct, since it is unreasonable to assume a correlation if there is no evidence for one
- This is part of a general class of problems where we try to make an inference without enough information to evaluate a unique probability distribution
- **Principle**: "of the possible probability distributions which agree with a set of constraints, choose the one which is maximally non-committal with regard to missing information" [2]
- So how can we be maximally non-committal?
- The greater the missing information, the more uncertain the estimate. Therefore, make estimates that maximize the uncertainty in the probability distribution while being maximally constrained by the given information

Uncertainty Maximization I

Example

Suppose we run an experiment with only two possible outcomes. Which of the three probability distributions below have the most uncertain outcome?

- $p_1 = p_2 = \frac{1}{2}$
- $p_1 = \frac{1}{4}, p_2 = \frac{3}{4}$
- $p_1 = \frac{1}{100}, p_2 = \frac{99}{100}$

Uncertainty Maximization I

Example

Suppose we run an experiment with only two possible outcomes. Which of the three probability distributions below have the most uncertain outcome?

- $p_1 = p_2 = \frac{1}{2}$
- $p_1 = \frac{1}{4}, p_2 = \frac{3}{4}$
- $p_1 = \frac{1}{100}, p_2 = \frac{99}{100}$

The outcome is most uncertain if $p_1 = p_2$.

Uncertainty Maximization II

Example

Suppose we run an experiment with different numbers of outcomes. Which of the three probability distributions below have the most uncertain outcome?

- $p_1 = p_2 = \frac{1}{2}$
- $p_1 = p_2 = p_3 = p_4 = \frac{1}{4}$
- $p_1 = p_2 = \ldots = p_8 = \frac{1}{8}$

Uncertainty Maximization II

Example

Suppose we run an experiment with different numbers of outcomes. Which of the three probability distributions below have the most uncertain outcome?

- $p_1 = p_2 = \frac{1}{2}$
- $p_1 = p_2 = p_3 = p_4 = \frac{1}{4}$
- $p_1 = p_2 = \ldots = p_8 = \frac{1}{8}$

The third distribution is the most uncertain. If there are n equally probable outcomes, the uncertainty goes like n.

Shannon Entropy

- You may recognize that we are restating what earlier in the semester we called the Principle of Indifference. If there are $n > 1$ indistinguishable outcomes in an experiment, each possibility should be assigned a probability $1/n$
- **Theorem**: the uncertainty of a discrete probability distribution $\{p_i\}$ is given by the entropy [4]

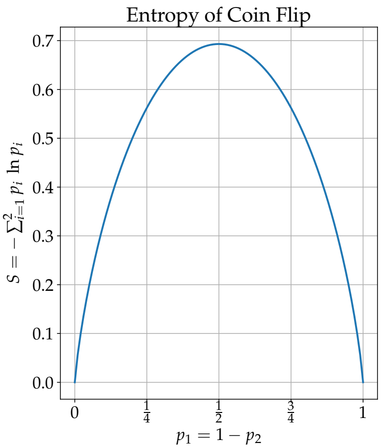
$$S(p_1, p_2, \ldots, p_n) = -\sum_{i=1}^{n} p_i \ln p_i.$$

- Assumptions: S exists, is a continuous function of the p_i, and is consistent – it gives the same answer if there are several ways of working it out.
- Moreover, more possibilities implies more uncertainty. For $p_i = 1/n$,

$$S\left(\frac{1}{n}, \ldots, \frac{1}{n}\right) = nf\left(\frac{1}{n}\right)$$

is a monotonic increasing function of n.

Entropy: Coin Flipping



- Consider the experiment with two outcomes of probability p_1 and p_2
- Our constraint is the sum rule: $p_1 + p_2 = 1$
- The Shannon entropy

$$S = -p_1 \ln p_1 - (1 - p_1) \ln (1 - p_1)$$

  is clearly maximized at $p_1 = \frac{1}{2}$
- This is where the outcome of the experiment is most uncertain

Entropy: Kangaroos

We can write down the entropy for the kangaroo problem in terms of z, the probability of kangaroos being left-handed and drinking Foster's:

$$S = -\sum_{i=1}^{4} p_i \ln p_i$$
$$= -z \ln z - 2\left(\frac{1}{3} - z\right) \ln \left(\frac{1}{3} - z\right) - \left(\frac{1}{3} + z\right) \ln \left(\frac{1}{3} + z\right)$$

Maximizing gives

$$\frac{\partial S}{\partial z} = 0 = 1 + \ln z + 2\left[\ln\left(\frac{1}{3} - z\right) + 1\right] - \left[\ln\left(\frac{1}{3} + z\right) + 1\right]$$
$$= \ln \frac{(1/3 - z)^2}{z(1/3 + z)}$$
$$1 = \frac{(1/3 - z)^2}{z(1/3 + z)} \implies \frac{z}{3} + z^2 = z^2 - \frac{2}{3}z + \frac{1}{9} \implies z = \frac{1}{9}$$

Intuition: Weighted Die

Example

Suppose we have a weighted die with unknown outcomes p_i, but we are told that

$$\text{mean number of dots} = \sum_{i=1}^{6} i\, p_i = 4.$$

(Note: for a fair die, the mean is 3.5.)

What is the unique set of outcomes p_i consistent with this constraint?

Thinking like we did in the kangaroo problem, we could make a contingency table of all possible outcomes in N tosses and reject everything which predicts a mean $\neq 4$. But this still leaves a huge number of outcomes, even for small N.

Weighted Die Contingency Table

Some hypotheses about the possible outcomes of tossing a die 10 times, from [2]:

n_{dots}	h_1	h_2	h_3	h_4	...	h_n
1	0/10	1/10	1/10	1/10		2/10
2	0/10	1/10	2/10	1/10		1/10
3	0/10	1/10	2/10	1/10		3/10
4	10/10	1/10	2/10	2/10		2/10
5	0/10	6/10	2/10	4/10		1/10
6	0/10	0/10	1/10	1/10		1/10
mean	4.0	4.0	3.5	4.0		3.2

There are quite a few combinations which produce a mean of 4 dots even for just $N = 10$ tosses of the die.

Intuition: Weighted Die

- The probability of a given set of outcomes $\boldsymbol{n} = (n_1, \ldots, n_6)$ is given by the multinomial distribution

$$p(n_1, \ldots, n_6 | N, p_1, \ldots, p_6) = \frac{N!}{n_1! \ldots n_6!} p_1^{n_1} \times \ldots \times p_6^{n_6},$$

where

$$N = \sum_i n_i = \text{total number of throws of the die.}$$

- The quantity $W = N!/(n_1! \ldots n_6!)$, or multiplicity, represents the number of states available to any given outcome h_i.
- Without knowing the $\{p_i\}$ we can't evaluate

$$p_1^{n_1} \times \ldots \times p_6^{n_6}$$

- **Claim**: the outcome h_i with the largest multiplicity W is the most probable.

Understanding the Multiplicity

- Consider the hypothesis h_1 where we roll ten 4's. There is only one way to do this, since

$$W_1 = \frac{10!}{0!0!0!10!0!0!} = 1$$

- Now consider h_4, where we roll one 1, one 2, one 3, two 4's, four 5's, and one 6:

$$W_4 = \frac{10!}{1!1!1!2!4!1!} = 75600$$

- In other words, we expect that h_4 will occur 75600 times more often than h_1 (in the absence of additional information)
- Since larger W implies a more probable hypothesis, we want to maximize W to get the most likely state h_{max} with probabilities $(p_1, \ldots, p_6)_{\text{max}}$

The Large N Limit

- In the large N limit, W increases such that there are essentially infinitely more ways W_{max} can be realized than another very similar probability distribution. It is very sharply peaked
- Using $\ln N! \approx N \ln N - N$ and $n_i = Np_i$, we can write

$$\begin{aligned}
\ln W &= \ln N! - \sum_{i=1}^{6} \ln n_i! \approx N \ln N - N - \sum_{i=1}^{6} n_i \ln n_i + \sum_{i=1}^{6} n_i \\
&= N \ln N - N - \sum_{i=1}^{6} Np_i \ln Np_i + \sum_{i=1}^{6} Np_i \\
&= N \ln N - N - N \left(\sum_{i=1}^{6} p_i \ln p_i + \ln N \right) + N \\
&= -N \sum_{i=1}^{6} p_i \ln p_i \\
&= NS
\end{aligned}$$

Behavior of the Maximum

- $\ln W = NS \implies W = \exp(NS)$
- The ratio of the maximum W_{max} to another distribution W with entropy S is

$$\frac{W_{\text{max}}}{W} = \exp[N(S_{\text{max}} - S] = \exp(N\Delta S).$$

  I.e., for large N there are effectively infinitely more ways for the max entropy solution to be realized, as we just claimed
- The quantity $2N\Delta S$ is distributed like χ^2_{M-k-1} where M is the possible number of outcomes and k is the number of constraints [5]
- Using the χ^2_{M-k-1} we can calculate the range of S about S_{max} to any desired confidence level in the usual way

Shannon-Jaynes Entropy

Up to now we have claimed total ignorance of the p_i, but what if there is some prior estimate m_i on the p_i? Then

$$
\begin{aligned}
p(n_1,\ldots,n_M|N,p_1,\ldots,p_M) &= \frac{N!}{n_1!\ldots n_M!} m_1^{n_1} \times \ldots \times m_M^{n_M} \\
\ln p(n_1,\ldots,n_M|N,p_1,\ldots,p_M) &= \sum_{i=1}^{M} n_i \ln m_i + \ln N! - \sum_{i=1}^{M} \ln n_i! \\
&= \sum_{i=1}^{M} n_i \ln m_i - N \sum_{i=1}^{M} p_i \ln p_i \\
&= N \left(\sum_{i=1}^{M} p_i \ln m_i - \sum_{i=1}^{M} p_i \ln p_i \right) \\
&= -N \sum_{i=1}^{M} p_i \ln (p_i/m_i) = NS
\end{aligned}
$$

Shannon-Jaynes Entropy

We are left with the generalized Shannon-Jaynes entropy

$$S = -\sum_{i=1}^{M} p_i \ln\left(p_i/m_i\right)$$

For the continuous case,

$$S = -\int p(x) \ln\left(\frac{p(x)}{m(x)}\right) dx$$

The quantity $m(x)$ is called the Lebesgue measure and ensures that S is invariant under the change of variables $x \to x' = f(x)$ since $m(x)$ and $p(x)$ transform in the same way.

MaxEnt and the Principle of Indifference

- We want to find a set of probabilities $p_1, \ldots, p_n$ that maximizes

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

- If all of the p_i are independent, this implies

$$dS = \frac{\partial S}{\partial p_1} dp_1 + \ldots + \frac{\partial S}{\partial p_n} dp_n = 0$$

- But if the p_i are independent, then all of the coefficients are individually equal to 0.
- Conclusion: all of the p_i are equal, i.e., $p_i = 1/n$
- Hence, the Principle of Maximum Entropy is just a formal statement of the Principle of Indifference.

MaxEnt and Constraints

Lagrange Undetermined Multipliers

- Suppose we impose a constraint on the p_i of the general form $C(p_1, \ldots, p_n) = 0$. Then

$$dC = \frac{\partial C}{\partial p_1} dp_1 + \ldots + \frac{\partial C}{\partial p_n} dp_n = 0$$

- We can combine dS and the constraint dC using a Lagrange multiplier:

$$dS - \lambda dC = 0$$

and therefore

$$dS - \lambda dC = \left(\frac{\partial S}{\partial p_1} - \lambda \frac{\partial C}{\partial p_1} \right) dp_1 + \ldots + \left(\frac{\partial S}{\partial p_n} - \lambda \frac{\partial C}{\partial p_n} \right) dp_n = 0$$

We set the first coefficient to zero, letting us solve for λ and giving M simultaneous equations for the p_i.

Normalization Constraint

Derivation of Uniform Distribution

- Let's start from the minimal possible constraint on the p_i:

$$C = \sum_{i=1}^{n} p_i = 1$$

- Therefore, from $dS - \lambda dC = 0$ we have

$$d\left[-\sum_{i=1}^{M} p_i \ln\left(p_i/m_i\right) - \lambda\left(\sum_{i=1}^{M} p_i - 1\right)\right] = 0$$

$$d\left[-\sum_{i=1}^{M} p_i \ln p_i + \sum_{i=1}^{M} p_i \ln m_i - \lambda\left(\sum_{i=1}^{M} p_i - 1\right)\right] = 0$$

$$\sum_{i=1}^{M}\left(-\ln p_i - p_i \frac{\partial \ln p_i}{\partial p_i} + \ln m_i - \lambda \frac{\partial p_i}{\partial p_i}\right) dp_i = 0$$

$$\sum_{i=1}^{M}\left(-\ln\left(p_i/m_i\right) - 1 - \lambda\right) dp_i = 0$$

Normalization Constraint

Derivation of Uniform Distribution

- Allowing the p_i to vary independently implies that all of the coefficients must vanish, so that

$$-\ln(p_i/m_i) - 1 - \lambda = 0 \implies p_i = m_i e^{-(1+\lambda)}$$

- Since $\sum p_i = 1$ and $\sum m_i = 1$,

$$1 = \sum_{i=1}^{M} p_i = \sum_{i=1}^{M} m_i e^{-(1+\lambda)} = e^{-(1+\lambda)} \sum_{i=1}^{M} m_i = e^{-(1+\lambda)}$$

  Thus, $\lambda = -1$ and

$$p_i = m_i$$

- If our prior information tells us that $m_i =$ constant, then p_i describe a uniform distribution.

Exponential: Mean Constraint

Suppose we have two constraints:

- $\sum_{i=1}^{M} p_i = 1$ (usual normalization constraint)
- $\sum_{i=1}^{M} y_i p_i = \mu$ (known mean of observations y_i)

For example, we might know the average number of dots on many throws of a die but not the results of individual throws. In this case, we need two multipliers for the two constraints:

$$d\left[-\sum_{i=1}^{M} p_i \ln\left(p_i/m_i\right) - \lambda\left(\sum_{i=1}^{M} p_i - 1\right) - \lambda_1\left(\sum_{i=1}^{M} y_i p_i - \mu\right)\right] = 0$$

$$\sum_{i=1}^{M}\left(-\ln p_i/m_i - p_i\frac{\partial \ln p_i}{\partial p_i} - \lambda\frac{\partial p_i}{\partial p_i} - y_i\lambda_1\frac{\partial p_i}{\partial p_i}\right) dp_i = 0$$

$$\sum_{i=1}^{M}\left(-\ln p_i/m_i - 1 - \lambda - y_i\lambda_1\right) dp_i = 0$$

Exponential: Mean Constraint

For all p_i, we require

$$-\ln p_i/m_i - 1 - \lambda - y_i\lambda_1 = 0$$

$$p_i = m_i\, e^{-(1+\lambda)}\, e^{-\lambda_1 y_i}$$

Applying the two constraints gives

$$\sum_{i=1}^{M} p_i = 1 = e^{-(1+\lambda)} \sum_{i=1}^{M} e^{-\lambda_1 y_i}$$

$$\sum_{i=1}^{M} y_i p_i = \mu = \frac{\sum_{i=1}^{M} y_i m_i\, e^{-\lambda_1 y_i}}{\sum_{i=1}^{M} m_i\, e^{-\lambda_1 y_i}}$$

For a given μ, one can numerically solve for λ_1.

Exponential: Mean Constraint

The discrete condition

$$p_i = m_i \, e^{-(1+\lambda)} \, e^{-\lambda_1 y_i}$$

can be generalized to the continuous distribution $p(y|I)$:

$$p(y|I) = m(y) \, e^{-(1+\lambda)} \, e^{-\lambda_1 y}$$

If $m(y) = $ constant then

$$\begin{aligned} p(y|I) &\propto e^{-\lambda_1 y} \\ &= \frac{1}{\mu} e^{-y/\mu}, \qquad y \geq 0 \end{aligned}$$

Hence, given a fixed mean observable the maximum entropy distribution of p_i (or $p(y)$) is an exponential distribution.

Gaussian: Mean and Variance Constraint

- Suppose you have a continous variable x and you constrain the mean to be μ and the variance to be σ^2:

$$\int_{x_L}^{x_H} p(x)\,dx = 1$$
$$\int_{x_L}^{x_H} x\,p(x)\,dx = \mu$$
$$\int_{x_L}^{x_H} (x-\mu)^2\,p(x)\,dx = \sigma^2$$

- In the limit that the variance is small compared to the range of the parameter, i.e.,

$$\frac{x_H-\mu}{\sigma} \gg 1 \quad \text{and} \quad \frac{\mu-x_L}{\sigma} \gg 1$$

then it turns out the maximum entropy distribution with this variance is Gaussian:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$

Utility of the Gaussian

- Suppose your data are scattered around your model with an unknown error distribution.
- It turns out that the most conservative thing you can assume (in a maximum entropy sense) is the Gaussian distribution.
- By "conservative" we mean that the Gaussian will give a greater uncertainty than what you would get from a more appropriate distribution based on more information.
- Wait, isn't that bad?
- No: for model fitting, a Gaussian model of the uncertainties is a safe choice. Other distributions may give you artificially tight constraints unless you have appropriate prior information.

Application: Lossy Image Reconstruction

- The most well-known use of maximum entropy techniques is the reconstruction of images with noise or lost pixels
- This is quite a common issue in physics and astronomy
- **Example**: consider a CCD image of a star field contaminated with pixel noise. How can we ID real stars on top of the background of noise fluctuations?
- Reconstructing damaged film is also a plot point in countless movies...

Application: Hollywood

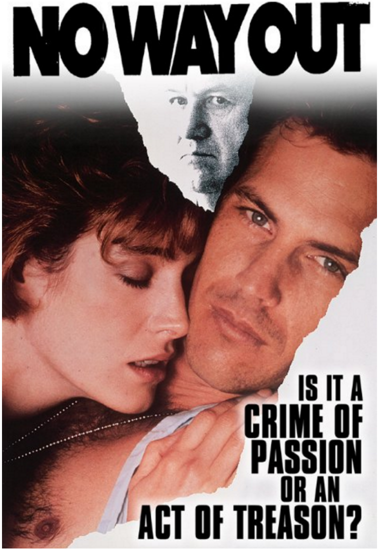


Image Reconstruction

We want to obtain the most probable image for incomplete and noisy data. Using notation from Gregory [2],

$$B \equiv \text{proposition representing prior information}$$
$$I_i \equiv \text{proposition representing a particular image}$$

Applying Bayes' Theorem, we want to solve

$$p(I_i|D,B) \propto p(D|I_i,B)\, p(I_i|B)$$

If the image consists of M pixels ($j = 1 \to M$) with measurement d_j, prediction I_{ij}, and indepdent identically distributed Gaussian noise σ_j,

$$p(d_j|I_{ij},B) \propto \exp\left[-\frac{1}{2}\left(\frac{d_j - I_{ij}}{\sigma_j}\right)^2\right]$$
$$p(D|I_i,B) \propto \prod_{j=1}^{m} \exp\left[-\frac{1}{2}\left(\frac{d_j - I_{ij}}{\sigma_j}\right)^2\right] = \exp\left[-\frac{\chi^2}{2}\right]$$

Image Reconstruction

Suppose we make trial images I_i by taking N quanta/blobs and randomly populating the M pixels. Then

$$p(I_i|B) = \frac{N!}{n_1! \ldots n_M!} \frac{1}{M^N} = \frac{W}{M^N}$$

For large N, $\ln W \rightarrow NS$, and therefore

$$p(I_i|B) = \frac{1}{M^N} e^{NS}.$$

Since we don't know the number of quanta/blobs in the image in general, we write $p(I_i|B) = \exp{(\alpha S)}$ and maximize

$$p(I_i|D, B) = \exp\left(\alpha S - \frac{\chi^2}{2}\right).$$

Example Image

- Fully Bayesian approach: marginalize α, use prior information (p_i/m_i) to enforce smoothness
- Example images from Gregory [2], taken from S.F. Gull
- Left, middle: blurred image with high noise, reconstructed at right
- Left, bottom: blurred image with low noise, reconstructed at right
- Does this seem familiar?
- This is just the Bayesian form of unfolding, accounting for blur (instrumental "smearing") and noise

References I

[1] D.S. Sivia and John Skilling. *Data Analysis: A Bayesian Tutorial*. New York: Oxford University Press, 1998.

[2] P. Gregory. *Bayesian Logical Data Analysis for the Physical Sciences*. Cambridge, UK: Cambridge University Press, 2005.

[3] Gull, S.F. and Skilling, J. "The Maximum Entropy Method". In: *Indirect Imaging*. Ed. by Roberts, J.A. Cambridge, UK: Cambridge University Press, 1984.

[4] Claude E. Shannon. "A Mathematical Theory of Communication". In: *Bell Sys. Tech. J.* 27 (1948), pp. 379–423.

[5] E.T. Jaynes. "On the Rationale of Maximum Entropy Methods". In: *Proc. IEEE* 70 (1982), pp. 939–952.