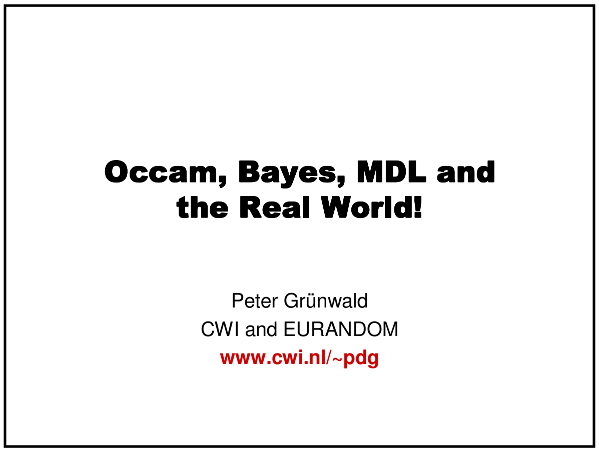
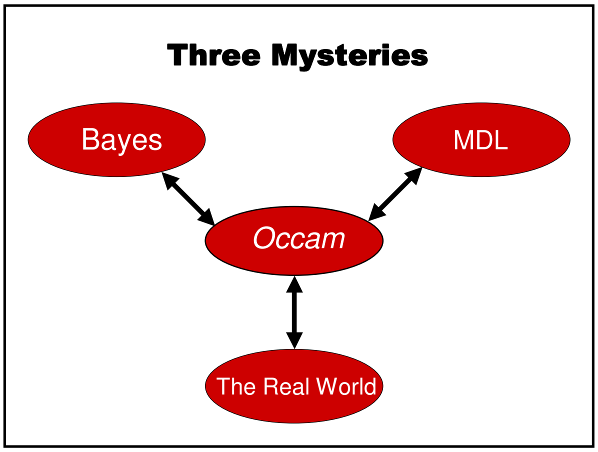


Occam, Bayes, MDL and the Real World!

Peter Grünwald
CWI and EURANDOM
www.cwi.nl/~pdg



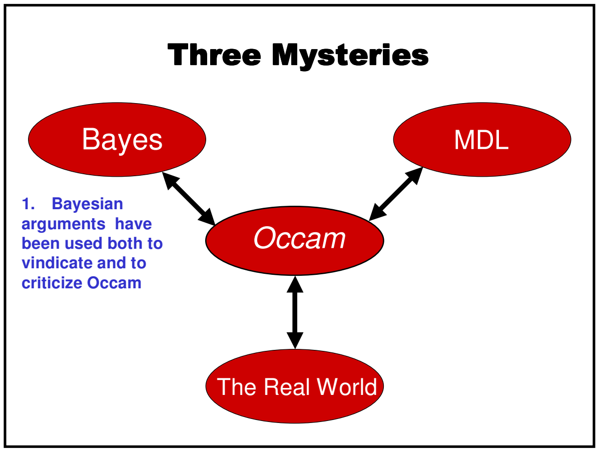
Three Mysteries

Bayes

MDL

Occam

The Real World



Three Mysteries

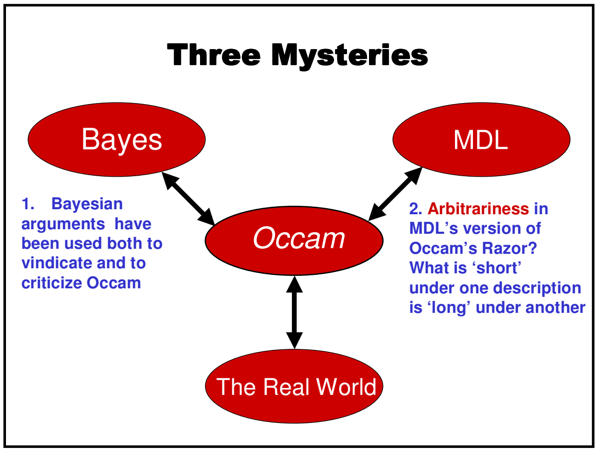
Bayes

MDL

Occam

The Real World

1. Bayesian arguments have been used both to vindicate and to criticize Occam



Three Mysteries

Bayes

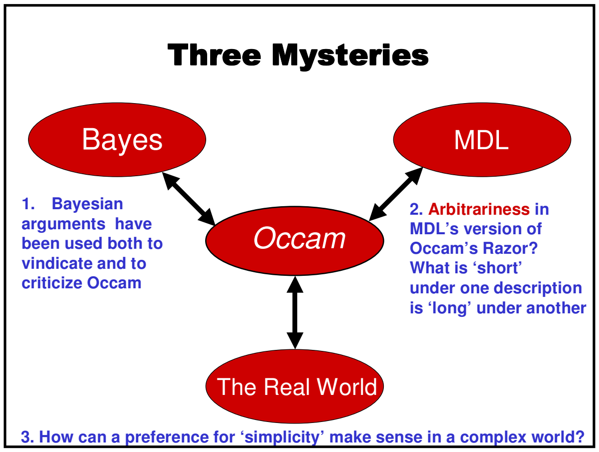
MDL

Occam

The Real World

1. Bayesian arguments have been used both to vindicate and to criticize Occam

2. Arbitrariness in MDL's version of Occam's Razor? What is 'short' under one description is 'long' under another



Three Mysteries

Bayes

MDL

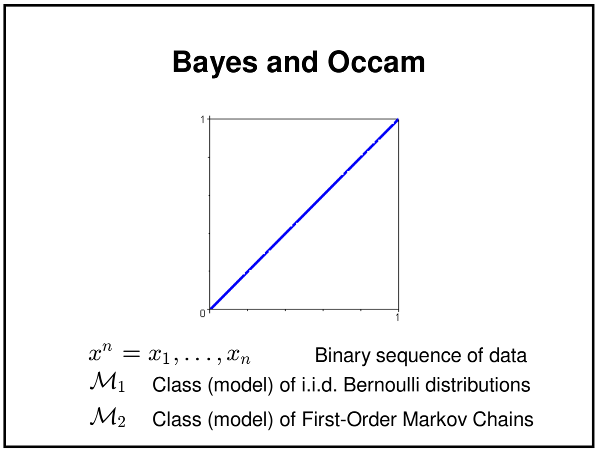
Occam

The Real World

1. Bayesian arguments have been used both to vindicate and to criticize Occam

2. Arbitrariness in MDL's version of Occam's Razor? What is 'short' under one description is 'long' under another

3. How can a preference for 'simplicity' make sense in a complex world?



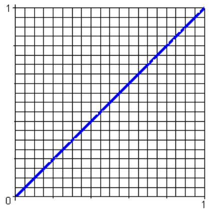
Bayes and Occam

$x^n = x_1, \ldots, x_n$ Binary sequence of data

$\mathcal{M}_1$ Class (model) of i.i.d. Bernoulli distributions

$\mathcal{M}_2$ Class (model) of First-Order Markov Chains

Bayes and Occam



$\dddot{\mathcal{M}}_1$ **Discretized** i.i.d. Bernoulli distributions

$\dddot{\mathcal{M}}_2$ **Discretized** First-Order Markov Chains

Bayesian **justification** of Occam

'Occam Factor'-type
Argument (Gull '88)

- No prior preference for $\dddot{\mathcal{M}}_1$ or $\dddot{\mathcal{M}}_2$.
  - expressed as $P(\dddot{\mathcal{M}}_1) = P(\dddot{\mathcal{M}}_2) = \frac{1}{2}$
- *Given* $\dddot{\mathcal{M}}_j$, no preference for any of the distributions in $\dddot{\mathcal{M}}_j$:
  - i.e. for all θ indexing a distribution in $\dddot{\mathcal{M}}_j$,
    $P(\theta|\dddot{\mathcal{M}}_j) = \text{const.} = \frac{1}{|\dddot{\mathcal{M}}_j|}$.
  - for example:
    $P(\theta|\dddot{\mathcal{M}}_1) = \frac{1}{100}$; $P(\theta|\dddot{\mathcal{M}}_2) = \frac{1}{100 \times 100} = \frac{1}{10000}$

Bayesian **justification** of Occam

`Occam Factor'-type
Argument (Gull '88)

- Bayesian model selection selects

$$\arg\max_j P(\dddot{\mathcal{M}}_j|x^n) = \arg\max_j P(x^n|\dddot{\mathcal{M}}_j)P(\dddot{\mathcal{M}}_j) =$$

$$\arg\max_j \sum_{\theta : P(\cdot|\theta) \in \dddot{\mathcal{M}}_j} P(x^n|\theta) \underset{\substack{\uparrow \\ \text{a lot smaller for } \dddot{\mathcal{M}}_2}}{P(\theta|\dddot{\mathcal{M}}_j)}$$

Bayesian **justification** of Occam

'Occam Factor'-type
Argument (Gull '88)

Prior for individual distribution *within* 'complex' model is much smaller. Therefore, if the simple and the complex model fit the data about equally well, Bayes selects 'simple' model.

Bayesian **criticism** of Occam

'No Free Lunch'-type Argument
(Van Allen, Greiner '00)

- No prior preference for $\dddot{\mathcal{M}}_1$ or $\dddot{\mathcal{M}}_2$.
  - previously expressed as
    $P(\dddot{\mathcal{M}}_1) = P(\dddot{\mathcal{M}}_2) = \frac{1}{2}$
  - now expressed as

$$P'(\theta) = \text{const.} = \frac{1}{|\dddot{\mathcal{M}}_1 \cup \dddot{\mathcal{M}}_2|} = \frac{1}{|\dddot{\mathcal{M}}_2|}.$$

  - for example:

$$P'(\dddot{\mathcal{M}}_1) = \textstyle\sum_{\theta \,:\, P(\cdot|\theta) \in \dddot{\mathcal{M}}_1} P'(\theta) = 100 \times \frac{1}{10000} = \frac{1}{100}$$
$$P'(\dddot{\mathcal{M}}_2) = \textstyle\sum_{\theta \,:\, P(\cdot|\theta) \in \dddot{\mathcal{M}}_2 \setminus \dddot{\mathcal{M}}_1} P'(\theta) = 1 - \frac{1}{100} = \frac{99}{100}$$

Bayesian **criticism** of Occam

- As discretization gets finer and finer, $\dddot{\mathcal{M}}_1$ gets swamped by $\dddot{\mathcal{M}}_2$ in the sense that

$$\frac{P'(\dddot{\mathcal{M}}_1)}{P'(\dddot{\mathcal{M}}_2)} \rightarrow 0$$

- Therefore, with prior $P'(\dddot{\mathcal{M}}_j)$, Bayesian model selection will always select $\dddot{\mathcal{M}}_2$, no matter what data/sample size we actually observe!

Bayesian **criticism** of Occam

'No Free Lunch'-type Argument
(Van Allen, Greiner '00)

- No prior preference for $\ddot{\mathcal{M}}_1$ or $\ddot{\mathcal{M}}_2$.
  - previously expressed as

    uniform prior over things you are interested in
  - now expressed as

    uniform prior over possible states of the world

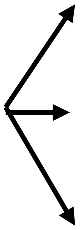
Who's Right??

short answer:

The validity of either argument depends entirely on what you mean by *'Bayesian Statistics'*!

Savage, De Finetti, Jeffreys

'modern' Bayesian Statistics has (at least) three founding fathers, each with (quite) different ideas



L. Savage
The Foundations of Statistics (1954)

B. De Finetti
Theory of Probability ('1937',1974)

H. Jeffreys
Theory of Probability (1939, 1961)

Subjective vs Pragmatic Priors

- **Savage**
  - most influential of the three
  - $P(\theta)$ is quite literally 'degree of belief that θ is true'
- **De Finetti**
  - Allows *pragmatic* priors
  - $P(\theta)$ *cannot* be interpreted as 'degree of belief that θ is true' (nevertheless, subjectivist)

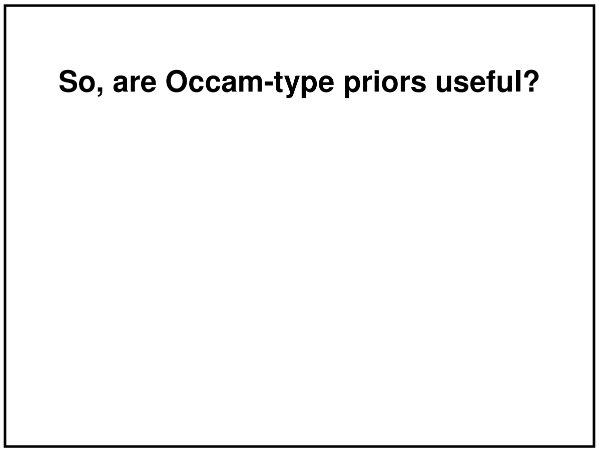
'Probabilities do not exist'
B. De Finetti, 1974, page 1

Purely Subjective vs Pragmatic Priors

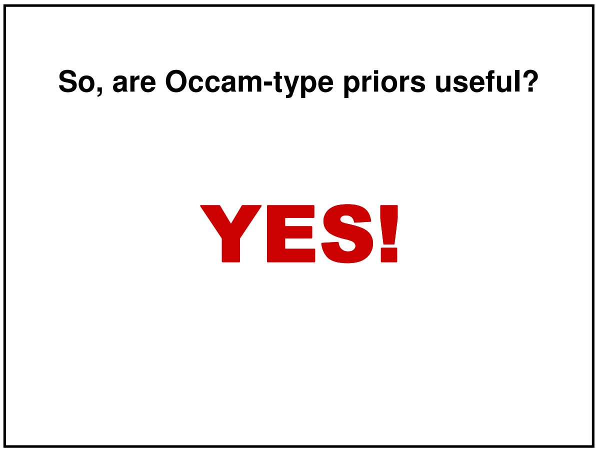
- If you insist on Savage's interpretation, and you believe that the distributions in $\ddot{\mathcal{M}}_1$ are not a priori more likely than those in $\ddot{\mathcal{M}}_2$, then you end up with NFL-type argument
- If you accept De Finetti/Jeffreys, you may choose to use Occam-type prior if it is *useful.*

Purely Subjective vs Pragmatic Priors

- If you insist on Savage's interpretation, and you believe that the distributions in $\ddot{\mathcal{M}}_1$ are not a priori more likely than those in $\ddot{\mathcal{M}}_2$, then you end up with NFL-type argument
  - IMHO, Savage's interpretation is untenable when viewed as `sole valid interpretation' of Bayesian inference: naïve Bayes, speech recognition…
- If you accept De Finetti/Jeffreys, you may choose to use Occam-type prior if it is *useful.*

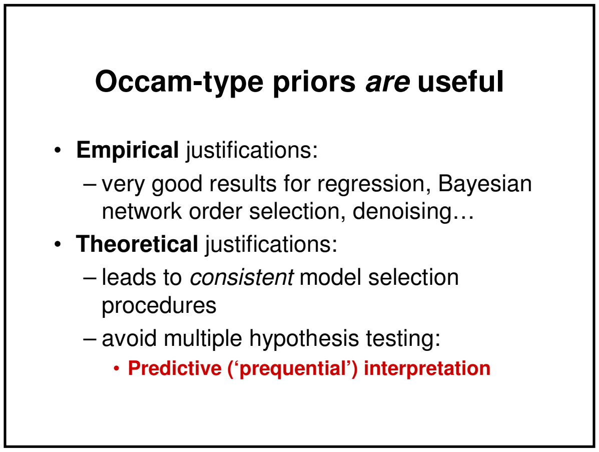


So, are Occam-type priors useful?



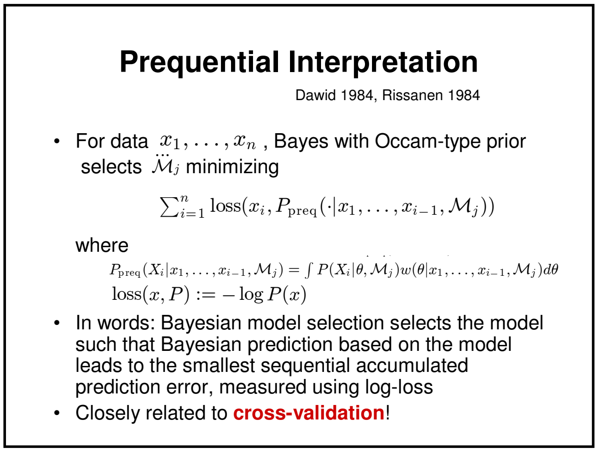
So, are Occam-type priors useful?

YES!



Occam-type priors *are* useful

- **Empirical** justifications:
  - very good results for regression, Bayesian network order selection, denoising…
- **Theoretical** justifications:
  - leads to *consistent* model selection procedures
  - avoid multiple hypothesis testing:
    - **Predictive ('prequential') interpretation**



Prequential Interpretation

Dawid 1984, Rissanen 1984

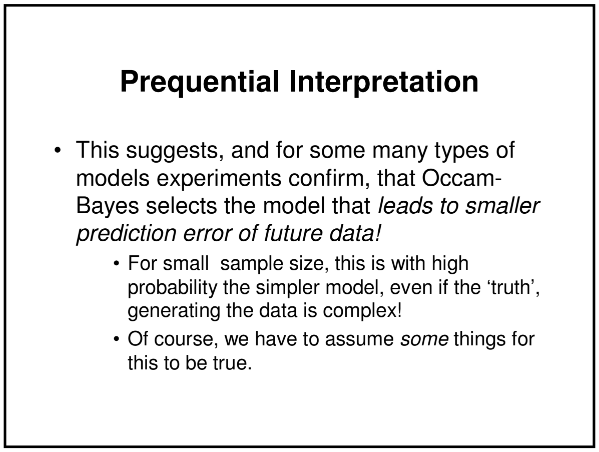
- For data $x_1, \ldots, x_n$, Bayes with Occam-type prior selects $\mathcal{M}_j$ minimizing

  $$\sum_{i=1}^{n} \mathrm{loss}(x_i, P_{\mathrm{preq}}(\cdot|x_1, \ldots, x_{i-1}, \mathcal{M}_j))$$

  where

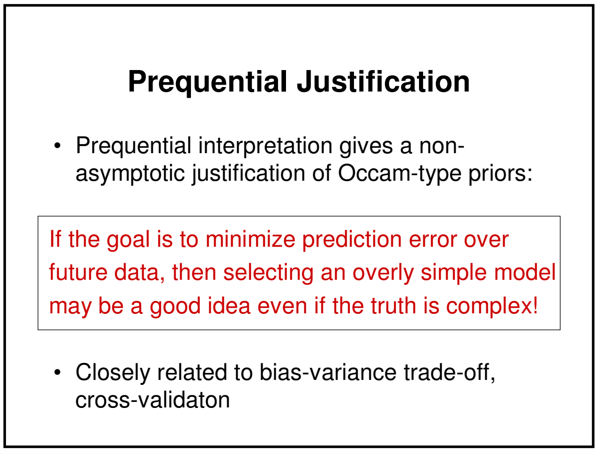
  $$P_{\mathrm{preq}}(X_i|x_1, \ldots, x_{i-1}, \mathcal{M}_j) = \int P(X_i|\theta, \mathcal{M}_j) w(\theta|x_1, \ldots, x_{i-1}, \mathcal{M}_j) d\theta$$

  $$\mathrm{loss}(x, P) := -\log P(x)$$
- In words: Bayesian model selection selects the model such that Bayesian prediction based on the model leads to the smallest sequential accumulated prediction error, measured using log-loss
- Closely related to **cross-validation**!



Prequential Interpretation

- This suggests, and for some many types of models experiments confirm, that Occam-Bayes selects the model that *leads to smaller prediction error of future data!*
  - For small sample size, this is with high probability the simpler model, even if the 'truth', generating the data is complex!
  - Of course, we have to assume *some* things for this to be true.



Prequential Justification

- Prequential interpretation gives a non-asymptotic justification of Occam-type priors:

> If the goal is to minimize prediction error over future data, then selecting an overly simple model may be a good idea even if the truth is complex!

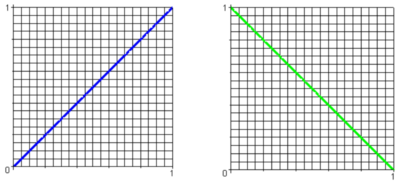
- Closely related to bias-variance trade-off, cross-validaton

Occam-type priors and MDL

- with Occam-type priors, Bayesian model selection becomes very similar to 'modern' versions of MDL…
- … but not the same (tutorial tomorrow!)
- gives alternate justification for MDL
- yet one nagging problem remains:
  - Possible *arbitrariness* in definition of 'Occam-type prior' (and hence MDL…)

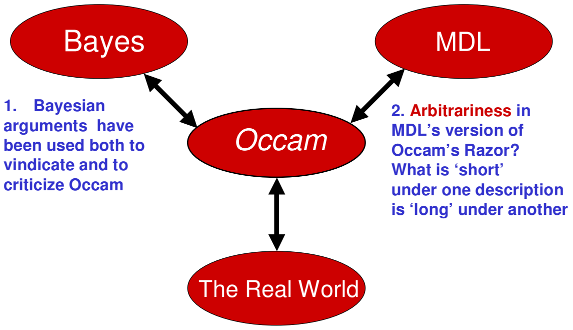
Arbitrariness?

Given $\dddot{\mathcal{M}}_2$, how should we construct $\dddot{\mathcal{M}}_1$?



Why is Bernoulli (left) more natural candidate for $\dddot{\mathcal{M}}_1$ than `reverse Bernoulli' (right) or *any other 1-dimensional submodel* of $\dddot{\mathcal{M}}_2$, for that matter?

One Mystery to Go



1. Bayesian arguments have been used both to vindicate and to criticize Occam

2. Arbitrariness in MDL's version of Occam's Razor? What is 'short' under one description is 'long' under another

3. How can a preference for 'simplicity' be useful in a complex world?

Simple models in a complex world

- Remark:
  Occam's Razor seems no good, because, after all, `What good are simple models in a complex world?'
  G. Webb (as **quoted** in KDD Nuggets 96:2)
- Answer:
  Occam's Razor is useful after all, because it is `mostly true in most real world situations'
  G. Piatetski-Shapiro (KDD Nuggets 96:2)
  (Piatetski later retracted this statement)
  (thanks to Pedro Domingos for telling me this)

'truth' of Occam's Razor is not the point!

- MDL and Bayes with pragmatic priors are *strategies* for inductive inference …
  - **Strategies are not 'true' or 'false', but 'clever' or 'stupid'!**
- …these strategies are *not at all* based on belief that 'simple models are a priori more likely to be true'
  - that idea derives from (untenable yet very influential) purely Savagian interpretation of Bayesian inference
  - much work on MDL based on assumption that 'truth is infinitely complex' (Barron and Cover, 1991)

Simple models in a complex world

- A preference for simplicity can lead to algorithms achieving better predictions for small samples, *even if truth is complex*
  - Of course *some* regularity conditions are needed!
  - Criticisms usually mention boosting, decision trees. *These are very special (yet interesting) cases!*

Thank you for your attention!