1.5 Probabilistic Reasoning

Adding non-exclusive and non-exhaustive statements:

generalized sum rule: $P(A+B) = P(A) + P(B) - P(AB)$

Product rule: $P(A,B) = P(A|B)\,P(B) = P(B|A)\,P(A)$

Bayes' theorem: $$P(A|B) = \frac{P(A,B)}{P(B)} = \frac{P(B|A)\,P(A)}{P(B)}$$

cause A $\underset{\text{deduction}}{\overset{\text{causes}}{\rightleftarrows}}$ result B $\qquad$ $P(A) \xrightarrow{P(B|A)} P(A|B)$

$P(A|B)$ "posterior" probability of A given B
$P(A)$ "prior" probability of A
$P(B|A)$ "likelihood" for A, probability of outcome of causal process $A \to B$
$P(B)$ "evidence", normalization constant, $P(B) = P(B|I)$ is likelihood for model I

1.5.1 Deductive Logic

Does probabilistic reasoning contain the syllogisms of Aristotelian logic?

strong syllogism: $I =$ "$A \Rightarrow B$" $\Rightarrow$(i) $P(B|AI) = 1$, (ii) $P(A|\overline{B}I) = 0$

proof: "$A \Rightarrow B$" $=$ "$A = AB$" $\Rightarrow P(AB|I) = P(A|I)$

(i) $P(B|AI) = \frac{P(AB|I)}{P(A|I)} = 1,$ (ii) $P(A|\overline{B}I) = \frac{P(A\overline{B}|I)}{P(\overline{B}|I)} = \frac{P(AB\overline{B}|I)}{P(\overline{B}|I)} = 0$

unless $P(\overline{B}|I) = 0$, which turns r.h.s. into empty statement□

weak syllogism: $I =$ "$A \Rightarrow B$" $\Rightarrow P(A|BI) \geq P(A|I)$

proof: $P(B|AI) = 1$ was shown above

$P(A|BI) = \frac{P(B|AI)\,P(A|I)}{P(B|I)} = \frac{P(A|I)}{P(B|I)} \geq P(A|I)$ since $P(B|I) \leq 1$□

weaker syllogism: $J =$ "$B \Rightarrow A$ more plausible", $P(A|BJ) > P(A|J)$

claim: $J \Rightarrow$ "A $\Rightarrow B$ more plausible", $P(B|AJ) > P(B|J)$

proof: $P(B|AJ) = \underbrace{\frac{P(A|BJ)}{P(A|J)}}_{>1} P(B|J) > P(B|J)$ □

1.5.2 Assigning Probabilities

I background information, $A_1, \ldots A_n$ mutually exclusive, exhausting I
$\Rightarrow$ "one and only one A_i with $i \in \{1, \ldots \text{ n}\}$ is true", $\sum_{i=1}^{n} P(A_i|I) = 1$

If knowledge in I about $A_1, \ldots A_n$ is symmetric $\Rightarrow P(A_i|I) = P(A_j|I)$

$$\text{uniform probability distribution:} \quad P(A_i|B) \quad = \frac{1}{n}$$

Laplace's principle of the insufficient reason

Canonical examples:

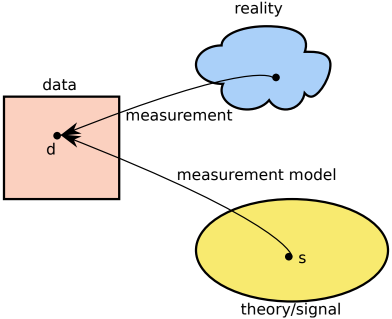
- $P(\boxdot \mid \text{fair die}) = \frac{1}{6}$
- $P(\boxdot \mid \text{loaded die}) = \frac{1}{6}$
- $P(\boxdot \mid$ previous results, loaded die) may differ from $^1/_6$

$\Rightarrow$ Conditional probabilities describe learning from data.

1.6 Statistical Inference

1.6.1 Measurement process

$$\text{physical state} \underset{\text{inference}}{\overset{\text{measurement}}{\rightleftarrows}} \text{resulting data}$$



Potential problems:

- Theory incorrect.
- Theory insufficient for reality.
- Data is not uniquely determined, $P(d|s) \neq \delta(d - R(s))$.
- Signal is not uniquely determined, $P(s|d) \neq \delta(s - s^*(d))$.

1.6.2 Bayesian Inference

I= background information: on signal s, on measurement yielding data d
I assumed impicitly in the following, $P(s) := P(s|I)$ etc.

$$\text{Bayes' theorem: } P(s|d) = \frac{P(d,\ s)}{P(d)} = \frac{P(d|s)}{P(d)} P(s)$$

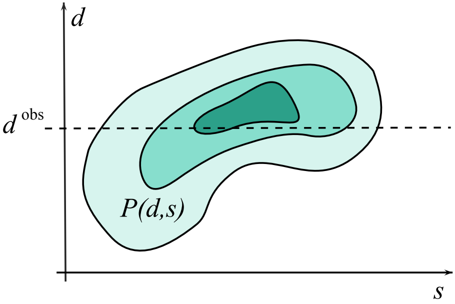
Sloppy notation: $P(s) = P(s_{\text{var}} = s_{\text{val}}|I)$, s_{var} unknown variable, s_{val} concrete value

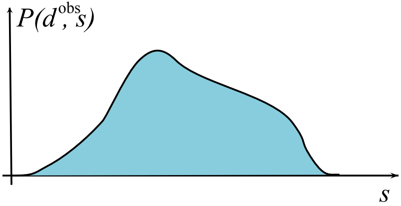
Observations:

- Joint probability $P(d,\ s)$ decomposed in likelihood and prior
- Prior $P(s)$ summarizes knowledge on s prior to measurment
- Likelihood $P(d|s)$ describes measurement process, updates prior, $P(s) \xrightarrow{P(d|s)} P(s|d)$
- Evidence $P(d) = \sum_s P(d,\ s)$ normalizes posterior

$$\sum_s P(s|d) = \sum_s \frac{P(d,s)}{P(d)} = \frac{\sum_s P(d,s)}{\sum_{s'} P(d,s')} = 1$$

Picturing Bayesian Inference





Observations:

- After measurement only hyperplane $d = d^{\text{obs}}$ relevant
- Any deduction relying on unobserved data $d^{\text{mock}} \neq d^{\text{obs}}$ is suboptimal, inconsistent, or just wrong
- Normalization of restricted probability $P(d = d^{\text{obs}}, s)$ by area under curve: $\sum_s P(d^{\text{obs}}, s) = P(d^{\text{obs}})$

1.7 Coin tossing

1.7.1 Recognizing the unfair coin

I_1 = "Outcome of coin tosses stored in data $d = (d_1, d_2, \ldots)$,
$d_i \in \{\text{head}, \text{tail}\} := \{1, 0\}$ of i^{th} toss, $d^{(n)} = (d_1, \ldots d_n)$ = data up to toss n"

Question 1: What is our knowledge on $d^{(1)} = (d_1)$ given I_1?
Due to symmetry in knowledge: $P(d_1 = 0|I_1) = P(d_1 = 1|I_1) = 1/2$

Question 2: What is our knowledge about d_{n+1} given $d^{(n)}$, I_1?

$$P(d_{n+1}|d^{(n)}, I_1) = \frac{P(d^{(n+1)}|I_1)}{P(d^{(n)}|I_1)} \text{ with } d^{(n+1)} = (d_{n+1}, d^{(n+1)})$$

I_1 symmetric w.r.t. 2^n possible sequences $d^{(n)} \in \{0, 1\}^n$ of length $n \Rightarrow P(d^{(n)}|I_1) = 2^{-n}$

$$P(d_{n+1}|d^{(n)}, I_1) = \frac{2^{-n-1}}{2^{-n}} = \frac{1}{2}$$

Statistical Independence

Given I_1, the data $d^{(n)}$ contains no useful information on d_{n+1}. What did we miss?
It seems $I_1 \Rightarrow$ "All tosses are statistically independent of each other."

A and B **statistically independent** under $C \Leftrightarrow P(A|BC) = P(A|C)$
$\Rightarrow P(AB|C) = P(A|BC)\,P(B|C) = P(A|C)\,P(B|C)$

Additional information $I_2 =$ "Tosses done with same coin, which might be loaded, meaning heads occur with frequency f"
$\exists f \in [0,1] : \forall i \in \mathbb{N} : \; P(d_i = 1|f, I_1, I_2) = f, I = I_1 I_2$

$$P(d_i|f,\, I) = \begin{cases} f & d_i = 1 \\ 1-f & d_i = 0 \end{cases} = f^{d_i}\,(1-f)^{1-d_i}$$

1.7.2 Probability Density Functions

Question 3: What do we know about f given I and our data $d^{(n)}$ after n tosses?
f is a continuous parameter!

Notation: $P(f \in F|I)$ with $F \subset \Omega$. In the above case $\Omega = [0, 1]$
$P(f \in F|I)$ must increase monotonically with $|F| = \int_F df\, 1$ until $P(f \in \Omega|I) = 1$
If I symmetric for $\forall f \in \Omega$ we request

$$P(f \in F|I) := \frac{|F|}{|\Omega|} = \frac{\int_F df\, 1}{\int_\Omega df\, 1}$$

If I implies weights $w : \Omega \mapsto \mathbb{R}_0^+$, we use $|F|_w := \int_F df\, w(f)$

$$P(f \in F|I) := \frac{|F|_w}{|\Omega|_w} = \frac{\int_F df\, w(f)}{\int_\Omega df\, w(f)} =: \int_F df\, \mathcal{P}(f|\, I)$$

$\mathcal{P}(f|\, I) := w(f)/|\Omega|_w$ is called **probability density function** (PDF)

Normalization of PDFs

Normalization:

$$P(f \in \Omega | I) = \int_\Omega df\, \mathcal{P}(f|I) = \int_\Omega df\, \frac{w(f)}{|\Omega|_w} = \frac{|\Omega|_w}{|\Omega|_w} = 1$$

Coordinate transformation: $T: F \mapsto F',\, T^{-1}: F' \mapsto F$ with $F' = T(F)$

Coordinate in-variance of probabilities: $P(f \in F|I) = P(f' \in F'|I)$ with $f' = T(f)$

$$\begin{aligned} \Rightarrow \int_F df\, \mathcal{P}(f|I) &= \int_{F'} df'\, \mathcal{P}(f'|I) \text{ for } \forall F \subset \Omega \\ \Rightarrow \mathcal{P}(f'|I) &= \mathcal{P}(f|I) \left\| \frac{df}{df'} \right\|_{f=T^{-1}(f')} \end{aligned}$$

PDF are not coordinate invariant!

Bayes Theorem for PDFs

Joint PDFs: $\mathcal{P}(x, y|I)$ joint PDF of $x \in \mathbb{R}$ and $y \in \mathbb{R}$, *i.e.*

$$P(x \in X,\ y \in Y|I) := \int_X dx \int_Y dy\, \mathcal{P}(x, y|I) \text{ for } \forall\, X, Y \subset \mathbb{R}$$

Marginal PDF: $\mathcal{P}(x|I) := \int dy\, \mathcal{P}(x, y|I) \qquad \mathcal{P}(y|I) := \int dx\, \mathcal{P}(x, y|I)$

Conditional PDF: $\mathcal{P}(x|y, I) := \dfrac{\mathcal{P}(x, y|I)}{\mathcal{P}(y|I)} \qquad \mathcal{P}(y|x, I) := \dfrac{\mathcal{P}(x, y|I)}{\mathcal{P}(x|I)}$

$\Rightarrow$ product rule for PDFs: $\quad \mathcal{P}(x, y|I) = \mathcal{P}(x|y, I)\mathcal{P}(y|I) = \mathcal{P}(y|x, I)\mathcal{P}(x|I)$

$\Rightarrow$ Bayes theorem for PDFs: $\quad \mathcal{P}(y|x, I) = \dfrac{\mathcal{P}(x|y, I)\mathcal{P}(y|I)}{\mathcal{P}(x|I)}$

To be shown: quantities defined above are indeed PDFs

Marginal & Conditional PDFs

Marginalized PDF:

$$\begin{aligned} P(x \in X|I) &\overset{?}{=} \int_X dx\, \mathcal{P}(x|I) = \int_X dx \int_{\mathbb{R}} dy\, \mathcal{P}(x,y|I) \\ &= P(x \in X, y \in \mathbb{R}|I) = P(x \in X|I) \end{aligned}$$

as $I \Rightarrow y \in \mathbb{R}$, similarly, $P(y \in Y|I) = \int_Y dy\, \mathcal{P}(y|I)$. $\square$

Conditional PDF: *e.g.* for x conditioned on y ($y_{\text{var}} = y_{\text{val}}$)

$$P(x \in X|y,I) \overset{?}{=} \int_X dx\, \mathcal{P}(x|y,I) = \int_X dx\, \frac{\mathcal{P}(x,y|I)}{\mathcal{P}(y|I)} = \frac{\int_X dx\, \mathcal{P}(x,y|I)}{\int_{\mathbb{R}} dx\, \mathcal{P}(x,y|I)} = \frac{|X|_{\mathcal{P}(x,y|I)}}{|\mathbb{R}|_{\mathcal{P}(x,y|I)}}$$

is ratio of weighted measures, as used to define PDFs. $\square$

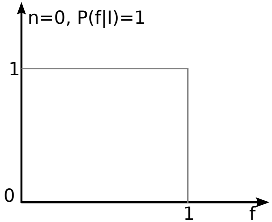
PDF $\mathcal{P}(x,y)$ uniquely defines probabilities $P(x \in X, y \in Y)$, but reverse is not true.

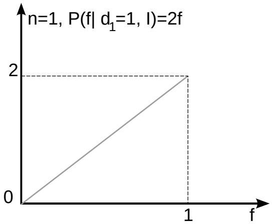
1.7.3 Infering the coin load

Question 3: What do we know about f given I and our data $d^{(n)}$ after n tosses?

$$n = 0: \ \mathcal{P}(f|I) = 1$$

$$n = 1: \ \mathcal{P}(f|d = (1), I) = \frac{\mathcal{P}(d_1 = 1|f, I)\mathcal{P}(f|I)}{\int_0^1 df\, \mathcal{P}(d_1 = 1|f, I)\mathcal{P}(f|I)} = \frac{f \times 1}{\int_0^1 df\, f} = \frac{f}{1/2} = 2f$$





Several Tosses

$$\mathcal{P}(f|d^{(n)},I) = \frac{\mathcal{P}(d^{(n)}|f,I)\,\mathcal{P}(f,I)}{\mathcal{P}(d^{(n)}|I)} = \frac{\mathcal{P}(d^{(n)},f|I)}{\mathcal{P}(d^{(n)}|I)}$$

$$\mathcal{P}(d^{(n)},f|I) = \prod_{i=1}^{n} \mathcal{P}(d_i|f,I) \times 1 = \prod_{i=1}^{n} f^{d_i}\,(1-f)^{1-d_i} = f^{n_1}\,(1-f)^{n_0}$$

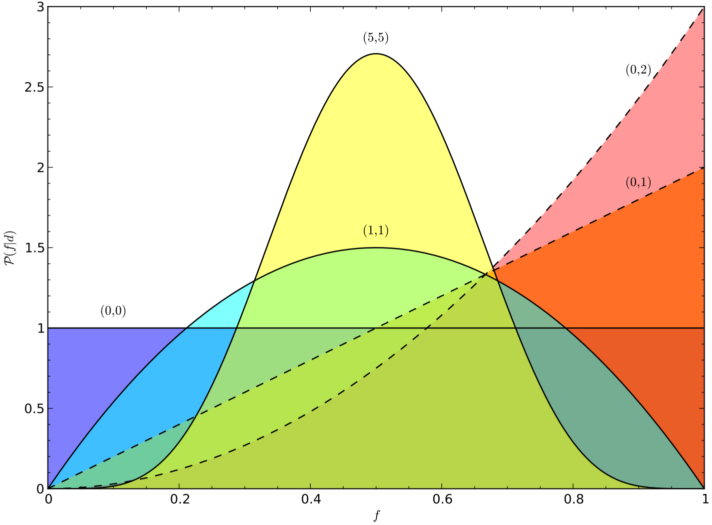
$$\#\,\text{heads} = n_1 = n_1(d^{(n)}) = \sum_{i=1}^{n} d_i, \quad \#\,\text{tails} = n_0 = n - n_1$$

$$\mathcal{P}(d^{(n)}|I) = \int_0^1 df\,\mathcal{P}(d^{(n)},f|I) = \int_0^1 df\,f^{n_1}(1-f)^{n_0} = \mathcal{B}(n_0+1,n_1+1) = \frac{n_0!\,n_1!}{(n+1)!}$$

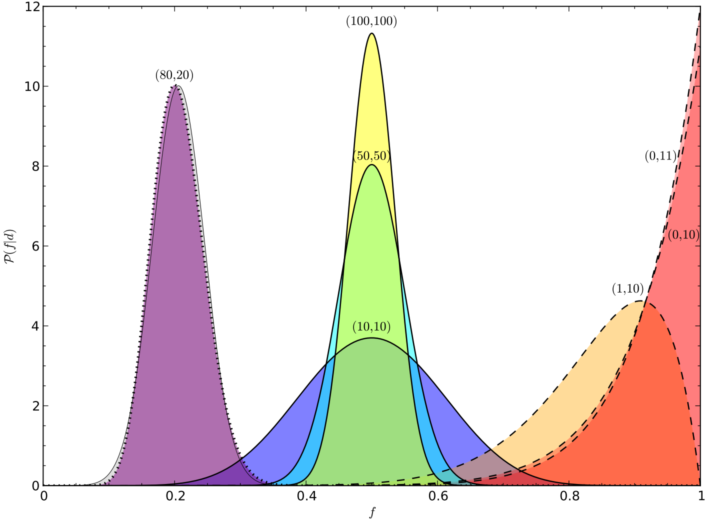
$$\mathcal{B}(a,b) = \int_0^1 dx\,x^{a-1}(1-x)^{b-1} = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \stackrel{a,b\in\mathbb{N}}{=} \frac{(a-1)!\,(b-1)!}{(a+b-1)!} \quad \text{Beta function}$$

$$\boxed{\mathcal{P}(f|d^{(n)},I) = \frac{P(d^{(n)},f|I)}{P(d^{(n)}|I)} = \frac{(n+1)!}{n_1!\,n_0!} f^{n_1}(1-f)^{n_0}}$$

Load Posterior $\mathcal{P}(f|(n_0, n_1), I)$ for Few Tosses



Load Posterior $\mathcal{P}(f|(n_0, n_1), I)$ for Many Tosses



Laplace's rule of succession

Question 2: What is our knowledge about d_{n+1} given $d^{(n)}$, $I = I_1 I_2$?

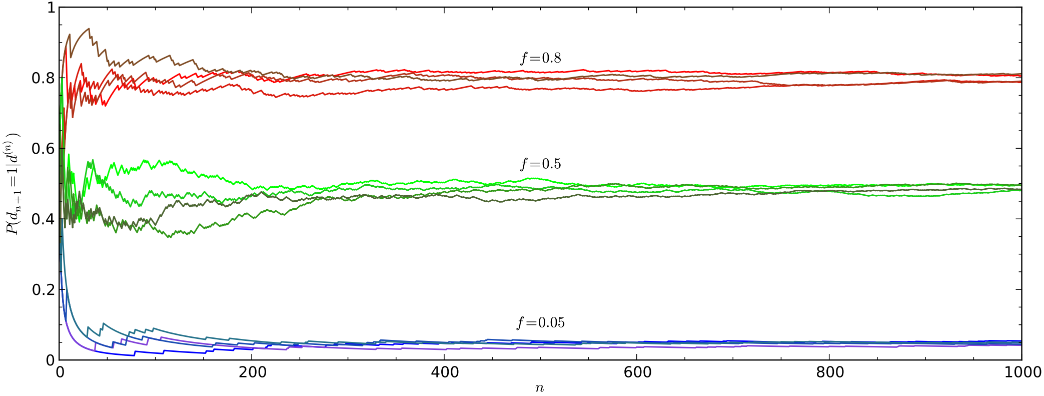
$$
\begin{aligned}
P(d_{n+1} = 1|d^{(n)}, I) &= \int_0^1 df\, P(d_{n+1} = 1, f|\, d^{(n)}, I) \\
&= \int_0^1 df\, P(d_{n+1} = 1|f,\, d^{(n)}, I)\, \mathcal{P}(f|d^{(n)}, I) \\
&= \int_0^1 df\, f\, \mathcal{P}(f|d^{(n)}, I) =: \langle f \rangle_{(f|d^{(n)}, I)} \\
&= \frac{(n+1)!}{n_1!\, n_0!} \int_0^1 df\, f^{n_1+1} (1-f)^{n_0} \\
&= \frac{(n+1)!}{n_1!\, n_0!} \frac{(n_1+1)!\, n_0!}{(n+2)!} = \frac{n_1+1}{n+2}
\end{aligned}
$$

$$
P(d_{n+1} = 1|d^{(n)}, I) = \langle f \rangle_{(f|d^{(n)}, I)} = \frac{n_1+1}{n+2} \neq \frac{n_1}{n}
$$

Laplace's rule can save your life!

$$
P(d_{n+1} = 0|d^{(n)}, I) = \langle 1-f \rangle_{(f|d^{(n)}, I)} = \frac{n_0+1}{n+2}
$$

Learning Sequence



1.7.4 Large Number of Tosses

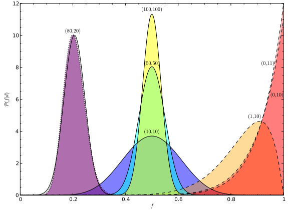
Central limit theorem: $\mathcal{P}(f|d^{(n)})$ becomes Gaussian for $n_0, n_1 \gg 1$

$$\mathcal{P}(f|d^{(n)}, I) \approx \mathcal{G}(f-\bar{f}, \sigma_f^2) = \frac{1}{\sqrt{2\pi\sigma_f^2}} \exp\left(-\frac{(f-\bar{f})^2}{2\sigma_f^2}\right)$$

$$\text{Mean: } \bar{f} = \langle f \rangle_{(f|d^{(n)}, I)} = \frac{n_1+1}{n+2}$$

$$\text{Variance: } \sigma_f^2 = \langle (f-\bar{f})^2 \rangle_{(f|d^{(n)})} = \langle f^2 - 2\bar{f}f + \bar{f}^2 \rangle_{(f|d^{(n)})} = \langle f^2 \rangle_{(f|d^{(n)})} - \bar{f}^2$$

$$= \frac{(n_1+2)(n_1+1)}{(n+3)(n+2)} - \left(\frac{n_1+1}{n+2}\right)^2 = \frac{\bar{f}(1-\bar{f})}{n+3} \sim \frac{1}{n}$$



Gaussian approx. needs $f, \bar{f}$ to be away from 0 and 1

1.7.5 The Evidence for the Load

hypotheses: $I =$ "loaded coin, $f \in [0,1] \backslash \{\frac{1}{2}\}$" , $J =$ "a fair coin, $f = \frac{1}{2}$", $M = I + J$
hyper-priors for hypotheses: $P(I|M) = P(J|M) = 1/2$

$$\text{a posteriori odds: } O(d^{(n)}) := \frac{P(I|d^{(n)}, M)}{P(J|d^{(n)}, M)} = \frac{P(d^{(n)}|I, M)\, P(I|M)/P(d^{(n)}|M)}{P(d^{(n)}|J\, M)\, P(J|M)/P(d^{(n)}|M)}$$

$$\text{loaded coin evidence: } P(d^{(n)}|I) = \frac{n_1!\, n_0!}{(n+1)!}$$

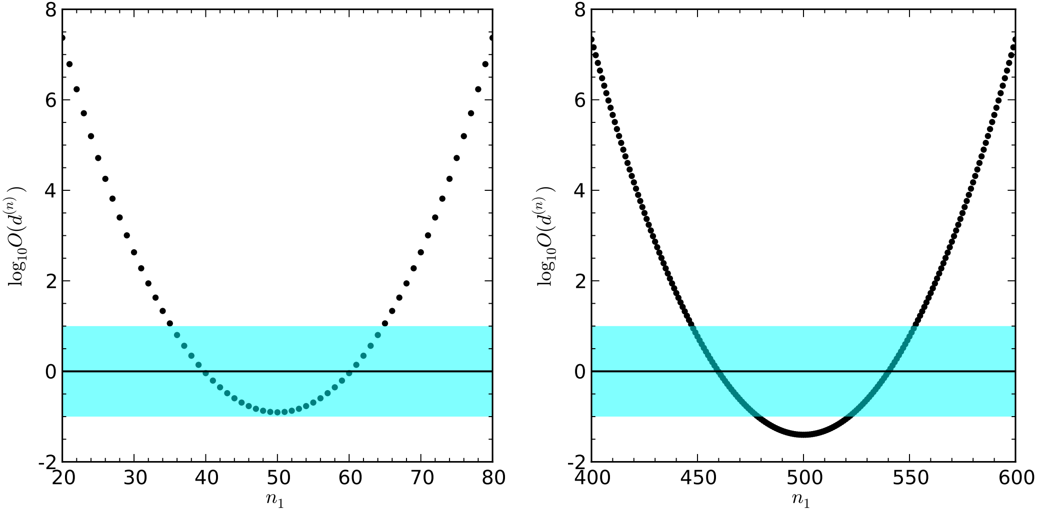
$$\text{fair coin evidence: } P(d^{(n)}|J) = \frac{1}{2^n}$$

$$O(d^{(n)}) = \frac{2^n\, n_1!\, n_0!}{(n+1)!}$$

Only heads:

$n_1 = n$	0	1	2	3	4	5	6	7	8	9	10	100	1000
$O(d^{(n)})$	1	1	$4/3$	2	$3\,1/5$	$5\,1/3$	$9\,1/7$	16	$28\,4/9$	$51\,1/5$	$93\,1/11$	$10^{28.1}$	10^{298}

Load Odds for $n = 100, 1000$



1.7.6 Lessons Learned

1. **Probabilities** described knowledge states
2. **Frequencies** are probabilities if known, $P(d = 1|f, I) = f$
3. **Joint probability** contain all relevant information
4. **Posterior** summarizes knowledge of signal given data and model knowledge
5. **Evidence:** Signal-marginalized joint probability, "likelihood" for model
6. **Background information matters:** $P(d_{n+1}|d^{(n)}, I_1) \neq P(d_{n+1}|d^{(n)}, I_1 I_2)$, if $I_2 \not\subseteq I_1$
7. **Intelligence needs models:** coins having a constant head frequency f
8. **Probability Density Functions** (PDFs) serve to construct probabilies
9. **Learning & forgetting:** Posterior changes with new data, usually sharpens thereby
10. **Sufficient statistics** are compressed data, giving the same information as original data on the quantity of interest, e.g. $P(f|d^{(n)}, I) = P(f|(n_0, n_1), I)$
11. **Nested models** contain each other; fair coin model is included in unfair coin model
12. **Occam's razor:** *Among competing hypotheses, the one with the fewest assumptions should be selected.*
13. **Uncertainty** of an inferred quantity may depend on data realization

1.8 Adaptive Information Retrieval

1.8.1 Inference from adaptive data retrieval

Data $d^{(n)} = (d_1, \ldots d_n)$ to infer signal s taken sequentially.
Action a_i chosen to measure d_i via $d_i \hookleftarrow P(d_i|a_i, s)$ can depend on previous data $d^{(i-1)}$ via data retrieval strategy function $A: \; d^{(i-1)} \to a_i$.

- A **predetermined strategy** is independent of the prior data: $A(d^{(i-1)}) \equiv a_i$ irrespective of $d^{(i-1)}$
- An **adaptive strategy** depends on the data: $\exists i, d^{(i-1)}, d'^{(i-1)} : \; A(d^{(i-1)}) \neq A(d'^{(i-1)})$

New datum d_i depends conditionally on previous data $d^{(i-1)}$ through strategy A,

$$P(d_i|a_i, s) = P(d_i|A(d^{(i-1)}),\, s) = P(d_i|d^{(i-1)},\, A,\, s)$$

Likelihood of the full data set $d = d^{(n)}$:

$$P(d|A,\, s) = P(d_n|d^{(n-1)},\, A,\, s) \cdots P(d^{(1)}|A,\, s) = \prod_{i=1}^{n} P(d_i|d^{(i-1)},\, A,\, s)$$

Different strategy $B \to$ different actions $b \to$ different data d'

Unknown strategy

Strategy $A \to$ actions a, data d; strategy $B \to$ actions b, data d'
predetermined strategy $B(d^{(i)}) \equiv a_i \to$ actions a, data d

$$
\begin{aligned}
\text{likelihood: } P(d|A,\, s) &= \prod_{i=1}^{n} P(d_i|A(d^{(i-1)}),\, s) = \prod_{i=1}^{n} P(d_i|a_i,\, s) \\
&= \prod_{i=1}^{n} P(d_i|B(d^{(i-1)}),\, s) = P(d|B,\, s) \\
\text{posterior: } P(s|d,\, A) &= \frac{P(d|A,\, s)P(s|A)}{P(d|A)} = \frac{P(d|A,\, s)P(s)}{P(d|A)} \\
&= \frac{P(d|A,\, s)P(s)}{\sum_s P(d|A,\, s)P(s)} = \frac{P(d|B,\, s)P(s)}{\sum_s P(d|B,\, s)P(s)} \\
&= P(s|d,\, B)
\end{aligned}
$$

Used assumption: $P(s|A) = P(s)$

Historical Inference

Why data was taken does not matter for Bayesian inference, only how and what it was. $P(s|d,A) = P(s|d,B)$, if strategies A, B provide identical actions for observed data, $A(d^{(i)}) = B(d^{(i)}) = a_i$, and if signal is independent of strategy, $P(s|A) = P(s)$.

Corollary: A **history**, a recorded sequence of interdependent observations (= actions and resulting data), is open to a Bayesian analysis without knowledge of the used strategy, but nearly useless for frequentists analysis as alternative realities are not available.

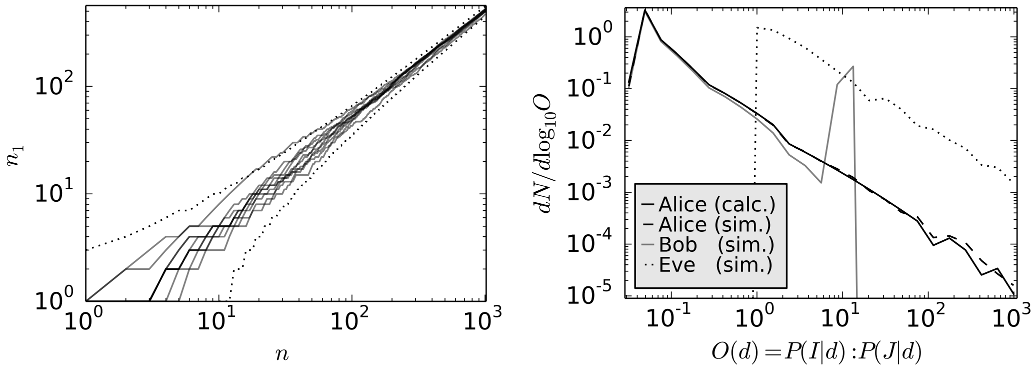
1.8.2 Adaptive Strategy to Maximize False Evidence

Can strategy choice create spurious evidence favouring false hypothesis I over right one J ?

$$\begin{aligned} \text{odds: } O(d) &= \frac{P(I|d)}{P(J|d)} = \frac{P(d|I)P(I)}{P(d|J)P(J)} \\ \text{expected odds: } \langle O(d)\rangle_{(d|J,A)} &= \sum_d P(d|A,J)\,O(d) = \sum_d P(d|A,J)\frac{P(d|A,I)\,P(I)}{P(d|A,J)\,P(J)} \\ &= \frac{P(I)}{P(J)} \underbrace{\sum_d P(d|A,I)}_{=1} = \frac{P(I)}{P(J)} = \text{prior odds, independent of } A \end{aligned}$$

Tuning of strategy can not create expected odds mass $\langle O(d)\rangle_{(d|J)}$ in favor of wrong hypothesis I, only redistribute it. Odds mass for right hypothesis J can be tuned, as $\left\langle \frac{1}{O(d)} \right\rangle_{(d|J,A)} = \left\langle \frac{P(J|d,A)}{P(I|d,A)} \right\rangle_{(d|J,A)} \geq \frac{P(J)}{P(I)}$ (nice exercise).

Adaptive Coin Tossing



Alice: serious scientist – predetermined sequence of $n = 1000$ tosses

Bob: ambitious scientist – stops when $O = \frac{P(I|d^{(n)})}{P(J|d^{(n)})} > 10$ or $n = 1000$

Eve: evil scientists – makes 1000 tosses and picks n retrospective without reporting this

End