Outline

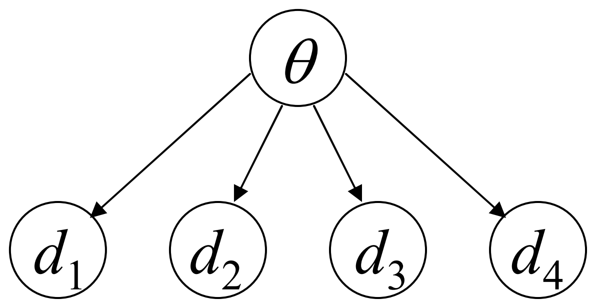
- Bayesian parameter estimation
- Hierarchical Bayesian models
- Metropolis-Hastings
  - A more general approach to MCMC

Coin flipping

- Comparing two simple hypotheses
  - $P(\texttt{H}) = 0.5$ vs. $P(\texttt{H}) = 1.0$
- Comparing simple and complex hypotheses
  - $P(\texttt{H}) = 0.5$ vs. $P(\texttt{H}) = \theta$
- Comparing infinitely many hypotheses
  - $P(\texttt{H}) = \theta$: Infer θ

Comparing infinitely many hypotheses

- Assume data are generated from a model:



$P(\text{H}) = \theta$

- What is the value of θ?
  - each value of θ is a hypothesis H
  - requires inference over infinitely many hypotheses

Comparing infinitely many hypotheses

- Flip a coin 10 times and see 5 heads, 5 tails.
- *P*(H) on next flip? 50%
- Why? 50% = 5 / (5+5) = 5/10.
- "Future will be like the past."

- Suppose we had seen 4 heads and 6 tails.
- *P*(H) on next flip? Closer to 50% than to 40%.
- Why? Prior knowledge.

Integrating prior knowledge and data

$$P(H \mid D) = \frac{P(H)P(D \mid H)}{P(D)}$$

$$P(\theta \mid D) \propto P(D \mid \theta)\, P(\theta)$$

- Posterior distribution $P(p \mid D)$ is a probability density over $\theta = P(\mathtt{H})$
- Need to work out likelihood $P(D \mid \theta)$ and specify prior distribution $P(\theta)$

Likelihood and prior

- Likelihood:

$$P(D \mid \theta) = \theta^{N_H} (1-\theta)^{N_T}$$

  - N_H: number of heads
  - N_T: number of tails

- Prior:

$$P(\theta) \propto \quad ?$$

A simple method of specifying priors

- Imagine some fictitious trials, reflecting a set of previous experiences
  - strategy often used with neural networks or building invariance into stat. machine vision.
- e.g., F ={1000 heads, 1000 tails} ~ strong expectation that any new coin will be fair

- In fact, this is a sensible statistical idea...

Likelihood and prior

- Likelihood:

$$P(D \mid \theta) = \theta^{N_H} (1-\theta)^{N_T}$$

  - N_H: number of heads
  - N_T: number of tails
- Prior:

$$P(\theta) \propto \theta^{F_H - 1} (1-\theta)^{F_T - 1} \quad \text{Beta}(F_H, F_T)$$

  - F_H: fictitious observations of heads
  - F_T: fictitious observations of tails

Likelihood and prior

- Likelihood:

$$P(D \mid \theta) = \theta^{N_H} (1-\theta)^{N_T}$$

  - N_H: number of heads
  - N_T: number of tails
- Prior:

$$P(\theta) = \frac{\Gamma(F_H+F_T)}{\Gamma(F_H)\,\Gamma(F_T)} \theta^{F_H-1} (1-\theta)^{F_T-1}$$

  - F_H: fictitious observations of heads
  - F_T: fictitious observations of tails

Likelihood and prior

- Likelihood:

$$P(D \mid \theta) = \theta^{N_{\mathrm{H}}} (1-\theta)^{N_{\mathrm{T}}}$$

  - N_{H}: number of heads
  - N_{T}: number of tails

- Prior:

$$\int_0^1 P(\theta)\, d\theta = \int_0^1 \frac{\Gamma(F_{\mathrm{H}}+F_{\mathrm{T}})}{\Gamma(F_{\mathrm{H}})\,\Gamma(F_{\mathrm{T}})}\, \theta^{F_{\mathrm{H}}-1} (1-\theta)^{F_{\mathrm{T}}-1} d\theta = 1$$

A very useful integral

Likelihood and prior

- Likelihood:

$$P(D \mid \theta) = \theta^{N_H} (1-\theta)^{N_T}$$

  - N_H: number of heads
  - N_T: number of tails

- Prior:

$$\int_0^1 P(\theta)\, d\theta = \int_0^1 \frac{\Gamma(F_H+F_T)}{\Gamma(F_H)\,\Gamma(F_T)}\, \theta^{F_H-1} (1-\theta)^{F_T-1} d\theta = 1$$

Also useful: $\Gamma(x) = (x-1)!$

$\Gamma(x+1) = x\,\Gamma(x)$

Shape of the Beta prior

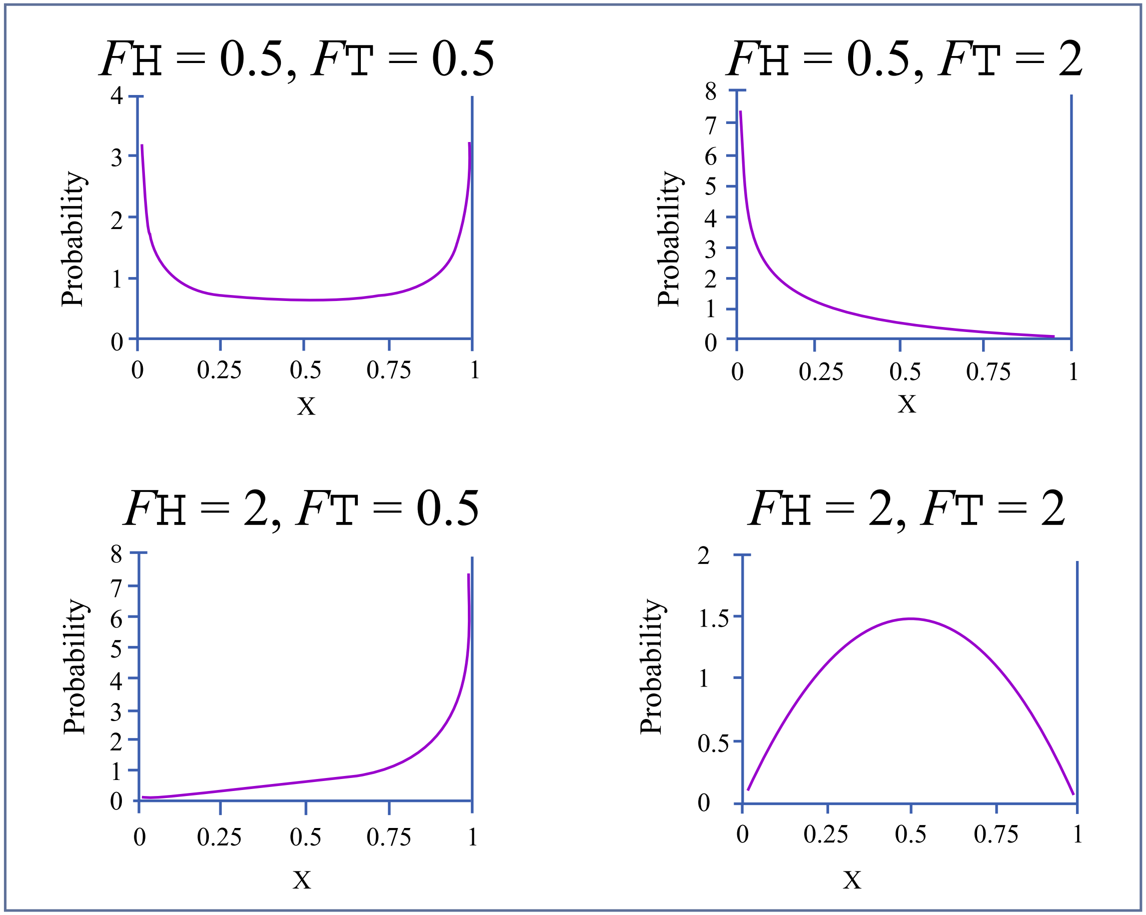


Figure by MIT OCW.

Shape of the Beta prior

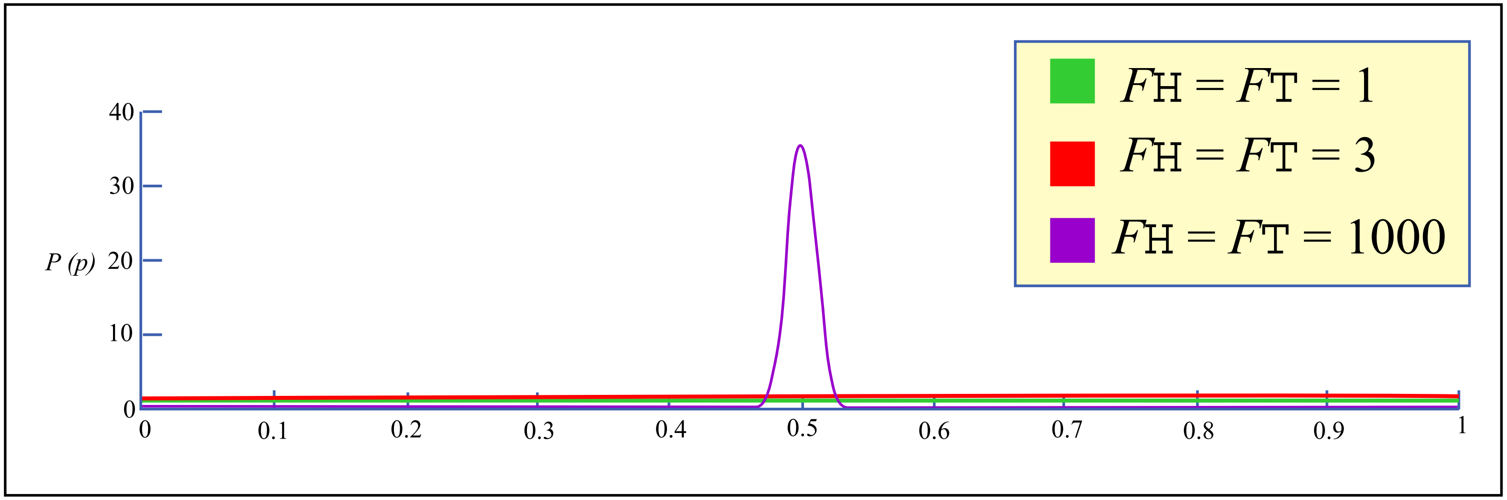


Figure by MIT OCW.

Bayesian parameter learning

- Likelihood: Bernoulli(θ)

$$P(D \mid \theta) = \theta^{N_H} (1-\theta)^{N_T}$$

  - N_H, N_T: number of heads, tails observed
- Prior: Beta(F_H,F_T)

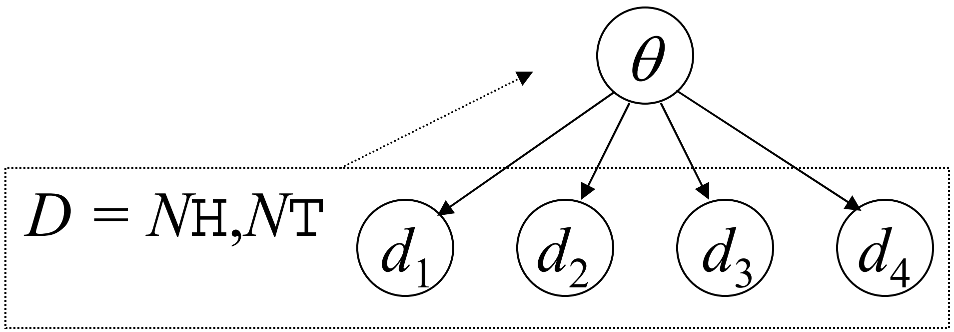
$$P(\theta) \propto \theta^{F_H - 1} (1-\theta)^{F_T - 1}$$

  - F_H, F_T: fictitious observations of heads, tails
- Posterior: Beta(N_H+F_H, N_T+F_T)

$$P(\theta \mid \mathrm{D}) \propto \theta^{N_H + F_H - 1} (1-\theta)^{N_T + F_T - 1}$$

$$= \frac{\Gamma(N_H + F_H + N_T + F_T)}{\Gamma(N_H + F_H)\,\Gamma(N_T + F_T)}\ \theta^{N_H + F_H - 1} (1-\theta)^{N_T + F_T - 1}$$

Bayesian parameter learning

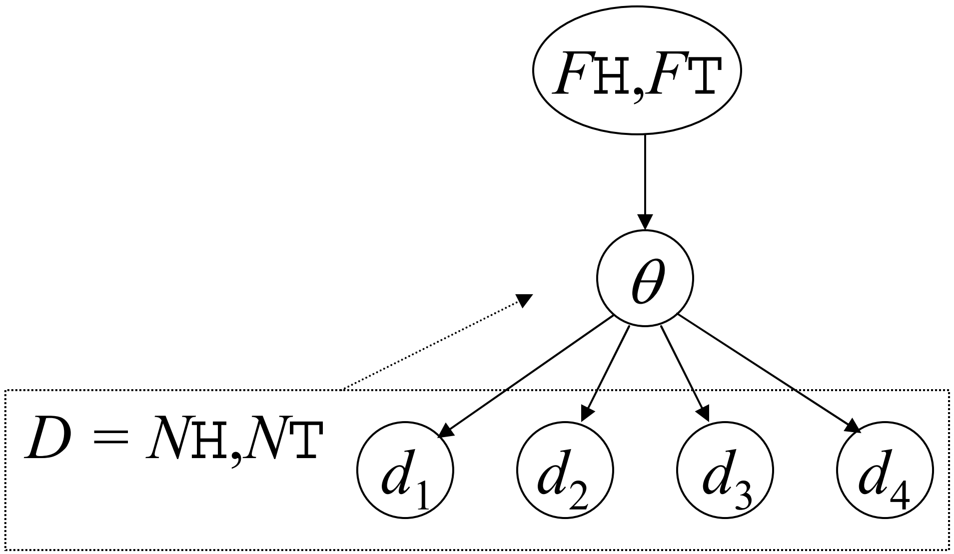


- Likelihood: Bernoulli(θ)

$$P(D \mid \theta) = \theta^{N_H} (1-\theta)^{N_T}$$

  – N_H, N_T: number of heads, tails observed

Bayesian parameter learning

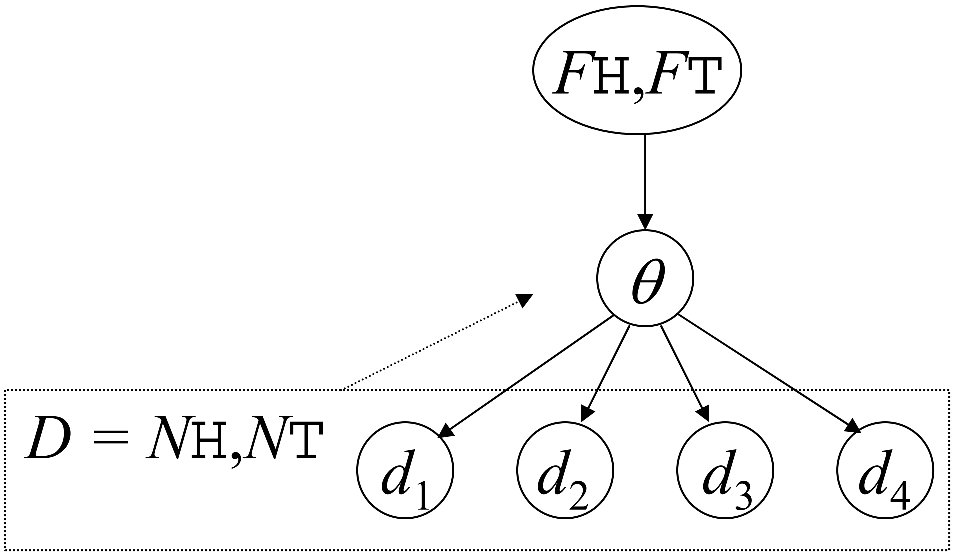


- Prior: Beta(*F*H,*F*T)

$$P(\theta \mid F\mathrm{H}, F\mathrm{T}) \propto \theta^{F\mathrm{H}-1} (1-\theta)^{F\mathrm{T}-1}$$

  - *F*H, *F*T: fictitious observations of heads, tails

Bayesian parameter learning



- Posterior: Beta(N_H+F_H, N_T+F_T)

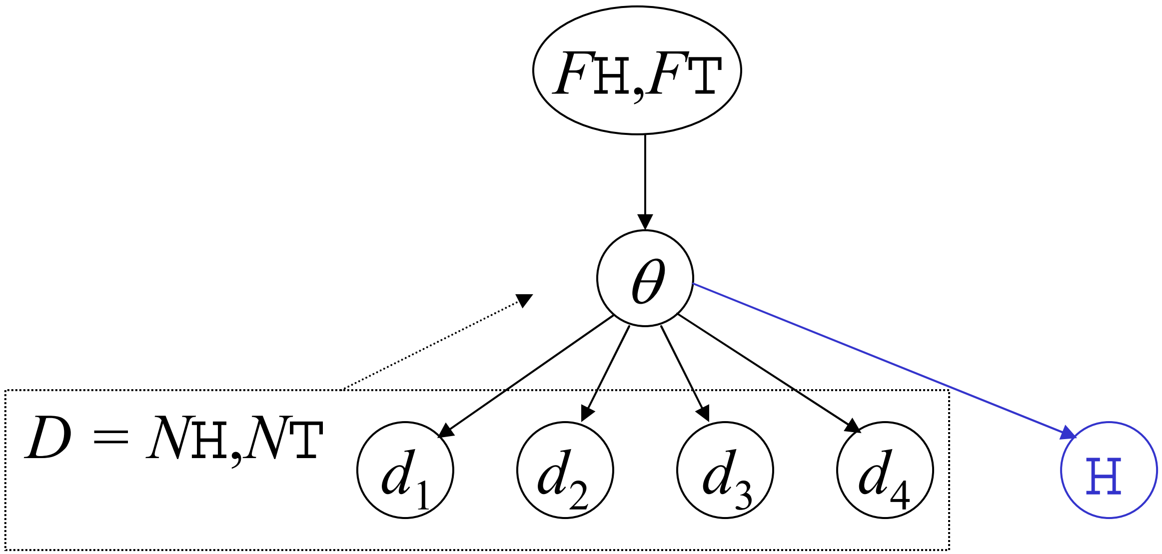
$$P(\theta \mid D, F_H, F_T) \propto \theta^{N_H+F_H-1} (1-\theta)^{N_T+F_T-1}$$

$$= \frac{\Gamma(N_H+F_H+N_T+F_T)}{\Gamma(N_H+F_H)\,\Gamma(N_T+F_T)} \theta^{N_H+F_H-1} (1-\theta)^{N_T+F_T-1}$$

Conjugate priors

- A prior $p(\theta)$ is *conjugate* to a likelihood function $p(D \mid \theta)$ if the posterior has the same functional form of the prior.
  - Different parameter values in the prior and posterior reflect the impact of observed data.
  - Parameter values in the prior can be thought of as a summary of “fictitious observations”.
- Exist for many standard distributions
  - all exponential family models
  - e.g., Beta is conjugate to Bernoulli (coin-flipping)

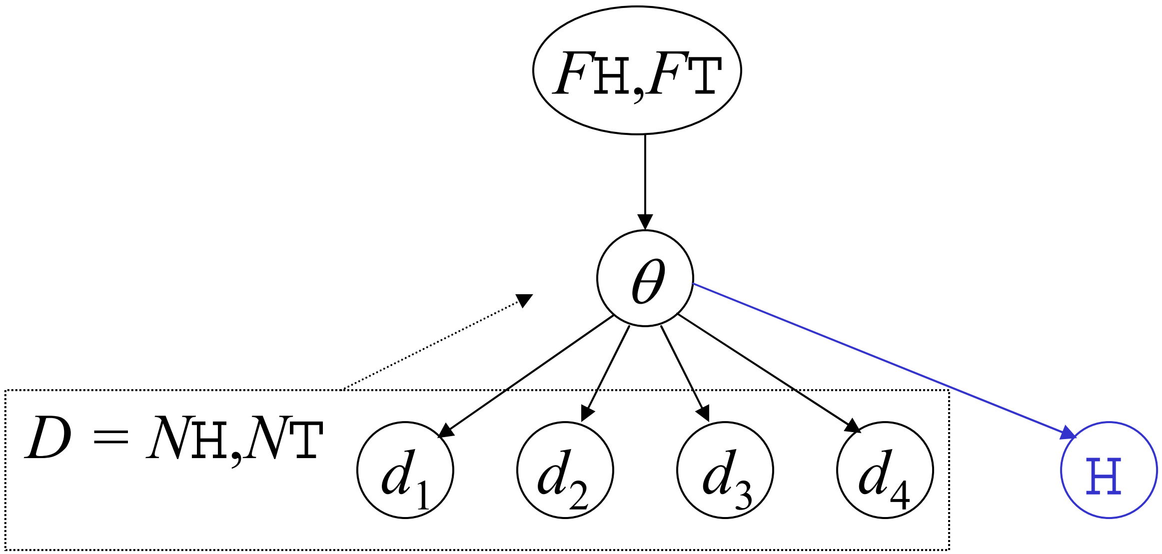
Bayesian parameter learning



- Posterior predictive distribution:

$$P(\mathtt{H} \mid D = N_{\mathtt{H}}, N_{\mathtt{T}}; F_{\mathtt{H}}, F_{\mathtt{T}}) = \, ?$$

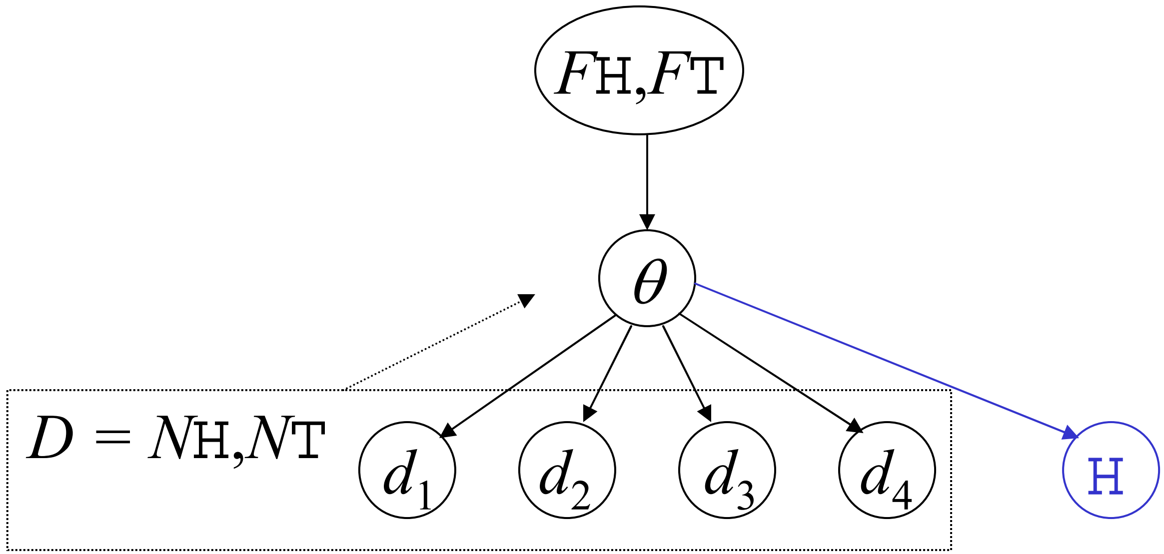
Bayesian parameter learning



- Posterior predictive distribution:

$$P(\mathrm{H} \mid D, F_\mathrm{H}, F_\mathrm{T}) = \int_0^1 P(\mathrm{H} \mid \theta)\, P(\theta \mid D, F_\mathrm{H}, F_\mathrm{T})\, d\theta$$

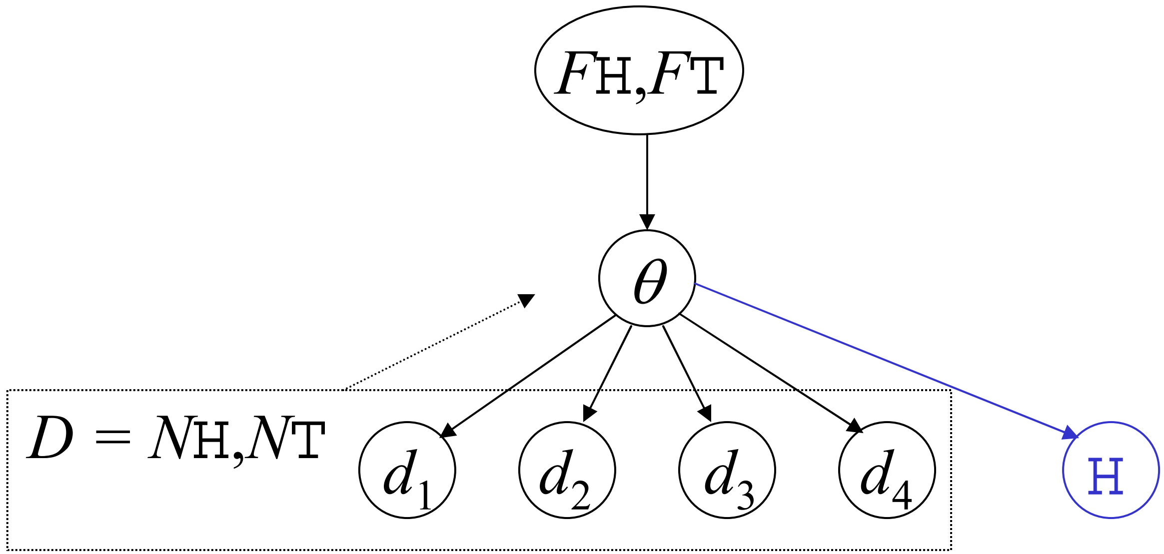
Bayesian parameter learning



- Posterior predictive distribution:

$$P(\mathtt{H} \mid D, F\mathtt{H}, F\mathtt{T}) = \int_0^1 \theta \; P(\theta \mid D, F\mathtt{H}, F\mathtt{T}) \, d\theta$$

Bayesian parameter learning

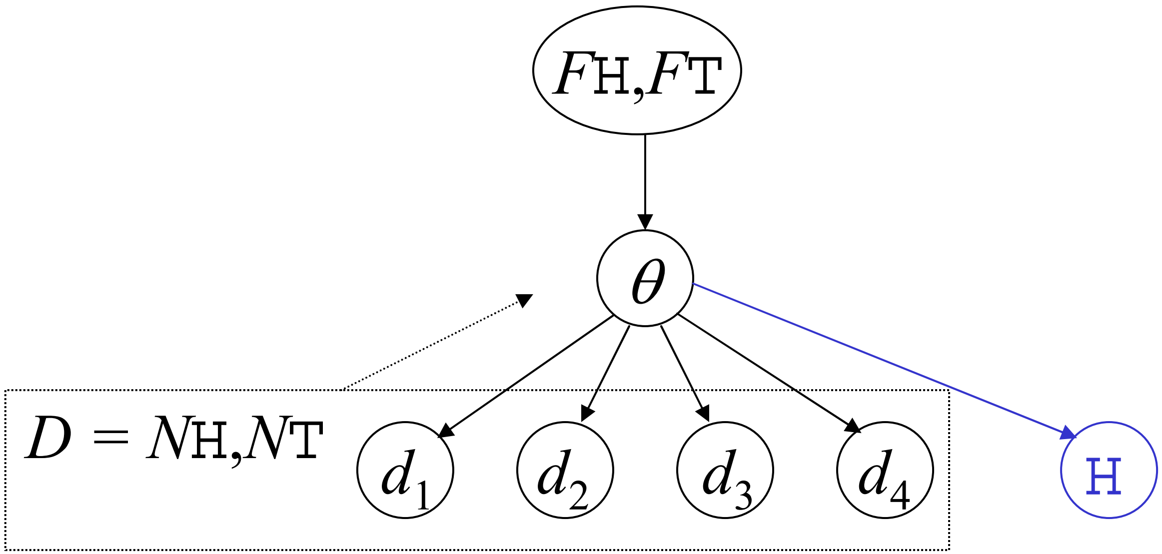


- Posterior predictive distribution:

$$P(\mathrm{H} \mid D, F\mathrm{H}, F\mathrm{T}) =$$

$$\int_0^1 \theta \, \frac{\Gamma(N\mathrm{H}+F\mathrm{H}+N\mathrm{T}+F\mathrm{T})}{\Gamma(N\mathrm{H}+F\mathrm{H})\,\Gamma(N\mathrm{T}+F\mathrm{T})} \, \theta^{N\mathrm{H}+F\mathrm{H}-1} (1-\theta)^{N\mathrm{T}+F\mathrm{T}-1} \, d\theta$$

Bayesian parameter learning

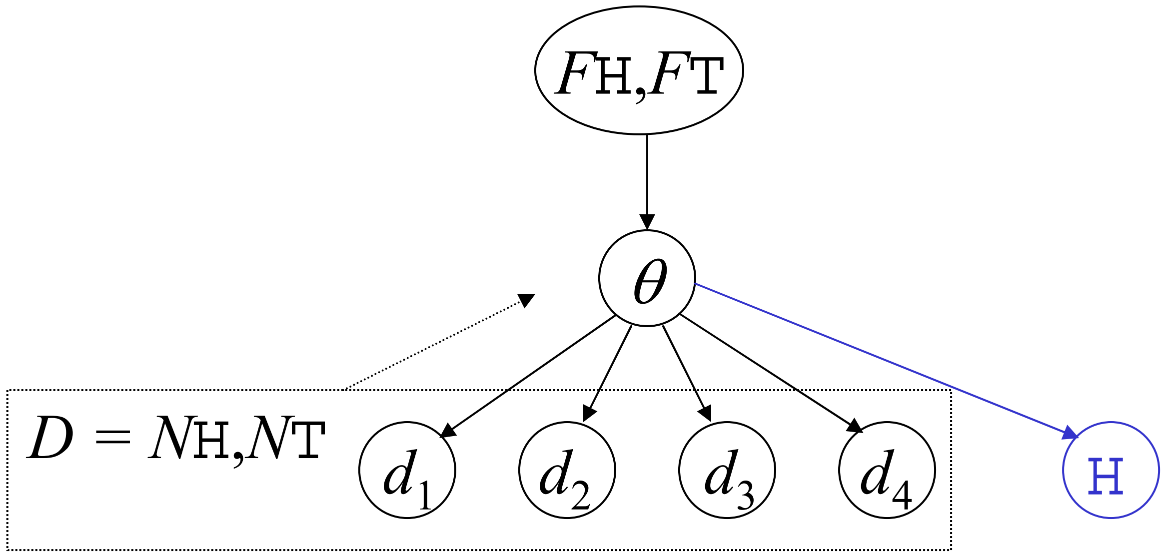


- Posterior predictive distribution:

$$P(\mathrm{H} \mid D, F\mathrm{H}, F\mathrm{T}) =$$

$$\frac{\Gamma(N\mathrm{H}+F\mathrm{H}+N\mathrm{T}+F\mathrm{T})}{\Gamma(N\mathrm{H}+F\mathrm{H})\,\Gamma(N\mathrm{T}+F\mathrm{T})} \int_0^1 \theta\, \theta^{N\mathrm{H}+F\mathrm{H}-1} (1-\theta)^{N\mathrm{T}+F\mathrm{T}-1}\, d\theta$$

Bayesian parameter learning

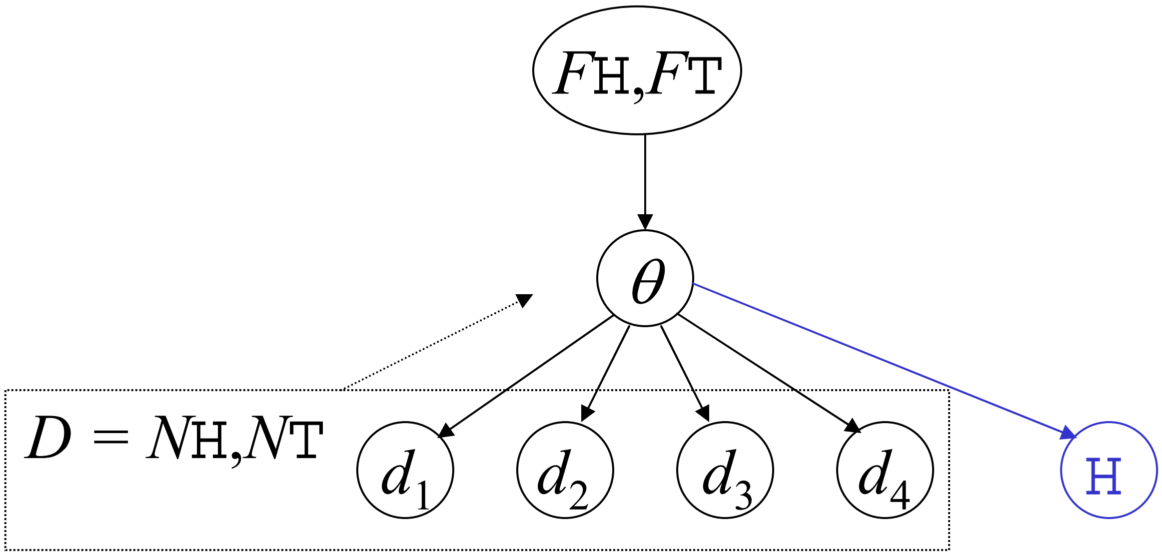


- Posterior predictive distribution:

$$P(\mathrm{H} \mid D, F_\mathrm{H}, F_\mathrm{T}) =$$

$$\frac{\Gamma(N_\mathrm{H}+F_\mathrm{H}+N_\mathrm{T}+F_\mathrm{T})}{\Gamma(N_\mathrm{H}+F_\mathrm{H})\,\Gamma(N_\mathrm{T}+F_\mathrm{T})} \int_0^1 \theta^{N_\mathrm{H}+F_\mathrm{H}} (1-\theta)^{N_\mathrm{T}+F_\mathrm{T}-1}\, d\theta$$

Bayesian parameter learning



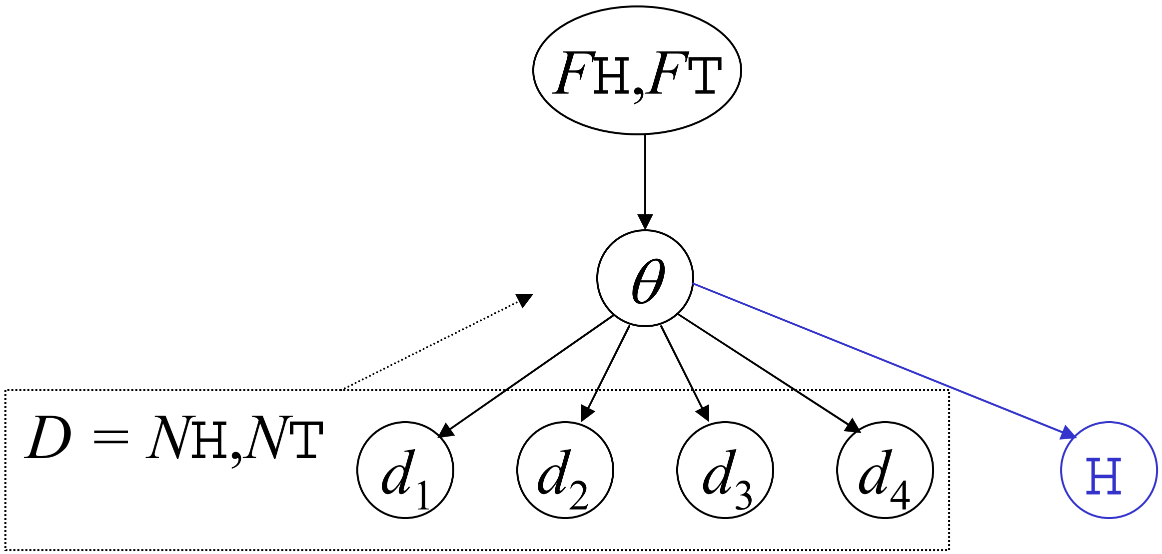
- Posterior predictive distribution:

$$P(H \mid D, F_H, F_T) =$$

$$\Gamma(x+1) = x\,\Gamma(x)$$

$$\frac{\Gamma(N_H+F_H+N_T+F_T)}{\Gamma(N_H+F_H)\,\Gamma(N_T+F_T)} \times \frac{\Gamma(N_H+F_H+1)\,\Gamma(N_T+F_T)}{\Gamma(N_H+F_H+N_T+F_T+1)}$$

Bayesian parameter learning

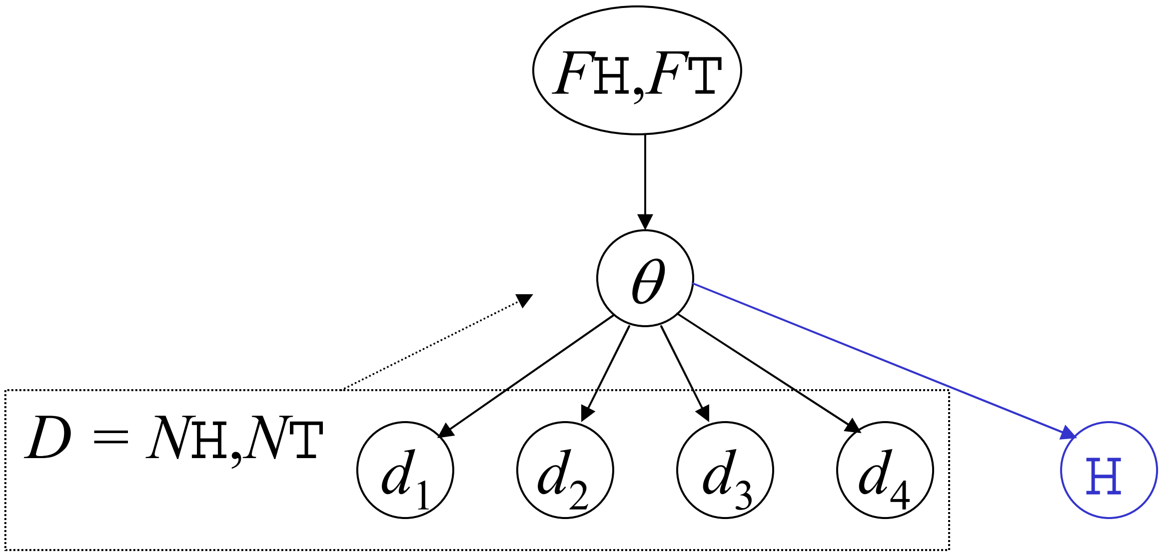


- Posterior predictive distribution:

$$P(\mathtt{H} \mid D, F\mathtt{H}, F\mathtt{T}) =$$

$$\frac{\Gamma(N\mathtt{H}+F\mathtt{H}+N\mathtt{T}+F\mathtt{T})}{\Gamma(N\mathtt{H}+F\mathtt{H})\,\Gamma(N\mathtt{T}+F\mathtt{T})} \times \frac{(N\mathtt{H}+F\mathtt{H})\,\Gamma(N\mathtt{H}+F\mathtt{H})\,\Gamma(N\mathtt{T}+F\mathtt{T})}{(N\mathtt{H}+F\mathtt{H}+N\mathtt{T}+F\mathtt{T})\,\Gamma(N\mathtt{H}+F\mathtt{H}+N\mathtt{T}+F\mathtt{T})}$$

Bayesian parameter learning



- Posterior predictive distribution:

$$P(\mathrm{H} \mid D, F\mathrm{H}, F\mathrm{T}) = \frac{(N\mathrm{H}+F\mathrm{H})}{(N\mathrm{H}+F\mathrm{H}+N\mathrm{T}+F\mathrm{T})}$$

Some examples

- e.g., F ={1000 heads, 1000 tails} ~ strong expectation that any new coin will be fair
- After seeing 4 heads, 6 tails, P(H) on next flip = 1004 / (1004+1006) = 49.95%
- e.g., F ={3 heads, 3 tails} ~ weak expectation that any new coin will be fair
- After seeing 4 heads, 6 tails, P(H) on next flip = 7 / (7+9) = 43.75%

Prior knowledge too weak

But… flipping thumbtacks

- e.g., F ={4 heads, 3 tails} ~ weak expectation that tacks are slightly biased towards heads
- After seeing 2 heads, 0 tails, P(H) on next flip = 6 / (6+3) = 67%

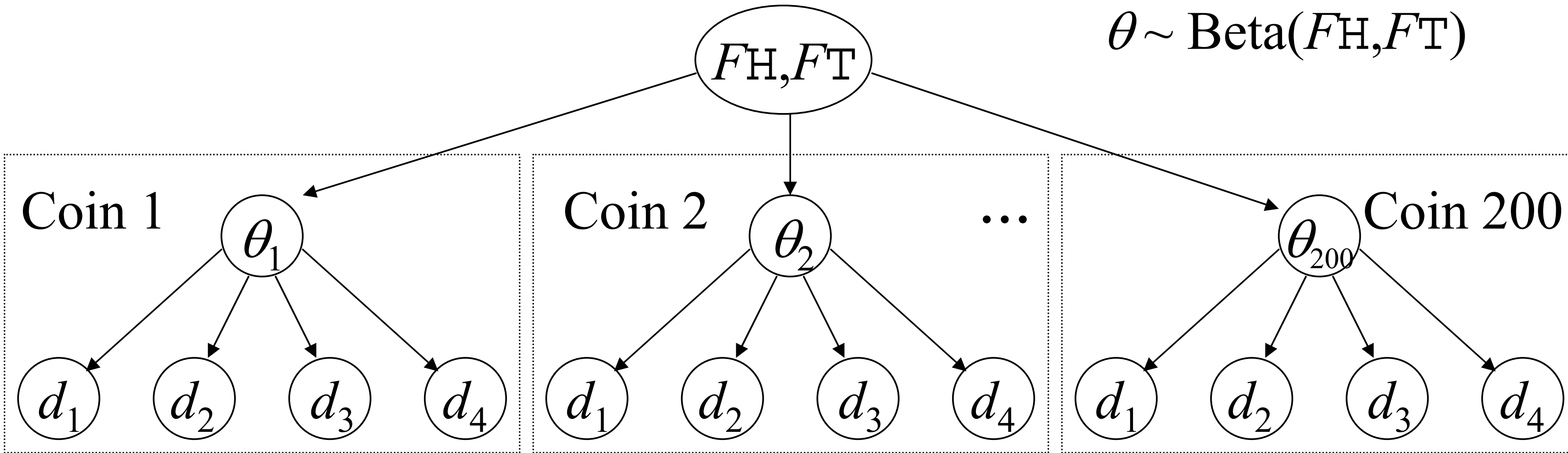
- Some prior knowledge is always necessary to avoid jumping to hasty conclusions...
- Suppose F = { }: After seeing 2 heads, 0 tails, P(H) on next flip = 2 / (2+0) = 100%

Origin of prior knowledge

- Tempting answer: prior experience
- Suppose you have previously seen 2000 coin flips: 1000 heads, 1000 tails

- By assuming all coins (and flips) are alike, these observations of *other* coins are as good as observations of the present coin
  - e.g., 200 coins flipped 10 times each.

Hierarchical priors



- Latent structure captures what is common to all coins, and also their individual variability

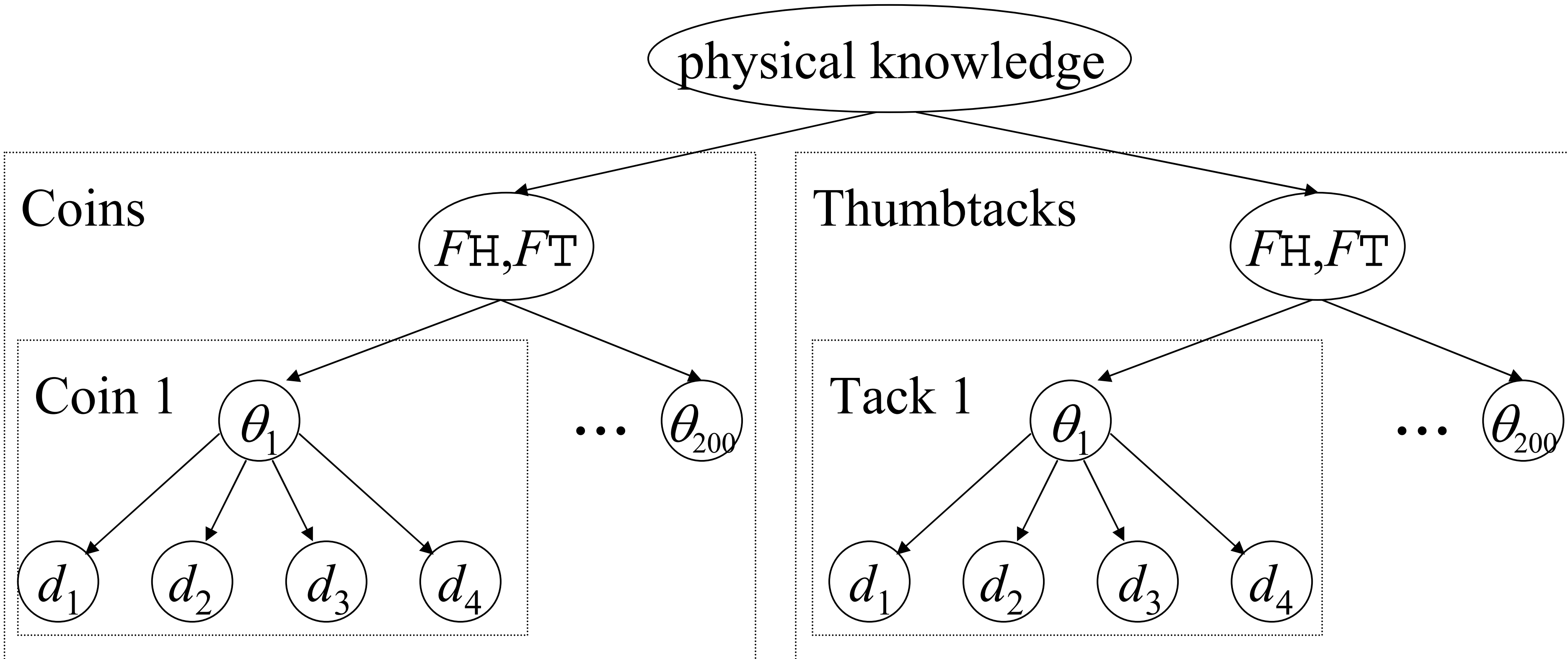
Problems with simple empiricism

- Haven't really seen 2000 coin flips, or *any* flips of a thumbtack
  - Prior knowledge is stronger than raw experience justifies
- Haven't seen exactly equal number of heads and tails
  - Prior knowledge is smoother than raw experience justifies
- Should be a difference between observing 2000 flips of a single coin versus observing 10 flips each for 200 coins, or 1 flip each for 2000 coins
  - Prior knowledge is more structured than raw experience

A simple theory

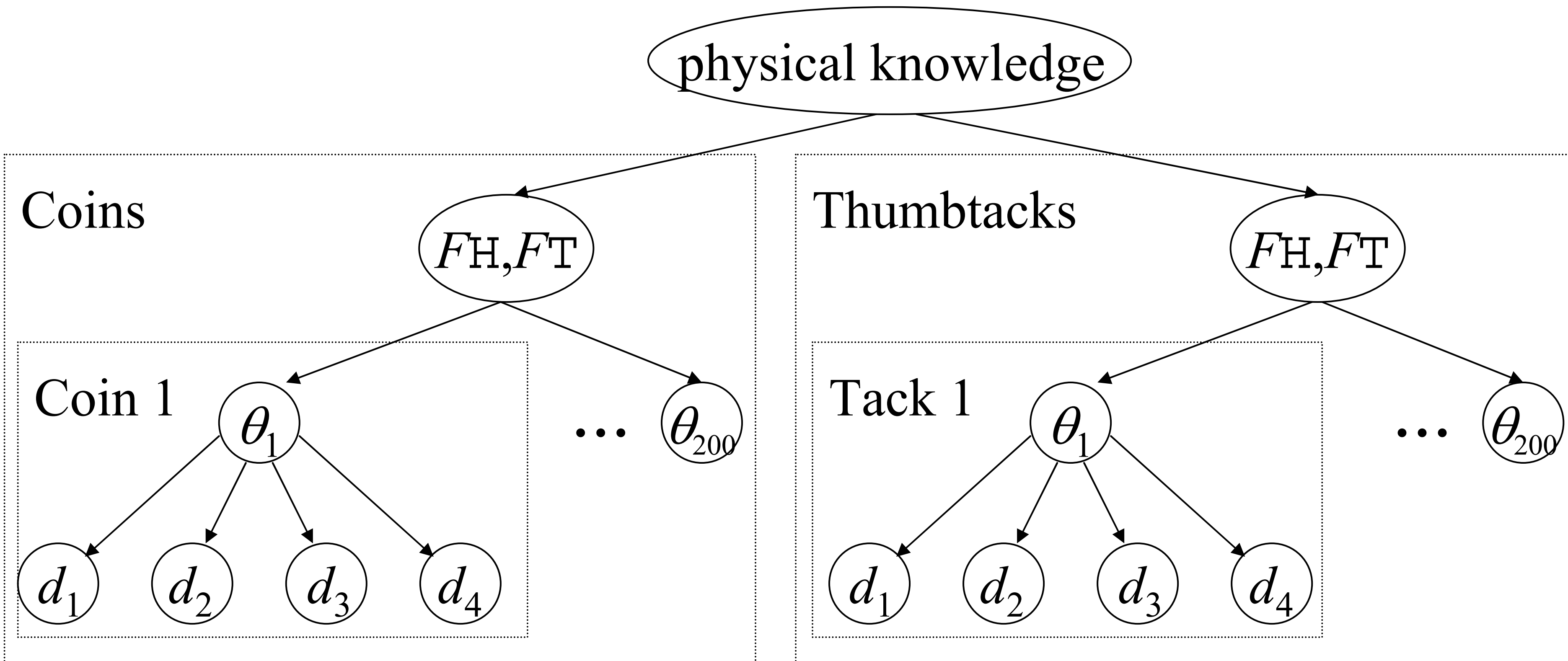
- "Coins are manufactured by a standardized procedure that is effective but not perfect."
  - Justifies generalizing from previous coins to the present coin.
  - Justifies smoother and stronger prior than raw experience alone.
  - Explains why seeing 10 flips each for 200 coins is more valuable than seeing 2000 flips of one coin.
- "Tacks are asymmetric, and manufactured to less exacting standards."

Hierarchical priors



- Qualitative beliefs (e.g. symmetry) can influence estimation of continuous properties (e.g. F_H, F_T)

Hierarchical priors



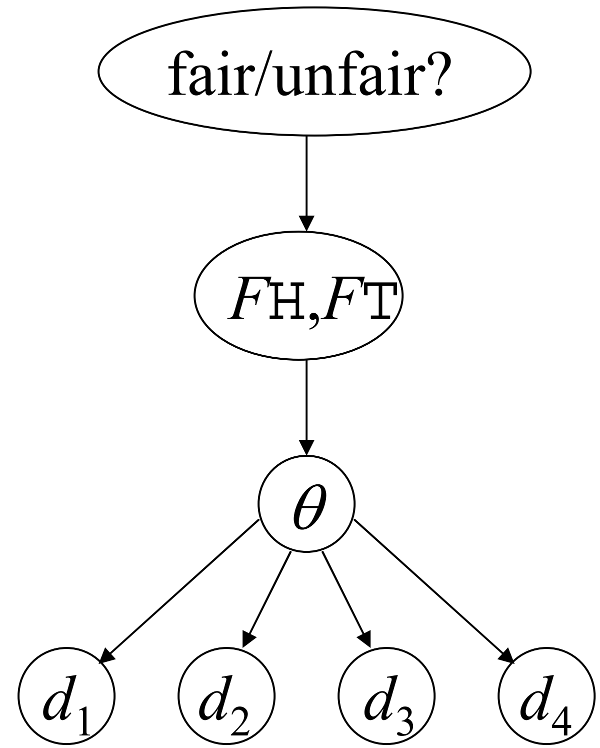
- Explains why 10 flips of 200 coin are better than 2000 flips of a single coin: more informative about FH, FT, assuming parameters not too large for new kind of coin.

Stability versus Flexibility

- Can all domain knowledge be represented with conjugate priors?
- Suppose you flip a coin 25 times and get all heads. *Something funny is going on* ...
- But with $F = \{1000 \text{ heads}, 1000 \text{ tails}\}$, P(heads) on next flip = 1025 / (1025+1000) = 50.6%. *Looks like nothing unusual.*
- How do we balance stability and flexibility?
  – Stability: 6 heads, 4 tails $\longrightarrow$ $\theta \sim 0.5$
  – Flexibility: 25 heads, 0 tails $\longrightarrow$ $\theta \sim 1$

Hierarchical priors

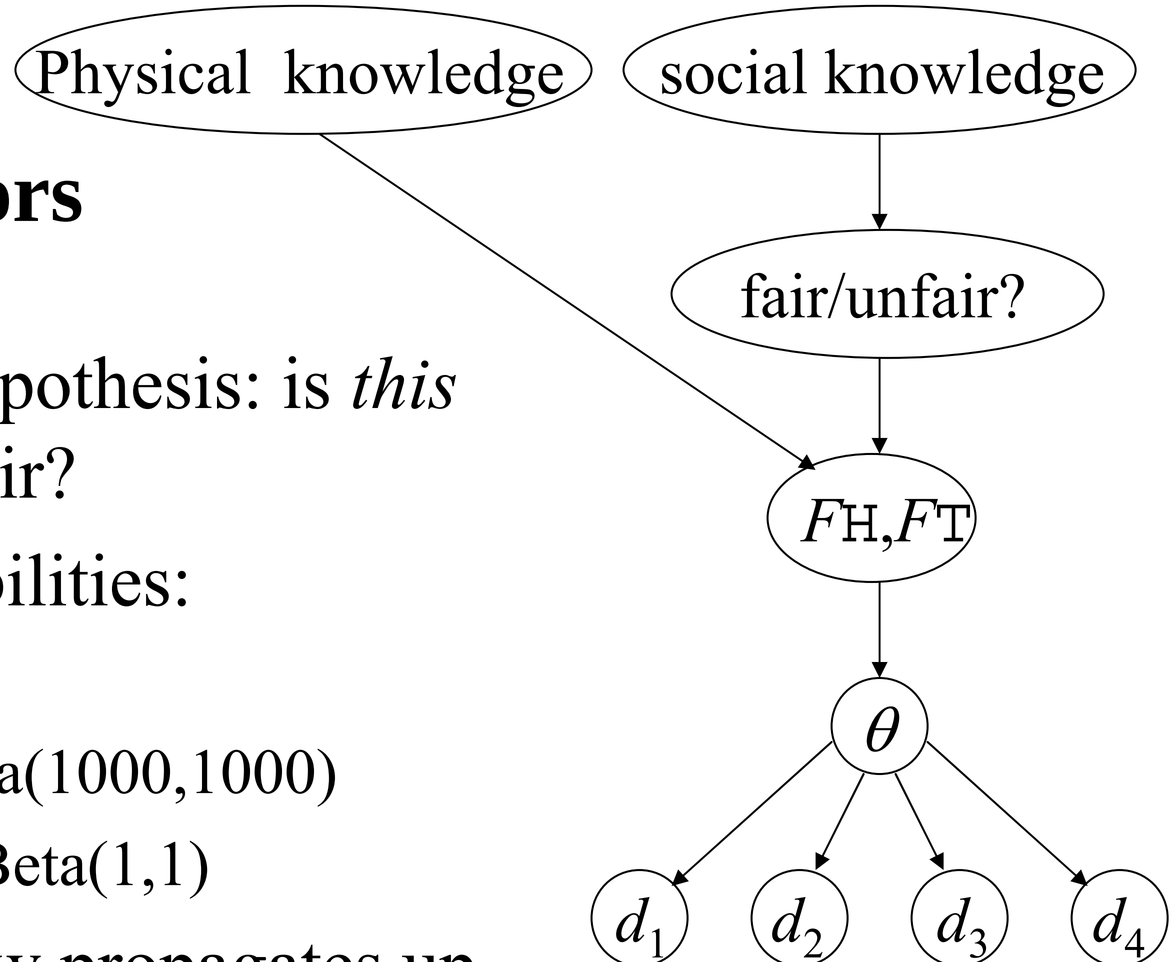
- Higher-order hypothesis: is *this* coin fair or unfair?
- Example probabilities:
  - $P(\text{fair}) = 0.99$
  - $P(\theta|\text{fair})$ is Beta(1000,1000)
  - $P(\theta|\text{unfair})$ is Beta(1,1)
- 25 heads in a row propagates up, affecting θ and then $P(\text{fair}|D)$



$$\frac{P(\text{fair}|\text{25 heads})}{P(\text{unfair}|\text{25 heads})} = \frac{P(\text{25 heads}|\text{fair})}{P(\text{25 heads}|\text{unfair})} \frac{\text{P(fair)}}{\text{P(unfair)}} = 9 \text{ x } 10^{-5}$$

$$P(D \mid \text{fair}) = \int_0^1 P(D \mid \theta)\, p(\theta \mid \text{fair})\, d\theta$$

Hierarchical priors



- Higher-order hypothesis: is *this* coin fair or unfair?
- Example probabilities:
  - $P(\text{fair}) = 0.99$
  - $P(\theta|\text{fair})$ is Beta(1000,1000)
  - $P(\theta|\text{unfair})$ is Beta(1,1)
- 25 heads in a row propagates up, affecting θ and then $P(\text{fair}|D)$

$$\frac{P(\text{fair}|\text{25 heads})}{P(\text{unfair}|\text{25 heads})} = \frac{P(\text{25 heads}|\text{fair})}{P(\text{25 heads}|\text{unfair})} \frac{\text{P(fair)}}{\text{P(unfair)}} = 9 \times 10^{-5}$$

Summary

- Learning the parameters of a generative model as Bayesian inference.
- Conjugate priors
  - an elegant way to represent simple kinds of prior knowledge.
- Hierarchical Bayesian models
  - integrate knowledge across instances of a system, or different systems within a domain.
  - can incorporate abstract theoretical knowledge.
  - inference may get difficult….

Other questions

- Learning isn't just about parameter estimation
  - How do we learn the functional form of a variable's distribution?
  - How do we learn model structure? Theories?
- Can we "grow" levels of abstraction?
- How do hierarchical Bayesian models address the Grue problem? Do we care?
- The "topics" model for semantics as an example of applying hierarchical Bayesian modeling to cognition. *Probably next time*.

Topic models of semantic structure: e.g., Latent Dirichlet Allocation (Blei, Ng, Jordan)

- Each document in a corpus is associated with a distribution θ over topics.
- Each topic t is associated with a distribution $\phi(t)$ over words.

Image removed due to copyright considerations. Please see:

Blei, David, Andrew Ng, and Michael Jordan. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3 (Jan 2003): 993-1022.

A selection of topics (TASA)

DISEASE	WATER	MIND	STORY	FIELD	SCIENCE	BALL	JOB
BACTERIA	FISH	WORLD	STORIES	MAGNETIC	STUDY	GAME	WORK
DISEASES	SEA	DREAM	TELL	MAGNET	SCIENTISTS	TEAM	JOBS
GERMS	SWIM	DREAMS	CHARACTER	WIRE	SCIENTIFIC	FOOTBALL	CAREER
FEVER	SWIMMING	THOUGHT	CHARACTERS	NEEDLE	KNOWLEDGE	BASEBALL	EXPERIENCE
CAUSE	POOL	IMAGINATION	AUTHOR	CURRENT	WORK	PLAYERS	EMPLOYMENT
CAUSED	LIKE	MOMENT	READ	COIL	RESEARCH	PLAY	OPPORTUNITIES
SPREAD	SHELL	THOUGHTS	TOLD	POLES	CHEMISTRY	FIELD	WORKING
VIRUSES	SHARK	OWN	SETTING	IRON	TECHNOLOGY	PLAYER	TRAINING
INFECTION	TANK	REAL	TALES	COMPASS	MANY	BASKETBALL	SKILLS
VIRUS	SHELLS	LIFE	PLOT	LINES	MATHEMATICS	COACH	CAREERS
MICROORGANISMS	SHARKS	IMAGINE	TELLING	CORE	BIOLOGY	PLAYED	POSITIONS
PERSON	DIVING	SENSE	SHORT	ELECTRIC	FIELD	PLAYING	FIND
INFECTIOUS	DOLPHINS	CONSCIOUSNESS	FICTION	DIRECTION	PHYSICS	HIT	POSITION
COMMON	SWAM	STRANGE	ACTION	FORCE	LABORATORY	TENNIS	FIELD
CAUSING	LONG	FEELING	TRUE	MAGNETS	STUDIES	TEAMS	OCCUPATIONS
SMALLPOX	SEAL	WHOLE	EVENTS	BE	WORLD	GAMES	REQUIRE
BODY	DIVE	BEING	TELLS	MAGNETISM	SCIENTIST	SPORTS	OPPORTUNITY
INFECTIONS	DOLPHIN	MIGHT	TALE	POLE	STUDYING	BAT	EARN
CERTAIN	UNDERWATER	HOPE	NOVEL	INDUCED	SCIENCES	TERRY	ABLE

A selection of topics (TASA)

DISEASE	WATER	MIND	STORY	**FIELD**	SCIENCE	BALL	JOB
BACTERIA	FISH	WORLD	STORIES	MAGNETIC	STUDY	GAME	WORK
DISEASES	SEA	DREAM	TELL	MAGNET	SCIENTISTS	TEAM	JOBS
GERMS	SWIM	DREAMS	CHARACTER	WIRE	SCIENTIFIC	FOOTBALL	CAREER
FEVER	SWIMMING	THOUGHT	CHARACTERS	NEEDLE	KNOWLEDGE	BASEBALL	EXPERIENCE
CAUSE	POOL	IMAGINATION	AUTHOR	CURRENT	WORK	PLAYERS	EMPLOYMENT
CAUSED	LIKE	MOMENT	READ	COIL	RESEARCH	PLAY	OPPORTUNITIES
SPREAD	SHELL	THOUGHTS	TOLD	POLES	CHEMISTRY	**FIELD**	WORKING
VIRUSES	SHARK	OWN	SETTING	IRON	TECHNOLOGY	PLAYER	TRAINING
INFECTION	TANK	REAL	TALES	COMPASS	MANY	BASKETBALL	SKILLS
VIRUS	SHELLS	LIFE	PLOT	LINES	MATHEMATICS	COACH	CAREERS
MICROORGANISMS	SHARKS	IMAGINE	TELLING	CORE	BIOLOGY	PLAYED	POSITIONS
PERSON	DIVING	SENSE	SHORT	ELECTRIC	**FIELD**	PLAYING	FIND
INFECTIOUS	DOLPHINS	CONSCIOUSNESS	FICTION	DIRECTION	PHYSICS	HIT	POSITION
COMMON	SWAM	STRANGE	ACTION	FORCE	LABORATORY	TENNIS	**FIELD**
CAUSING	LONG	FEELING	TRUE	MAGNETS	STUDIES	TEAMS	OCCUPATIONS
SMALLPOX	SEAL	WHOLE	EVENTS	BE	WORLD	GAMES	REQUIRE
BODY	DIVE	BEING	TELLS	MAGNETISM	SCIENTIST	SPORTS	OPPORTUNITY
INFECTIONS	DOLPHIN	MIGHT	TALE	POLE	STUDYING	BAT	EARN
CERTAIN	UNDERWATER	HOPE	NOVEL	INDUCED	SCIENCES	TERRY	ABLE

Joint models of syntax and semantics

(Griffiths, Steyvers, Blei & Tenenbaum, NIPS 2004)

- Embed topics model inside an *n*th order Hidden Markov Model:

Griffiths, T. L., M. Steyvers, D. M. Blei, and J. B. Tenenbaum.Integrating Topics and Syntax. *Advances in Neural Information Processing Systems* 17 (2005).

Image removed due to copyright considerations. Please see:
Griffiths, T. L., M. Steyvers, D. M. Blei, and J. B. Tenenbaum. Integrating Topics and Syntax. *Advances in Neural Information Processing Systems* 17 (2005).

Semantic classes

FOOD	MAP	DOCTOR	BOOK	GOLD	BEHAVIOR	CELLS	PLANTS
FOODS	NORTH	PATIENT	BOOKS	IRON	SELF	CELL	PLANT
BODY	EARTH	HEALTH	READING	SILVER	INDIVIDUAL	ORGANISMS	LEAVES
NUTRIENTS	SOUTH	HOSPITAL	INFORMATION	COPPER	PERSONALITY	ALGAE	SEEDS
DIET	POLE	MEDICAL	LIBRARY	METAL	RESPONSE	BACTERIA	SOIL
FAT	MAPS	CARE	REPORT	METALS	SOCIAL	MICROSCOPE	ROOTS
SUGAR	EQUATOR	PATIENTS	PAGE	STEEL	EMOTIONAL	MEMBRANE	FLOWERS
ENERGY	WEST	NURSE	TITLE	CLAY	LEARNING	ORGANISM	WATER
MILK	LINES	DOCTORS	SUBJECT	LEAD	FEELINGS	FOOD	FOOD
EATING	EAST	MEDICINE	PAGES	ADAM	PSYCHOLOGISTS	LIVING	GREEN
FRUITS	AUSTRALIA	NURSING	GUIDE	ORE	INDIVIDUALS	FUNGI	SEED
VEGETABLES	GLOBE	TREATMENT	WORDS	ALUMINUM	PSYCHOLOGICAL	MOLD	STEMS
WEIGHT	POLES	NURSES	MATERIAL	MINERAL	EXPERIENCES	MATERIALS	FLOWER
FATS	HEMISPHERE	PHYSICIAN	ARTICLE	MINE	ENVIRONMENT	NUCLEUS	STEM
NEEDS	LATITUDE	HOSPITALS	ARTICLES	STONE	HUMAN	CELLED	LEAF
CARBOHYDRATES	PLACES	DR	WORD	MINERALS	RESPONSES	STRUCTURES	ANIMALS
VITAMINS	LAND	SICK	FACTS	POT	BEHAVIORS	MATERIAL	ROOT
CALORIES	WORLD	ASSISTANT	AUTHOR	MINING	ATTITUDES	STRUCTURE	POLLEN
PROTEIN	COMPASS	EMERGENCY	REFERENCE	MINERS	PSYCHOLOGY	GREEN	GROWING
MINERALS	CONTINENTS	PRACTICE	NOTE	TIN	PERSON	MOLDS	GROW

Image removed due to copyright considerations. Please see:
Griffiths, T. L., M. Steyvers, D. M. Blei, and J. B. Tenenbaum.Integrating Topics and Syntax. *Advances in Neural Information Processing Systems* 17 (2005).

Syntactic classes

SAID	THE	MORE	ON	GOOD	ONE	HE	BE
ASKED	HIS	SUCH	AT	SMALL	SOME	YOU	MAKE
THOUGHT	THEIR	LESS	INTO	NEW	MANY	THEY	GET
TOLD	YOUR	MUCH	FROM	IMPORTANT	TWO	I	HAVE
SAYS	HER	KNOWN	WITH	GREAT	EACH	SHE	GO
MEANS	ITS	JUST	THROUGH	LITTLE	ALL	WE	TAKE
CALLED	MY	BETTER	OVER	LARGE	MOST	IT	DO
CRIED	OUR	RATHER	AROUND	*	ANY	PEOPLE	FIND
SHOWS	THIS	GREATER	AGAINST	BIG	THREE	EVERYONE	USE
ANSWERED	THESE	HIGHER	ACROSS	LONG	THIS	OTHERS	SEE
TELLS	A	LARGER	UPON	HIGH	EVERY	SCIENTISTS	HELP
REPLIED	AN	LONGER	TOWARD	DIFFERENT	SEVERAL	SOMEONE	KEEP
SHOUTED	THAT	FASTER	UNDER	SPECIAL	FOUR	WHO	GIVE
EXPLAINED	NEW	EXACTLY	ALONG	OLD	FIVE	NOBODY	LOOK
LAUGHED	THOSE	SMALLER	NEAR	STRONG	BOTH	ONE	COME
MEANT	EACH	SOMETHING	BEHIND	YOUNG	TEN	SOMETHING	WORK
WROTE	MR	BIGGER	OFF	COMMON	SIX	ANYONE	MOVE
SHOWED	ANY	FEWER	ABOVE	WHITE	MUCH	EVERYBODY	LIVE
BELIEVED	MRS	LOWER	DOWN	SINGLE	TWENTY	SOME	EAT
WHISPERED	ALL	ALMOST	BEFORE	CERTAIN	EIGHT	THEN	BECOME

Corpus-specific factorization (NIPS)

Image removed due to copyright considerations. Please see:

Griffiths, T. L., M. Steyvers, D. M. Blei, and J. B. Tenenbaum.Integrating Topics and Syntax. *Advances in Neural Information Processing Systems* 17 (2005).

Syntactic classes in PNAS

5	8	14	25	26	30	33
IN	ARE	THE	SUGGEST	LEVELS	RESULTS	BEEN
FOR	WERE	THIS	INDICATE	NUMBER	ANALYSIS	MAY
ON	WAS	ITS	SUGGESTING	LEVEL	DATA	CAN
BETWEEN	IS	THEIR	SUGGESTS	RATE	STUDIES	COULD
DURING	WHEN	AN	SHOWED	TIME	STUDY	WELL
AMONG	REMAIN	EACH	REVEALED	CONCENTRATIONS	FINDINGS	DID
FROM	REMAINS	ONE	SHOW	VARIETY	EXPERIMENTS	DOES
UNDER	REMAINED	ANY	DEMONSTRATE	RANGE	OBSERVATIONS	DO
WITHIN	PREVIOUSLY	INCREASED	INDICATING	CONCENTRATION	HYPOTHESIS	MIGHT
THROUGHOUT	BECOME	EXOGENOUS	PROVIDE	DOSE	ANALYSES	SHOULD
THROUGH	BECAME	OUR	SUPPORT	FAMILY	ASSAYS	WILL
TOWARD	BEING	RECOMBINANT	INDICATES	SET	POSSIBILITY	WOULD
INTO	BUT	ENDOGENOUS	PROVIDES	FREQUENCY	MICROSCOPY	MUST
AT	GIVE	TOTAL	INDICATED	SERIES	PAPER	CANNOT
INVOLVING	MERE	PURIFIED	DEMONSTRATED	AMOUNTS	WORK	REMAINED
AFTER	APPEARED	TILE	SHOWS	RATES	EVIDENCE	ALSO
ACROSS	APPEAR	FULL	SO	CLASS	FINDING	THEY
AGAINST	ALLOWED	CHRONIC	REVEAL	VALUES	MUTAGENESIS	BECOME
WHEN	NORMALLY	ANOTHER	DEMONSTRATES	AMOUNT	OBSERVATION	MAG
ALONG	EACH	EXCESS	SUGGESTED	SITES	MEASUREMENTS	LIKELY

Semantic highlighting

Darker words are more likely to have been generated from the topic-based "semantics" module:

In contrast to this approach, we study here how the overall network activity can control single cell parameters such as input resistance, as well as time and space constants, parameters that are crucial for excitability and spariotemporal (sic) integration.

The integrated architecture in this paper combines feed forward control and error feedback adaptive control using neural networks.

In other words, for our proof of convergence, we require the softassign algorithm to return a doubly stochastic matrix as *sinkhorn theorem guarantees that it will instead of a matrix which is merely close to being doubly stochastic based on some reasonable metric.

The aim is to construct a portfolio with a maximal expected return for a given risk level and time horizon while simultaneously obeying *institutional or *legally required constraints.

The left graph is the standard experiment the right from a training with # samples.

The graph G is called the *guest graph, and H is called the host graph.

Outline

- Bayesian parameter estimation
- Hierarchical Bayesian models
- Metropolis-Hastings
  - A more general approach to MCMC

Motivation

- Want to compute $P(h|evidence)$:

$$P(h \mid evidence) = \frac{P(evidence \mid h)P(h)}{\sum_{h'} P(evidence \mid h')P(h')}$$

- General problem with complex models: sum over alternative hypotheses is intractable.

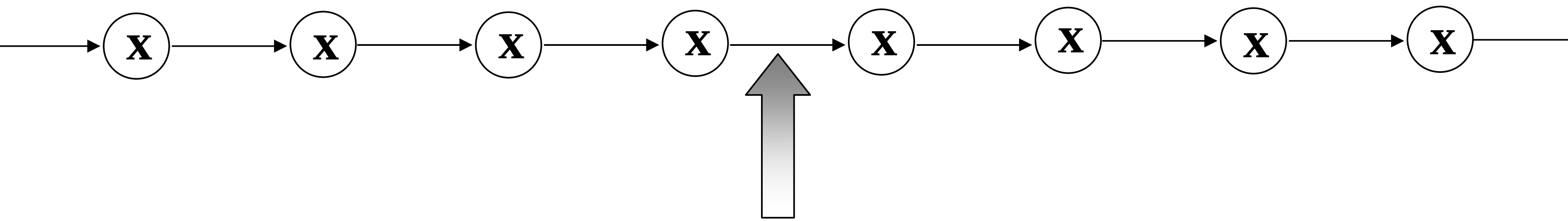
Markov chain Monte Carlo

- Sample from a Markov chain which converges to posterior distribution

- After an initial "burn in" period, samples are independent of starting conditions.

Image removed due to copyright considerations.

What's a Markov chain?



Transition matrix

$$P(\mathbf{x}^{(t+1)}|\mathbf{x}^{(t)}) = T(\mathbf{x}^{(t)},\mathbf{x}^{(t+1)})$$

- States of chain are variables of interest
- Transition matrix chosen to give posterior distribution as stationary distribution

Gibbs sampling

- Suppose (1) we can factor hypotheses into individual state variables, $h = <h_1, h_2, \ldots, h_n>$;
- and (2) we can easily compute

$$P(h_i|h_{-i}, evidence), \text{ where}$$

$$h_{-i} = h_1^{(t+1)}, h_2^{(t+1)}, \ldots, h_{i-1}^{(t+1)}, h_{i+1}^{(t)}, \ldots, h_n^{(t)}$$

- Then use Gibbs sampling:
  - Cycle through variables $h_1, h_2, \ldots, h_n$
  - Draw $h_i^{(t+1)}$ from $P(h_i|h_{-i}, evidence)$

Gibbs sampling

Image removed due to copyright considerations.

(MacKay, 2002)

Motivation for Metropolis-Hastings

- Want to compute $P(h|evidence)$:

$$P(h \mid evidence) = \frac{P(evidence \mid h)P(h)}{\sum_{h'} P(evidence \mid h')P(h')}$$

- We have a probabilistic model that allows us to compute $P(evidence|h)$ and $P(h)$.
- We can compute *relative posteriors*:

$$\frac{P(h_i \mid evidence)}{P(h_j \mid evidence)} = \frac{P(evidence \mid h_i)P(h_i)}{P(evidence \mid h_j)P(h_j)}$$

Metropolis-Hastings algorithm

- Transitions have two parts:
  - proposal distribution: $Q(h^{(t+1)} | h^{(t)})$
  - acceptance: take proposals with probability

$$\mathrm{A}(h^{(t+1)} | h^{(t)}) = \min\left\{ 1, \frac{P(h^{(t+1)} | evidence)\, Q(h^{(t)} | h^{(t+1)})}{P(h^{(t)} | evidence)\, Q(h^{(t+1)} | h^{(t)})} \right\}$$

Metropolis-Hastings algorithm

Complex unknown posterior distribution

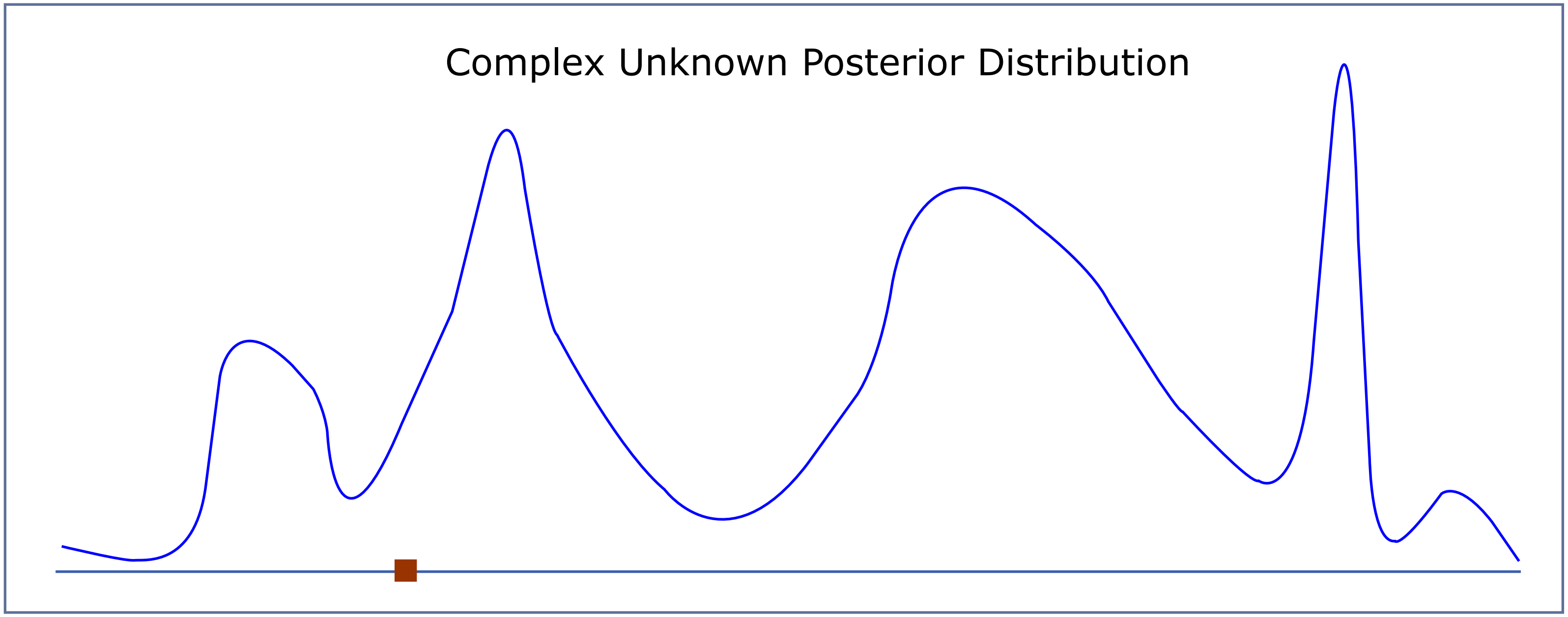


Figure by MIT OCW.

Metropolis-Hastings algorithm

Complex unknown posterior distribution

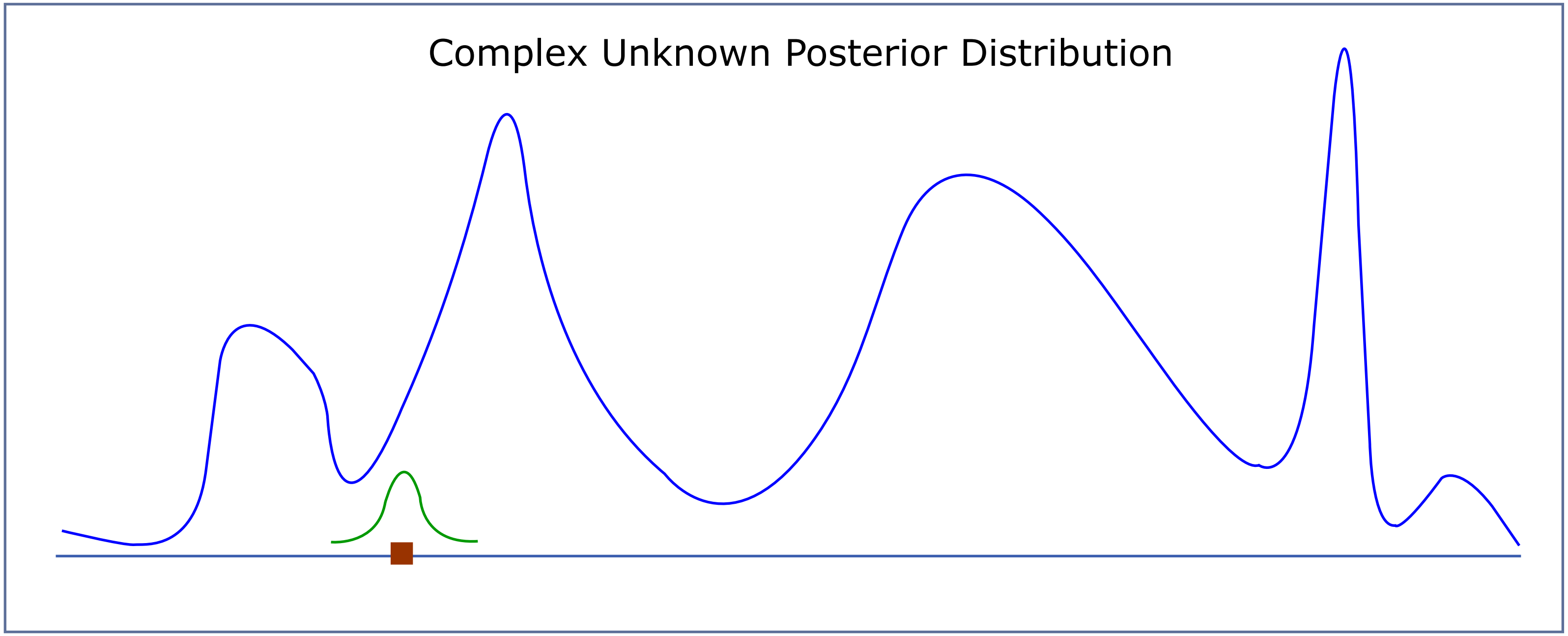


Figure by MIT OCW.

e.g., Gaussian proposal distribution

Metropolis-Hastings algorithm

Complex unknown posterior distribution

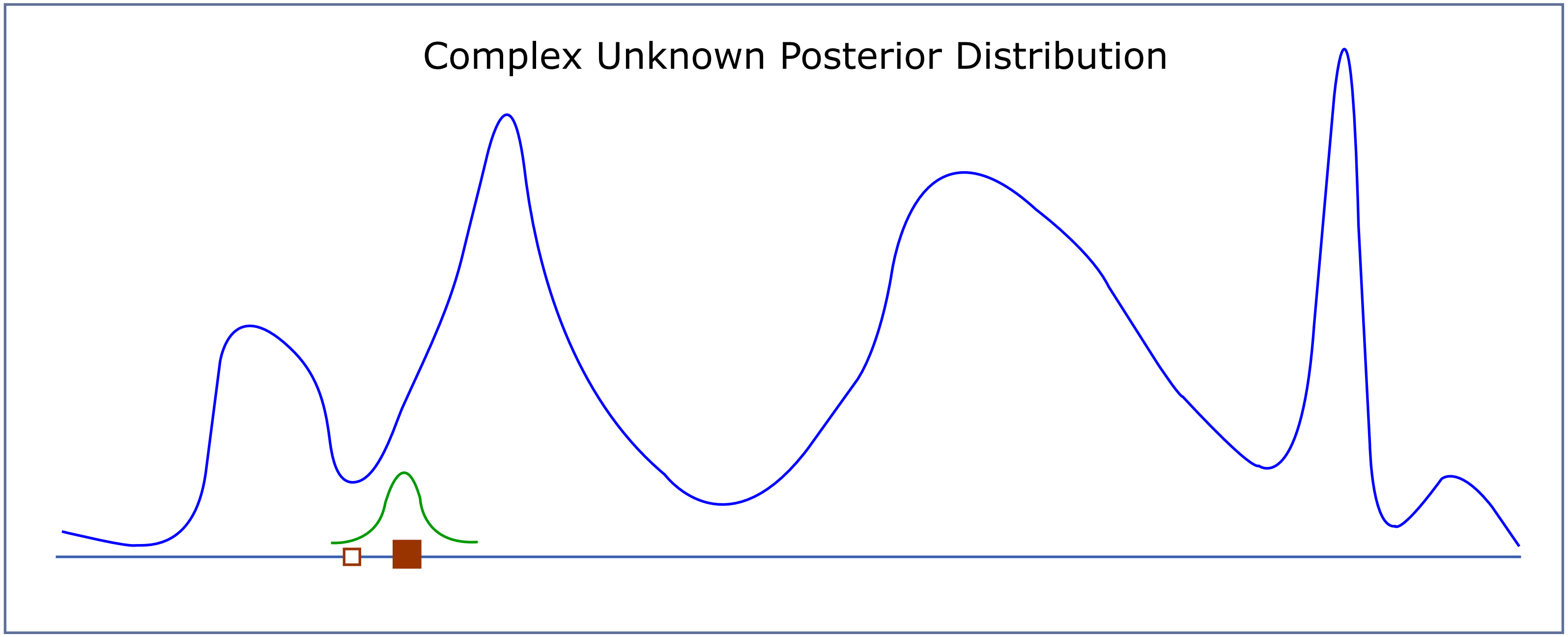


Figure by MIT OCW.

e.g., Gaussian proposal distribution

Metropolis-Hastings algorithm

Complex unknown posterior distribution

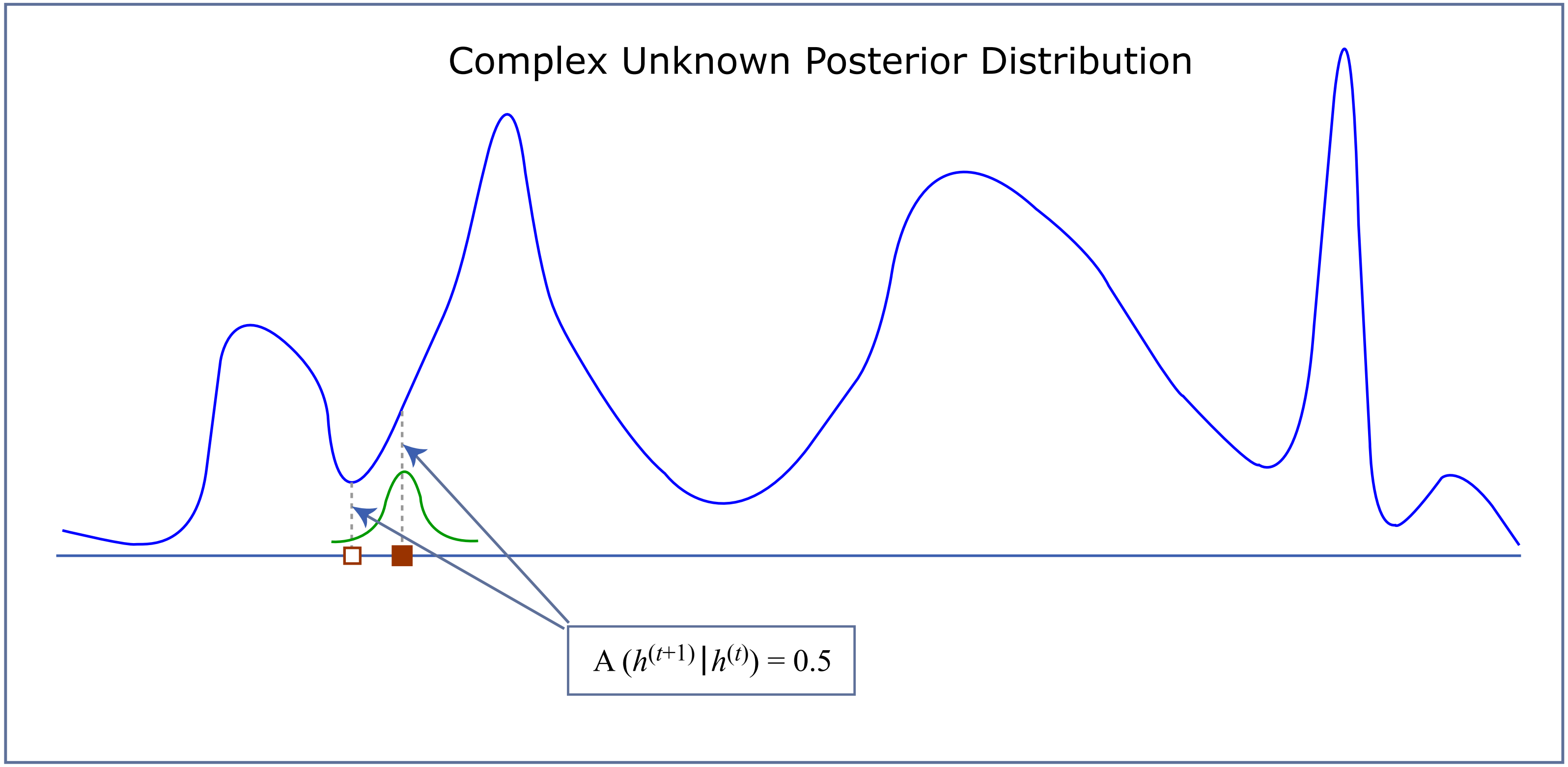


Figure by MIT OCW.

e.g., Gaussian proposal distribution

Metropolis-Hastings algorithm

Complex unknown posterior distribution

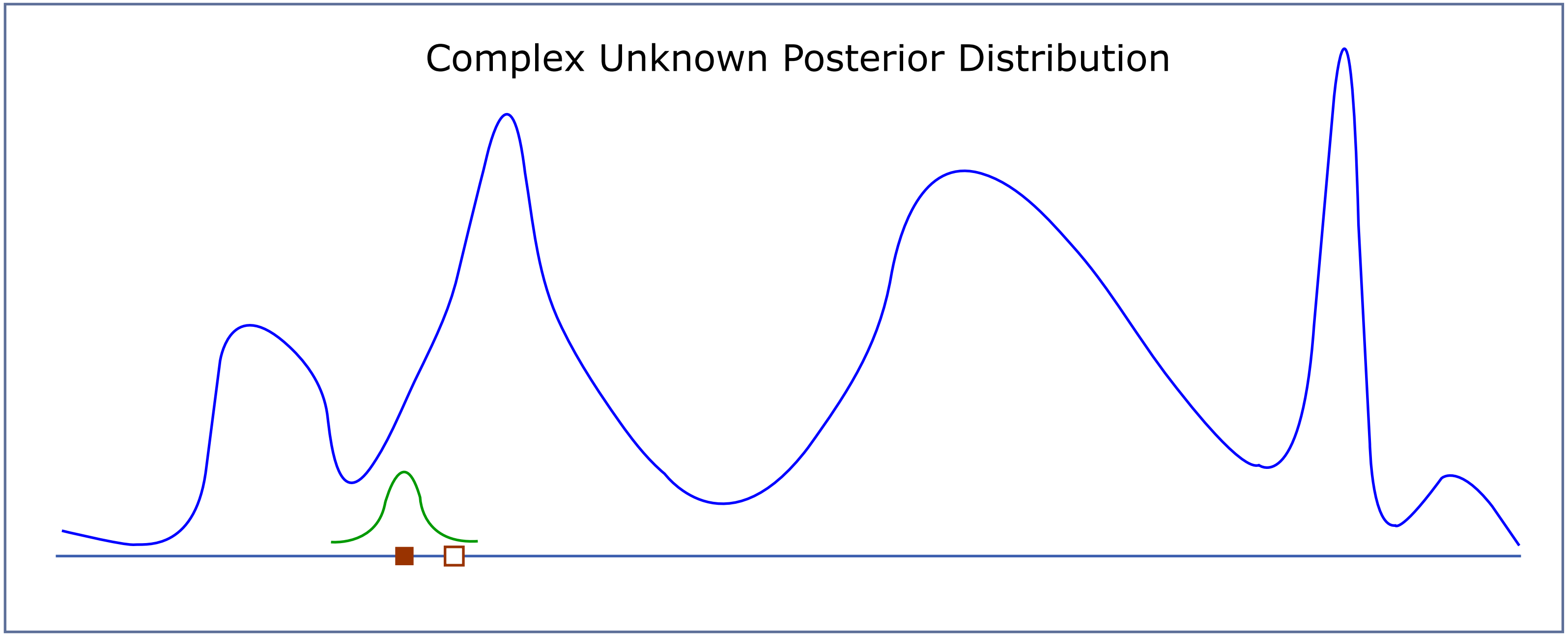


Figure by MIT OCW.

e.g., Gaussian proposal distribution

Metropolis-Hastings algorithm

Complex unknown posterior distribution

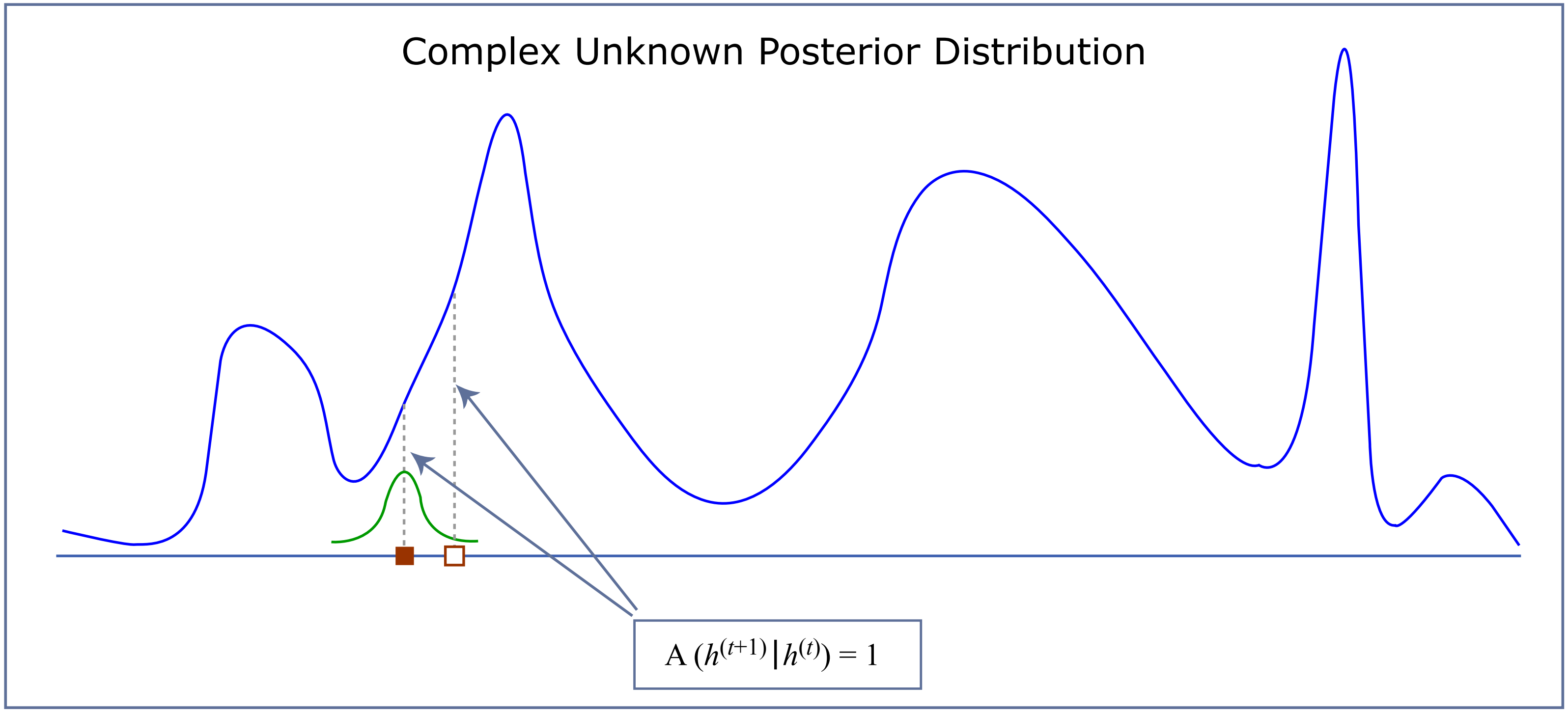


Figure by MIT OCW.

e.g., Gaussian proposal distribution

Advanced topics

- What makes a good proposal distribution?
  - "Goldilocks principle"
  - May be data-dependent
- Connections to simulated annealing
  - Integration versus optimization
  - MCMC at different temperatures
- MCMC over model structures
  - Reversible jump MCMC

Relation to simulated annealing

Complex unknown cost function

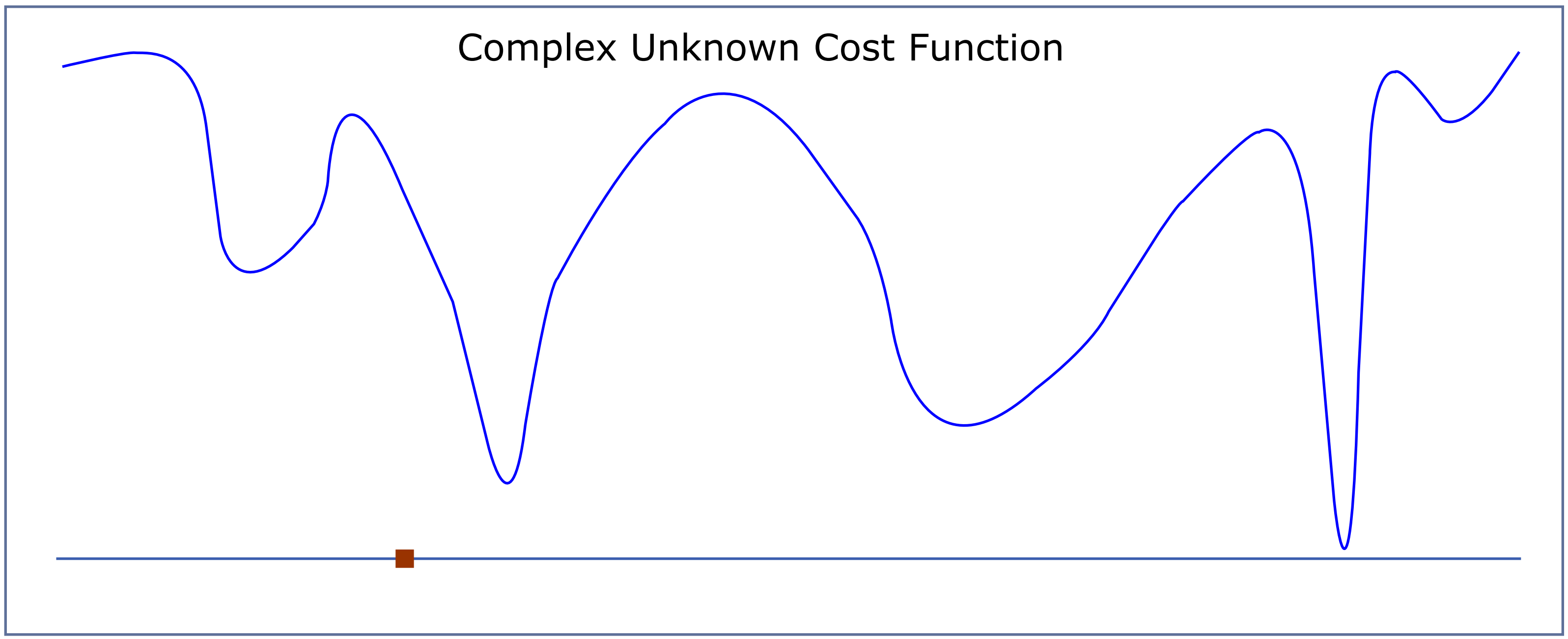


Figure by MIT OCW.

Why MCMC is important

- Simple
- Can be used with just about any kind of probabilistic model, including complex hierarchical structures
- Always works pretty well, if you’re willing to wait a long time

(cf. Backpropagation for neural networks.)

A model for cognitive development?

- Some features of cognitive development:
  - Small, random, dumb, local steps
  - Takes a long time
  - Can get stuck in plateaus or stages
  - "Two steps forward, one step back"
  - Over time, intuitive theories get consistently better (more veridical, more powerful, broader scope).
  - Everyone reaches basically the same state (though some take longer than others).