15-780: Grad AI
Lecture 17: Probability

Geoff Gordon (this lecture)
Tuomas Sandholm
TAs Erik Zawadzki, Abe Othman

Review: probability

- RVs, events, sample space Ω
- Measures, distributions
  - disjoint union property (law of total probability or "sum rule")
- Sample v. population
- Law of large numbers
- Marginals, conditionals

Suggested reading

- Bishop, Pattern Recognition and Machine Learning, p1–4, sec 1–1.2, sec 2–2.3

Terminology

- Experiment = planned obs.
- Prior = know ahead
- Posterior = after obs.

Example: model selection

- You're gambling to decide who has to clean the lab
- You are accused of using weighted dice!
- Two models:
  - .9 fair dice: all 36 rolls equally likely
  - .1 weighted: rolls summing to 7 more likely → 2/36 sum 7, 1/2 · 1/30 o/w

prior:

observation: 2-5 3-4

posterior:

fair 2-5 3-4 : $.9 \times \frac{1}{36} \times \frac{1}{36}$

w 2-5 3-4 : $.1 \times \frac{2}{36} \times \frac{2}{36}$

P(w, 2-5, 4-4)

$= .9 \times \frac{2}{36} \times \frac{1}{60}$

f 9/13

w 4/13

Independence

- X and Y are ***independent*** if, for all possible values of y, P(X) = P(X | Y=y)
  - equivalently, for all possible values of x, P(Y) = P(Y | X=x)
  - equivalently, P(X,Y) = P(X) P(Y)
- Knowing X or Y gives us no information about the other

Independence: probability = product of marginals

AAPL price

Weather

	up	same	down	
sun	0.09	0.15	0.06	0.3
rain	0.21	0.35	0.14	0.7
	0.3	0.5	0.2	

Expectations

- How much should we expect to earn from our AAPL stock? R

$$E(R) = \sum_{\text{atomic events } \omega} P(\omega) R(\omega)$$

$$= .09 \cdot 1 + .15 \cdot 0 + \ldots$$

$$= .1$$

AAPL price

Weather	up	same	down
sun	0.09	0.15	0.06
rain	0.21	0.35	0.14

Weather	up	same	down
sun	+1	0	-1
rain	+1	0	-1

Linearity of expectation

- Expectation is a linear function of numbers in bottom table
- E.g., suppose we own *k* shares

$E(kR) = .09k + \dots + .14(-k)$

$= k(.1) = k\,E(R)$

AAPL price

Weather	up	same	down
sun	0.09	0.15	0.06
rain	0.21	0.35	0.14

Weather	up	same	down
sun	+k	0	-k
rain	+k	0	-k

Conditional expectation

.3 .5 .2

- What if we know it's sunny?

$E(R \mid sun)$

$= .3(1) + .5(0) + .2(-1)$

$= .1$

AAPL price

Weather	up	same	down
sun	0.09	0.15	0.06
rain	0.21	0.35	0.14

Weather	up	same	down
sun	+1	0	-1
rain	+1	0	-1

Independence and expectation

- If X and Y are independent, E(XY) = E(X)E(Y)
- Proof:

$$\sum_{x,y} P(x,y)\,xy = \sum_{x,y} p(x)p(y)\,xy$$

$$= \sum_{x} p(x)\,x \sum_{y} p(y)\,y$$

Sample means

- Sample mean = $\bar{X} = \frac{1}{N}\sum_i X_i$
- Expectation of sample mean:

$$E(\bar{X}) = E\left(\frac{1}{N}\sum_i X_i\right) = \frac{1}{N}\sum_i E(X_i) = \frac{1}{N}\sum_i \mu = \mu$$

Estimators

- Common task: given a sample, infer something about the population
- An ***estimator*** is a function of a sample that we use to tell us something about the population
- E.g., sample mean is a good estimator of population mean
- E.g., linear regression

Law of large numbers
(more general form)

- For r.v. X: if we take a sample of size N from a distribution P(x) with mean μ and compute sample mean $\bar{X}$
- Then $\bar{X} \to \mu$ as $N \to \infty$

Bias

- Given estimator T of population quantity θ
- The ***bias*** of T is $E(T) - \theta$
- Sample mean is ***unbiased*** estimator of population mean
- $(1 + \sum x_i) / (N+1)$ is biased, but ***asymptotically unbiased***

Variance

- Two estimators of population mean: sample mean, mean of every 2nd sample
- Both unbiased, but one is more variable
- Measure of variability: variance

Variance

- If zero-mean: variance = $E(X^2)$
  - Ex: constant 0 v. coin-flip ±1

    ↳ 0 $\quad .5(1^2) + .5(-1^2) = 1$

- In general: $E([X - E(X)]^2)$
  - equivalently, $E(X^2) - E(X)^2$ (but note numerical problem)

Exercise

- What is the variance of 3X?

$E((3X)^2) = E(9X^2) = 9E(X^2)$

Sample variance

- Sample variance =
- Expectation: V(X)
- Sample size correction:

$$\frac{N-1}{N}\sum_i (x_i - \bar{x})^2 \quad \frac{1}{N-1}$$

$$\frac{N-1}{N}$$

Bias-variance decomposition

- Estimator T of population quantity θ
- ***Mean squared error*** = E((T – θ)²) =

$$E((T - E(T) + E(T) - \theta)^2)$$

$$= E((T - E(T))^2 + (T - E(T))(E(T) - \theta) + (E(T) - \theta)^2)$$

$$= \underbrace{E(T - E(T))^2)}_{\text{var}} + 0 + \underbrace{(E(T) - \theta)^2}_{\text{bias}^2}$$

Bias-variance tradeoff

- It's nice to have estimators w/ small MSE
- There is a ***smallest possible*** MSE for a given amount of data
  - limited data provides limited information
- Estimator which achieves min is ***efficient*** (close for large N: ***asymptotically eff.***)
- Often can adjust estimator so MSE is due to bias or variance—the famed ***tradeoff***

Covariance

- Suppose we want an approximate numeric measure of (in)dependence
- Let E(X) = E(Y) = 0 for simplicity
- Consider the random variable XY
  - if X, Y are typically both +ve or both -ve

    XY ">" 0 $E(XY) > 0$

  - if X, Y are independent

    $E(XY) = E(X)E(Y) = 0$

Covariance

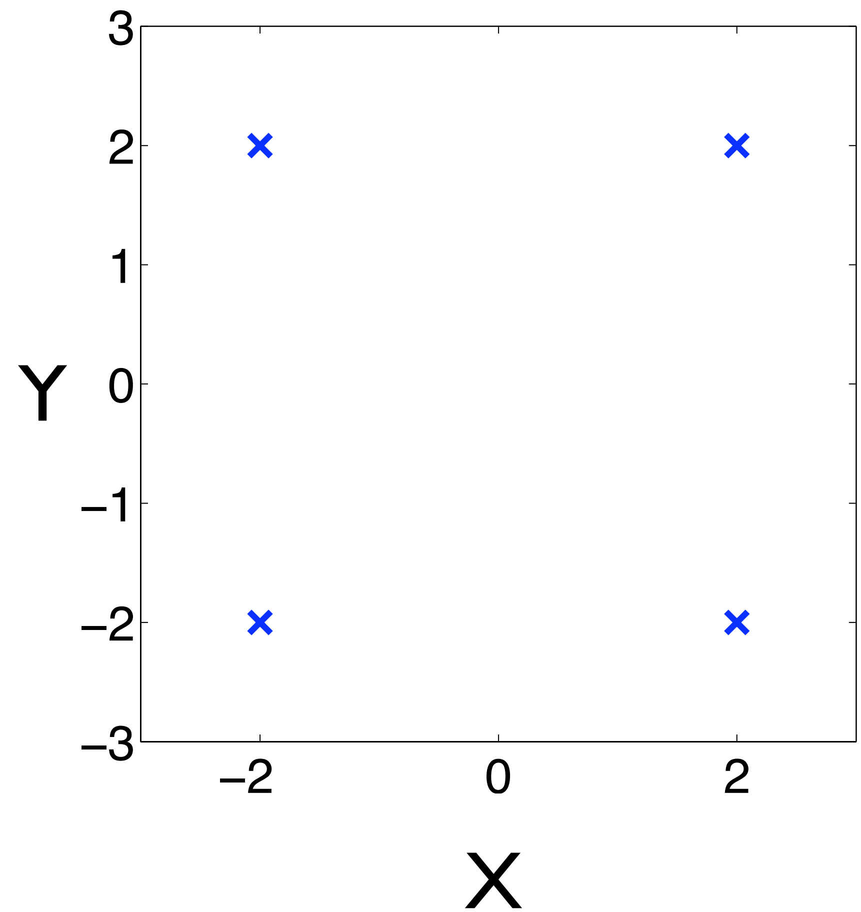
- $\text{cov}(X,Y) = E([X - E(X)][Y - E(Y)])$
- Is this a good measure of dependence?
  - Suppose we scale X by 10
  - $\text{cov}(10X,Y) = E([10X - E(10X)][Y - E(Y)])$
  - $\text{cov}(10X,Y) = 10\ \text{cov}(X,Y)$

Correlation

- Like covariance, but controls for variance of individual r.v.s
- $\text{cor}(X,Y) = \text{cov}(X,Y)/\sqrt{\text{var}(X)\text{var}(Y)}$
- $\text{cor}(10X,Y) =$ cor(X, Y)

Correlation & independence

- Equal probability on each point
- Are X and Y independent? Y
- Are X and Y uncorrelated? Y



Correlation & independence

- Do you think that all independent pairs of RVs are uncorrelated?

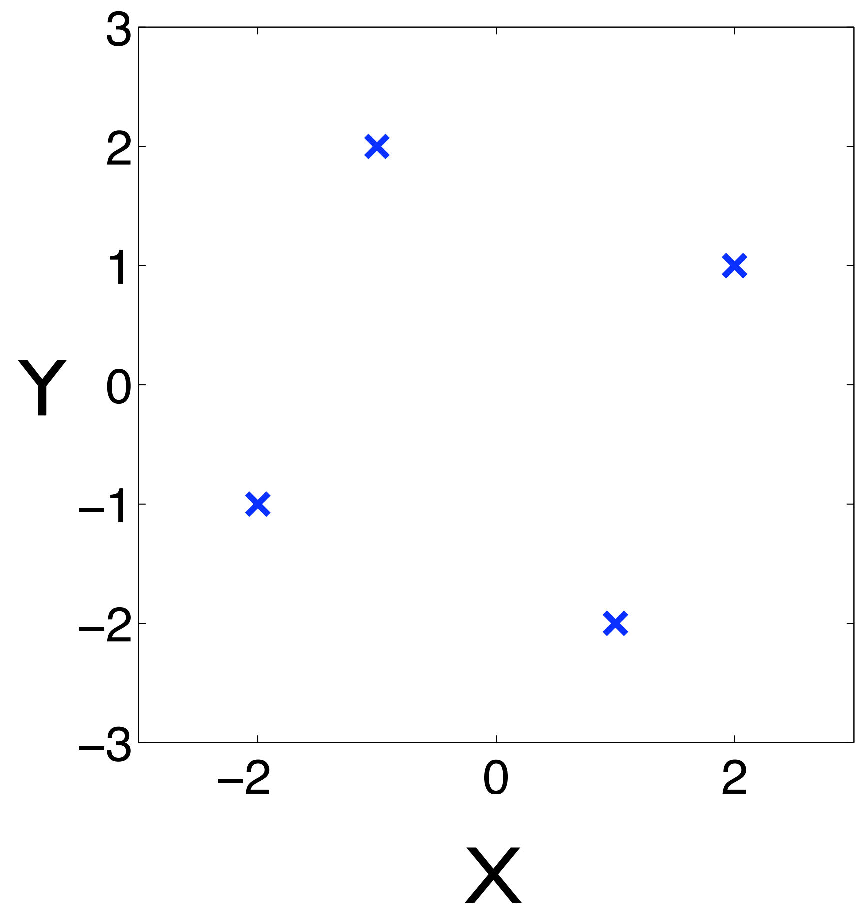
indep $\Rightarrow E(XY) = E(X)E(Y) = 0 \Rightarrow$ uncorr

- Do you think that all uncorrelated pairs of RVs are independent?

N

Correlation & independence

- Equal probability on each point
- Are X and Y independent?
- Are X and Y uncorrelated?



Law of iterated expectations

- For any two RVs, X and Y, we have:
  - $E_Y(E_X[X \mid Y]) = E(X)$
- Convention: note in subscript the RVs that are not yet conditioned on (in this E(.)) or marginalized away (inside this E(.))

Law of iterated expectations

- $E_X[X \mid Y] = \sum_x p(x|y)\, x$
- $E_Y(E_X[X \mid Y]) = \sum_y p(y)\, E_x[x|y]$

$$= \sum_y \sum_x \underbrace{p(y)\, p(x|y)}_{p(x,y)}\, x$$

$$= \sum_x p(x)\, x = E(x)$$

Bayes Rule

Rev. Thomas Bayes
1702–1761

- For any X, Y, C
  - P(X | Y, C) P(Y | C) = P(Y | X, C) P(X | C)
- Simple version (without context)
  - P(X | Y) P(Y) = P(Y | X) P(X)
  - more commonly, P(X | Y) = P(Y | X) P(X) / P(Y)
- Can be taken as definition of conditioning

Exercise

- You are tested for a rare disease, emacsitis—prevalence 3 in 100,000
- Your receive a test that is 99% ***sensitive*** and 99% ***specific***
  - sensitivity = P(yes | emacsitis) = 0.99
  - specificity = P(no | ¬emacsitis) = 0.99
- The test comes out ***positive***
- Do you have emacsitis? Probably not.

$$P(e \mid +) = P(+ \mid e)\,P(e) / P(+)$$
$$\approx .99 \cdot 3e\text{-}5 / .01$$
$$\approx .99 \cdot 3e\text{-}3$$

$$P(+) = P(+, e) + P(+, \bar{e})$$
$$\approx .99 \cdot 3e\text{-}5 + 1 \cdot .01$$
$$\approx .01$$

Revisit: weighted dice

- Fair dice: all 36 rolls equally likely
- Weighted: rolls summing to 7 more likely
- Data: 1-6 2-5

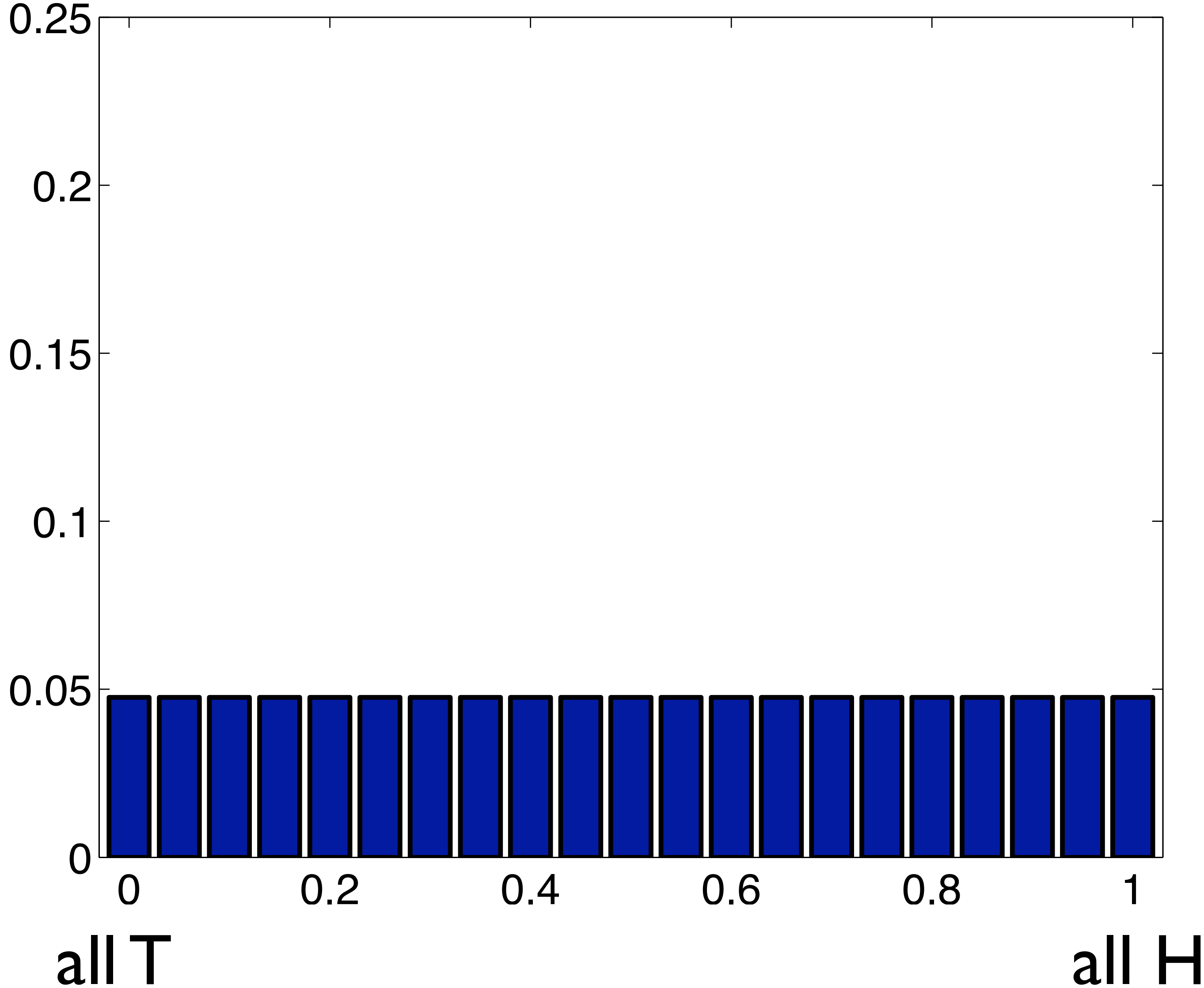
Learning from data

- Given a ***model class***
- And some data, sampled from a model in this class
- Decide which model best explains the sample

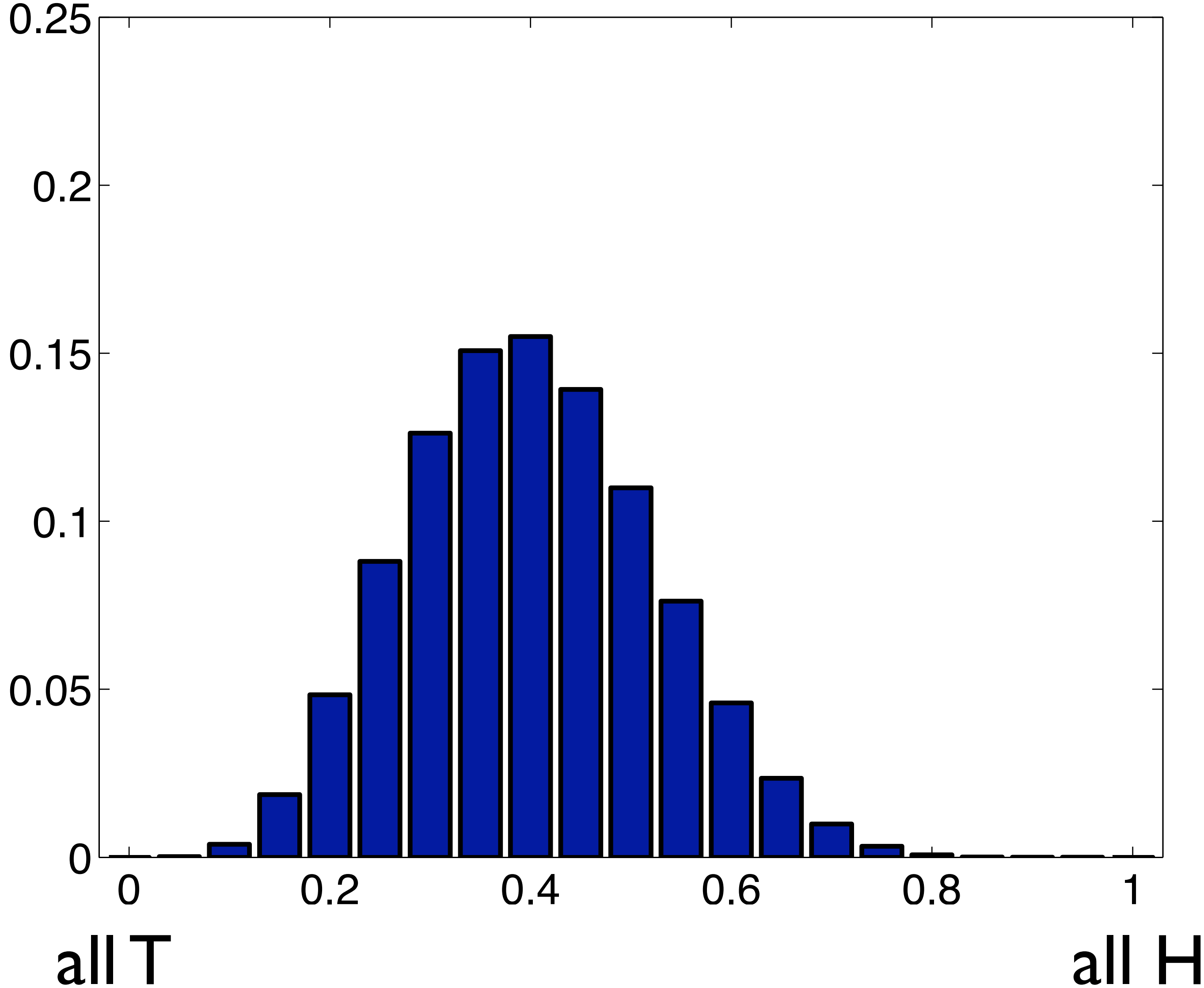
Bayesian model learning

- P(model | data) = P(data | model) P(model) / Z
- Z = P(data)
- So, for each model,
  - compute P(data | model) P(model)
  - normalize
- E.g., which parameters for face recognizer are best?
- E.g., what is P(H) for a biased coin?

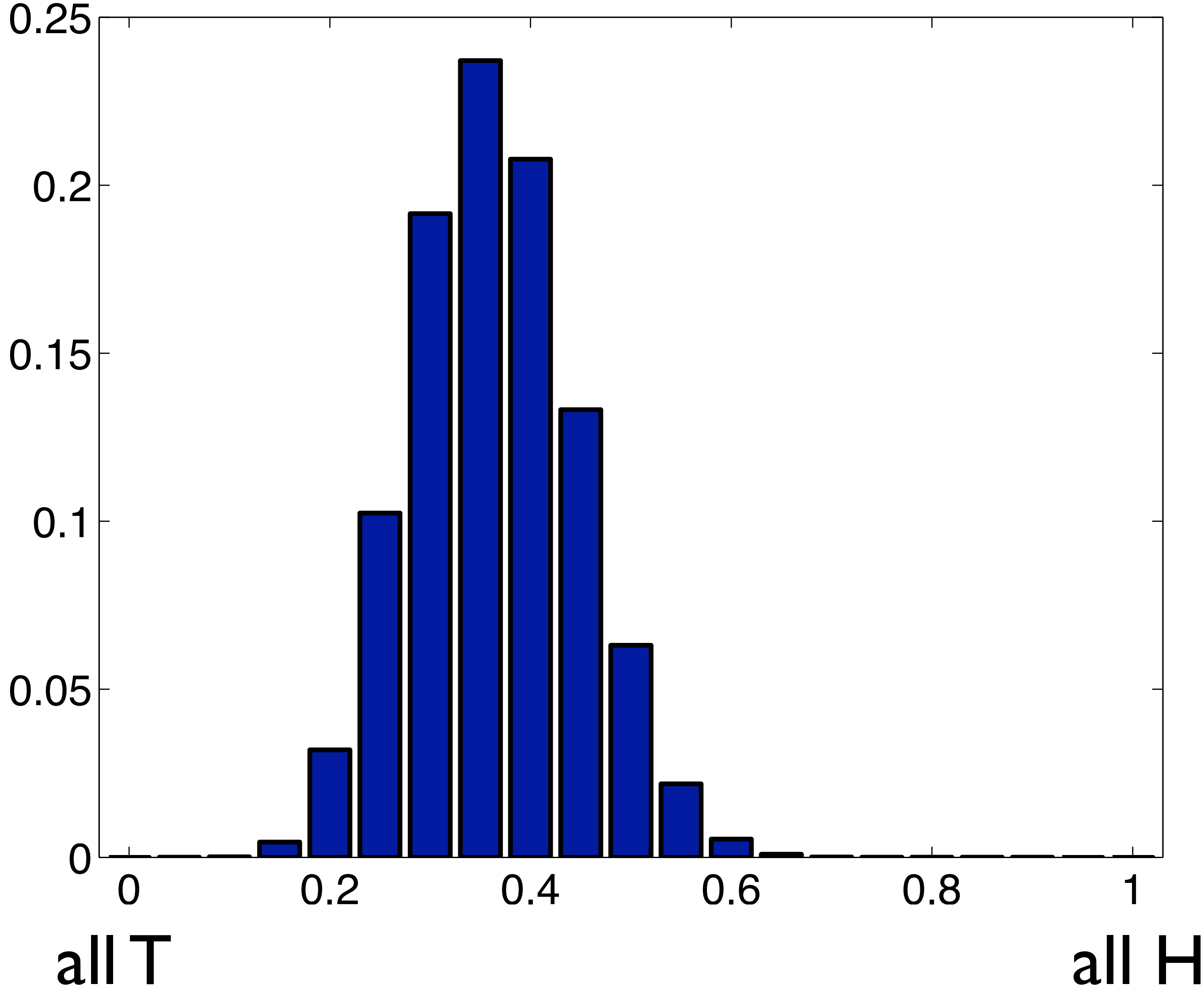
Prior: uniform



Posterior: after 5H, 8T



Posterior: 11H, 20T



Probability & AI

Why probability?

- Point of working with probability is to make ***decisions***
- E.g., find an open-loop ***plan*** or closed-loop ***policy*** with highest success probability or lowest expected cost
- Later: MDP, POMDP, …
- Now: simple motivating example
  - demonstrates that underlying problems are still familiar (related to SAT, PBI, MILP, #SAT)

Probabilistic STRIPS planning

- Same as ordinary STRIPS except each effect happens w/ (known, independent) probability

- Bake
  - pre: ¬have(Cake)
  - post: 0.8 have(Cake)
- Eat
  - pre: have(Cake)
  - post: ¬have(Cake), 0.9 eaten(Cake)

- Actions have no effect if ¬preconds
- Seek an (open-loop) plan with highest success probability

Translating to SAT-like problem

- Recall deterministic STRIPS → SAT:
  - $actA_{t+1} \Rightarrow preA1_t \wedge preA2_t \wedge \ldots$
  - $actA_{t+1} \Rightarrow postA1_{t+2} \wedge postA2_{t+2} \wedge \ldots$
  - $post_{t+2} \Rightarrow actA_{t+1} \vee actB_{t+1} \vee \ldots$
  - $goal1_T \wedge goal2_T \wedge \ldots$
  - $init1_1 \wedge init2_1 \wedge \ldots$
  - lots o' mutexes
- We need to modify 1–3 above, and handle maintenance and mutexes differently

Modified action constraints

- $[\text{actA}_{t+1} \wedge \text{preA1}_t \wedge \text{preA2}_t \wedge \ldots \wedge \text{gateA1}_t \Leftrightarrow \text{cA1}_{t+1}]$

  $\wedge\ \text{cA1}_{t+1} \Rightarrow \text{postA1}_{t+2}$
- $[\text{actA}_{t+1} \wedge \text{preA1}_t \wedge \text{preA2}_t \wedge \ldots \wedge \text{gateA2}_t \Leftrightarrow \text{cA2}_{t+1}]$

  $\wedge\ \text{cA2}_{t+1} \Rightarrow \text{postA2}_{t+2}$
- …
- $\text{pA1:gateA1}_t \wedge \text{pA2:gateA2}_t$

Modified literal constraints

- $lit_{t+2} \Rightarrow cA3_{t+1} \vee cB1_{t+1} \vee \ldots$
  $\vee [\neg c'A2_{t+1} \wedge \neg c'D5_{t+1} \wedge lit_t]$

Mutexes

- Need interference mutexes: if A deletes a precondition of B, $(\neg actA_t \vee \neg actB_t)$
- Other mutexes possible to generalize too (but we'll ignore, since they don't change semantics)

Example: causes for each postcondition

- $\neg have_1 \wedge gatebake_1 \wedge bake_2 \Leftrightarrow Cbake_2$
- $have_1 \wedge gateeat_1 \wedge eat_2 \Leftrightarrow Ceat_2$
- $have_1 \wedge eat_2 \Leftrightarrow Ceat'_2$
- $[Cbake_2 \Rightarrow have_3] \wedge [Ceat_2 \Rightarrow eaten_3] \wedge$

  $[Ceat'_2 \Rightarrow \neg have_3]$
- $0.8{:}gatebake_1 \wedge 0.9{:}gateeat_1$

Example: literal constraints

- $have_3 \Rightarrow [Cbake_2 \vee (\neg Ceat'_2 \wedge have_1)]$
- $\neg have_3 \Rightarrow [Ceat'_2 \vee (\neg Cbake_2 \wedge \neg have_1)]$
- $eaten_3 \Rightarrow [Ceat_2 \vee eaten_1]$
- $\neg eaten_3 \Rightarrow [\neg eaten_1]$

$\sim \neg Ceat_2$

Example: mutexes

- $\neg bake_2 \vee \neg eat_2$
- (pattern from past few slides is repeated for each pair of time slices)

Example: initial state and goals

- $\neg have_I \wedge \neg\ eaten_I$
- $have_T \wedge eaten_T$

Now what?

- Problem is to set decision variables so that, when random choices are set by Nature, P(formula satisfiable) is large
- I.e., if decision variables are X, Nature variables are Y, all other variables are Z, want:

$$\max_X \mathbb{E}_Y [\max_Z F(X, Y, Z)]$$

  - where F(X,Y, Z) is the formula we built on previous slides (with 1=true, 0=false)

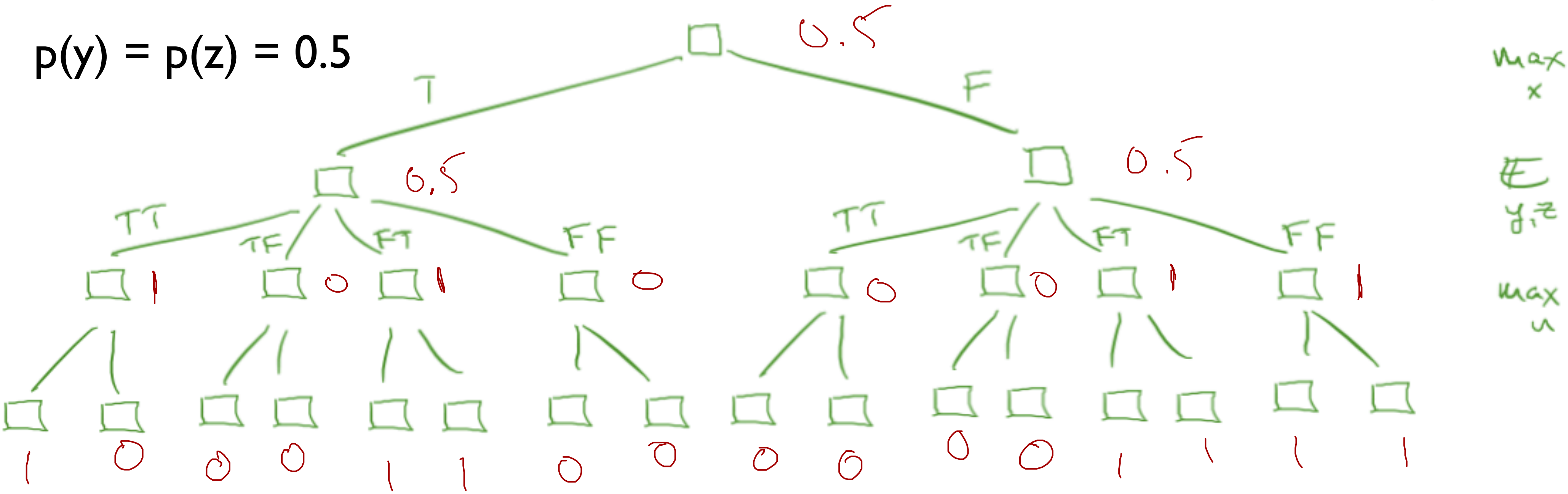
General class of problems

$$\mathbb{Q}_1 X_1 \ \mathbb{Q}_2 X_2 \ \mathbb{Q}_3 X_3 \ \ldots \ F(X_1, X_2, X_3, \ldots)$$

- where $\mathbb{Q}_i$ is max, min, or expectation
- Problem: test whether value ≥ threshold
- In general: difficulty determined by number of ***quantifier alternations***
- Contains QBF, so PSPACE-complete

Simpler example

p(y) = p(z) = 0.5



$$\max_x \; \mathbb{E}_{y,z} \; \max_u \; (\bar{x} \vee z) \wedge (\bar{y} \vee u) \wedge (x \vee \bar{y})$$

How can we solve?

- Scenario trick
  - transform to PBI or 0-1 ILP
- Dynamic programming
  - related to algorithms for SAT, #SAT
  - also to belief propagation in graphical models (next)