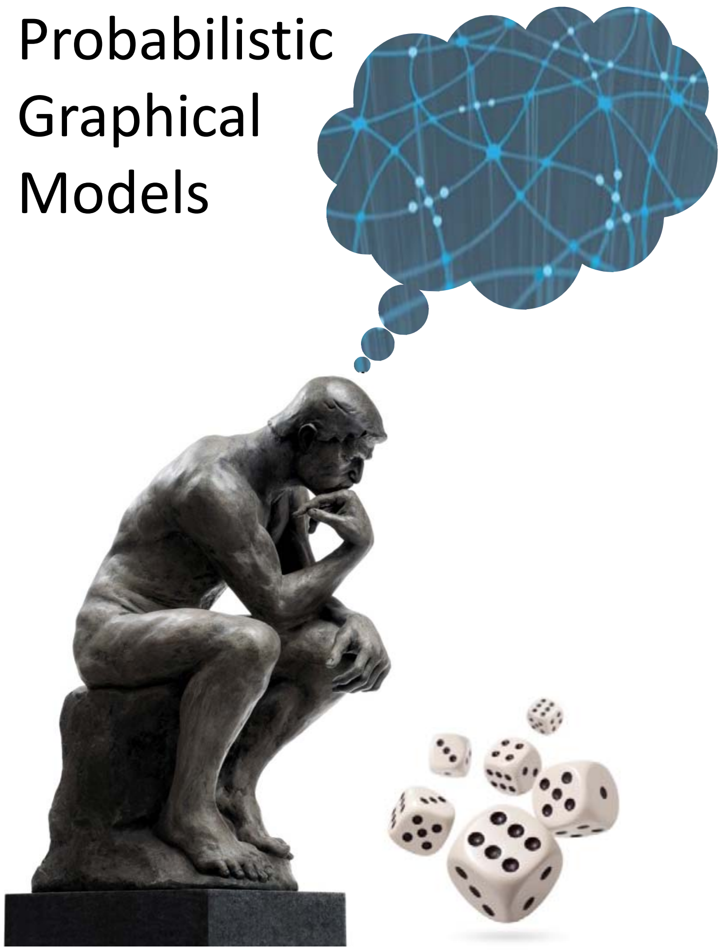
Probabilistic
Graphical
Models



Learning

Parameter Estimation

Maximum Likelihood Estimation

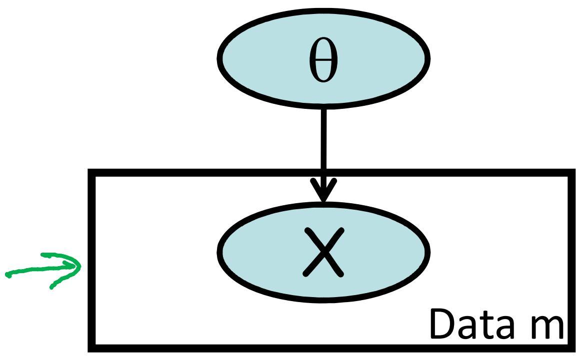
Biased Coin Example

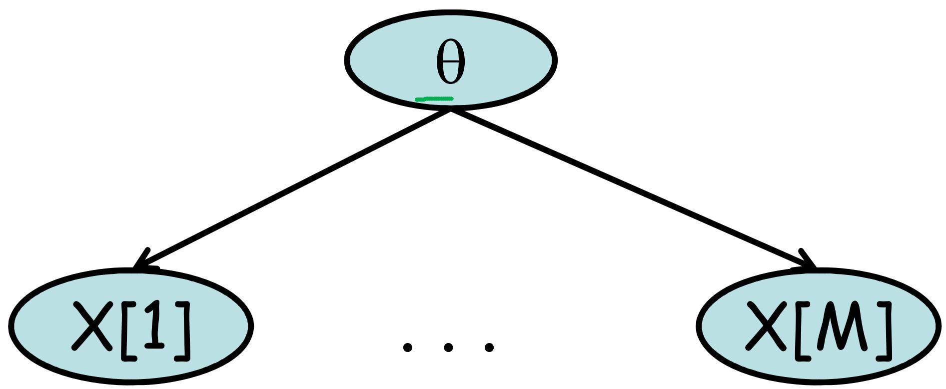
P is a Bernoulli distribution:
$P(X=1) = \theta$, $P(X=0) = 1-\theta$

$\mathcal{D} = \{x[1], \ldots, x[M]\}$ sampled IID from P

- Tosses are independent of each other
- Tosses are sampled from the same distribution (identically distributed)

IID as a PGM





$$P(x[m] \mid \theta) = \begin{cases} \theta & x[m] = x^1 \\ 1 - \theta & x[m] = x^0 \end{cases}$$

Maximum Likelihood Estimation

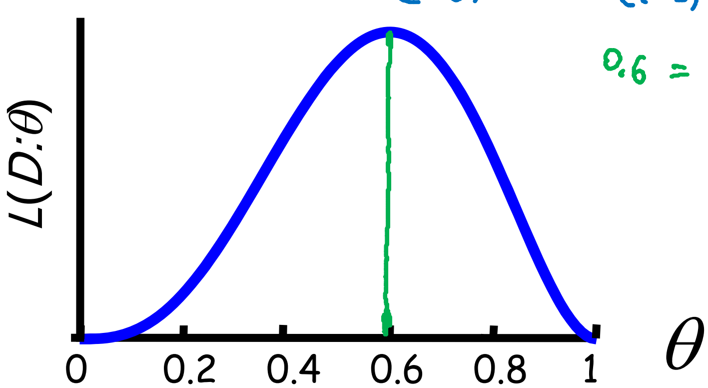
- Goal: find $\theta \in [0,1]$ that predicts D well
- Prediction quality = likelihood of D given θ

$$L(\theta : D) = P(D \mid \theta) = \prod_{m=1}^{M} P(x[m] \mid \theta)$$

$$L(\theta : \langle H, T, T, H, H \rangle)$$

$$\underbrace{P(H \mid \theta)}_{\theta} \cdot \underbrace{P(T \mid \theta)}_{(1-\theta)} \cdot \underbrace{P(T \mid \theta)}_{(1-\theta)} \cdot \underbrace{P(H \mid \theta)}_{\theta} \cdot \underbrace{P(H \mid \theta)}_{\theta} = \theta^3 (1-\theta)^2$$

$0.6 = \frac{3}{5}$



Maximum Likelihood Estimator

- Observations: M_H heads and M_T tails
- Find θ maximizing likelihood

$$L(\theta : M_H, M_T) = \theta^{M_H}(1-\theta)^{M_T}$$

log

- Equivalent to maximizing log-likelihood

$$l(\theta : M_H, M_T) = M_H \log \theta + M_T \log(1-\theta)$$

- Differentiating the log-likelihood and solving for θ: $\boxed{\hat{\theta} = \frac{M_H}{M_H + M_T}}$

Sufficient Statistics

- For computing θ in the coin toss example, we only needed M_H and M_T since

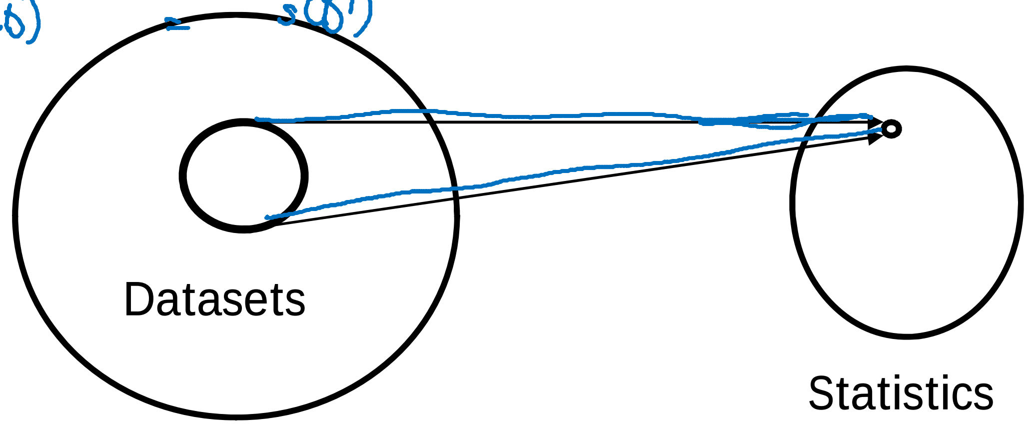
$$L(\theta : D) = \underbrace{\theta^{M_H}(1-\theta)^{M_T}}$$

- → M_H and M_T are sufficient statistics

Sufficient Statistics

- A function s(D) is a sufficient statistic from instances to a vector in $\Re^k$ if for any two datasets D and D′ and any $\theta \in \Theta$ we have

$$\underbrace{\sum_{x[i]\in D} s(x[i])}_{s(D)} = \underbrace{\sum_{x[i]\in D'} s(x[i])}_{s(D')} \quad \Rightarrow \quad L(\theta : D) = L(\theta : D')$$



Sufficient Statistic for Multinomial

- For a dataset D over variable X with k values, the sufficient statistics are counts $\langle M_1,\ldots,M_k \rangle$ where M_i is the # of times that $X[m]=x^i$ in D

- Sufficient statistic s(x) is a tuple of dimension k
  - $s(x^i)=(0,\ldots 0,1,0,\ldots,0)$ $\quad \sum_m s(x[m]) = (M_1, M_2, \ldots, M_k)$

i

$$L(\theta : D) = \prod_{i=1}^{k} \theta_i^{M_i}$$

param for $x = x^i$

Daphne Koller

Sufficient Statistic for Gaussian

- Gaussian distribution:

$$P(X) \sim N(\mu, \sigma^2) \quad \textit{if} \quad p(X) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

- Rewrite as

$$p(X) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-x^2 \frac{1}{2\sigma^2} + x\frac{\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right)$$

- Sufficient statistics for Gaussian: s(x)=<1,x,x^2>

$s(D) = \left(\sum_m x[m]^2, \sum_m x[m], M \right)$

Maximum Likelihood Estimation

- MLE Principle: Choose θ to maximize $L(D:\Theta)$

- Multinomial MLE: $$\hat{\theta}^i = \frac{M_i}{\sum_{i=1}^m M_i}$$ fraction of x^i in data

- Gaussian MLE: $$\hat{\mu} = \frac{1}{M}\sum_m x[m]$$ empirical mean
$$\hat{\sigma} = \sqrt{\frac{1}{M}\sum_m (x[m] - \hat{\mu})^2}$$ empirical st dev

Summary

- Maximum likelihood estimation is a simple principle for parameter selection given D
- Likelihood function uniquely determined by sufficient statistics that summarize D
- MLE has closed form solution for many parametric distributions