BN PARAMETER LEARNING

1. Setup:
   - Random Variables: \( X = \{X_1, \ldots, X_n\} \)
   - "True" Distribution: \( \sim P(X_1, \ldots, X_n) \)

   - Fully Observed Dataset of samples from \( P^* \)
     \[
     \begin{bmatrix}
     x_1^{(1)} & \cdots & x_n^{(1)} \\
     \vdots & \ddots & \vdots \\
     x_1^{(M)} & \cdots & x_n^{(M)} 
     \end{bmatrix}_{M \times n}
     \]

   - Known BN structure w/ assume \( P^* \) factorizes to \( G \)
     \( P^*(x_1, \ldots, x_n) = \prod_i P(x_i | Pa_i) \)

   Goal: estimate \( CP^* \)

   Tool: Frequentist: Maximum Likelihood
   - Bayesian: Maximum a Posteriori
     - full posterior estimation

2. Simplest BN: 1 RV: \( \{x\} \)
   \[
   CP^* P(X) = \begin{cases} 
   1 - \theta & \text{if } x = 0 / T \\
   \theta & \text{if } x = 1 / \# 
   \end{cases}
   \]

Note on next page from ECE 4984/5984
2. **Max Likelihood Estimation**

(Sample Space)

\[ D = \{ \text{Nadal Loses (L), Nadal Wins (W)} \} \]

Random Variable

\[ x \in \{ 0, 1 \} \]

\[ L, W \]

\[ T, H \]

\[ x \sim \text{Bernoulli} (\theta) \]

\[ P(x=1) = \theta \]

\[ P(x=0) = 1-\theta \]

Given:

\[ D = \{ 1, 0, 0, 1, 1 \} \] estimate \( \theta \in [0, 1] \)

Good \( \theta \) makes it likely for us to have observed \( D \)

E.g. if \( \theta = 0 \), we would never see 1?

Idea: let's maximize the probability of \( D \) under \( \theta \)

\[ \theta_{MLE} = \arg \max_{\theta \in [0,1]} P(D|\theta) \]

Called likelihood function

\[ L(\theta; D) = P(D|\theta) \]

IMP: Likelihood is a function of \( \theta \) (\( D \) is fixed to what we observed)

\[ L(\theta; D) = P(D|\theta) = \prod_{i=1}^{n} P(x_i; \theta) \] [Why? Hint: IID]
Example \( D = \{1\} \) \( \mathcal{L}(\theta) = 0 \)

\[
D = \{1, 1\} \quad \mathcal{L}(\theta) = \theta \theta = \theta^2
\]

\[
D = \{1, 0\} \quad \mathcal{L}(\theta) = \theta (1 - \theta)
\]

In general \( \mathcal{L}(\theta) = \theta (1 - \theta) \)

\( x_H = \# \text{Heads/Win} \)

\( x_T = \# \text{Tails/Loss} \)

\( \hat{\theta}_{MLE} = \arg \max_{\theta \in [0, 1]} \mathcal{L}(\theta) \)

\[ \text{called log-likelihood} \]

\[ = \arg \max_{\theta} \log \mathcal{L}(\theta) \] \[ \text{Why? \& log is a monotone function so preserves argmax} \]

How do we find \( \arg \max \) of \( \mathcal{L}(\theta) \)?

Take 1st derivative; set to zero

\[
\frac{\partial \mathcal{L}(\theta)}{\partial \theta} = \frac{2}{\theta} \left[ \frac{x_H}{\theta} + \frac{x_T}{1 - \theta} \right]
\]

\[ = \frac{x_H}{\theta} + \frac{x_T}{1 - \theta} \]
\[
\frac{\partial L(\theta)}{\partial \theta} - \frac{\alpha_H - \alpha_H \theta - \alpha_T \theta}{\theta (1-\theta)} = 0
\]

Ignoring boundary conditions,

\[
\theta_{\text{MLE}} = \frac{\alpha_H}{\alpha_H + \alpha_T}
\]

(3) Sufficient Statistics

\[D = \{1, 1, 1, 0, 0, 0\} \quad \alpha_T = \alpha_I = 3\]

\[D = \{1, 0, 1, 0, 1, 0\} \quad \alpha_H = \alpha_T = 3\]

Two datasets but look the same to likelihood.

4) MLE is OPT if model class is correct.

Consider \(\frac{1}{N} \sum_{i=1}^{n} \log P(x_i; \theta)\)

\[= \frac{1}{N} \left[ \#(x=1) P(x=1; \theta) + \#(x=2) P(x=2; \theta) + \ldots \right]
\]

(counting argument)
As data becomes infinite

\[ \lim_{N \to \infty} \frac{\#(X=1)}{N} = P(X=1 \mid \theta^*) \]

Shorthand:
\[
\begin{align*}
\text{P}^*(X) &= P(X \mid \theta^*) \\
P_0(X) &= P(X \mid \theta)
\end{align*}
\]

Now

\[
\frac{1}{N} \text{LL}(\theta) = \sum_{x=1}^{k} P^*(x) \log P_0(x)
\]

\[
= \sum_{x=1}^{k} P^*(x) \log \left[ \frac{P_0(x) \cdot P^*(x)}{P^*(x)} \right]
\]

\[
= \sum_{x=1}^{k} P^*(x) \log P^*(x) - \sum_{x=1}^{k} P^*(x) \log \frac{P^*(x)}{P_0(x)}
\]

\[
\frac{1}{N} \text{LL}(\theta) = -H(p^*) - KL(p^* \parallel p_0)
\]

So

\[
\max_{\theta} \frac{1}{N} \text{LL}(\theta) = \min_{\theta} KL(p^* \parallel p_0)
\]

[\text{a constant}]

POWERFUL RESULT: We did not specify \( P(X \mid \theta) \)
\text{Any distribution}

Consept: Inf data
\text{We must know the family } P_0 \text{, which we usually don't (e.g. is life Gaussian?)}
MAP + Bayesian Estimation

Let's think of $\Theta$ as a random quantity & apply Bayes' Rule:

$$P(\Theta | D) = \frac{P(D | \Theta) P(\Theta)}{P(D)}$$

$P(\Theta)$: what do we believe about $\Theta$ without any data?

: Prior Belief

$\Theta \in [0, 1]$ So we need a distribution over parameters of our distribution

**Beta distribution**

$$P(\Theta | B_+, B_r) = \Theta^{B_+-1} (1 - \Theta)^{B_r-1}$$

Hyper-parameters: parameters of the distribution

One parameter ($\Theta$)

**Important Facts:**

$$\text{constant} = \int_0^1 \Theta^{B_+-1} (1 - \Theta)^{B_r-1} d\Theta$$

Mode of distribution:

$$\text{mode of distribution} = \frac{B_+-1}{B_++B_r-2}$$
Maximum A Posteriori (MAP) Estimation

\[ \hat{\theta}_{MAP} = \arg \max \_ \theta \ P(\theta | D) \]

\[ = \arg \max \_ \theta \ \frac{P(D | \theta) P(\theta)}{P(D)} \]

\[ = \arg \max \_ \theta \ P(D | \theta) P(\theta) \]

\[ = \arg \max \_ \theta \ P(\theta) (1 - \theta) + \theta (1 - \theta) \]

\[ = \arg \max \_ \theta \ \theta \ (1 - \theta) \]

\[ = \text{mode of } \text{Beta}(\alpha_H + \beta_H, \alpha_T + \beta_T) \]

\[ = \frac{\alpha_H + \beta_T - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2} \]

Very nice, so \( \beta_H, \beta_T \) act as pseudo-flips/covers -> Phantom experiments not contained in our data!
Special Cases: \( \beta_h = \beta_T = 1 \)

\[
\hat{\theta}_{\text{MAP}} = \frac{\alpha + 1 - 1}{\alpha + 1 \cdot \beta + 1 - 2} = \hat{\theta}_{\text{MLE}}
\]

When \( \beta_h = \beta_T = 1 \)

\[
\beta \theta = \theta (1-\theta)
\]

uniform distribution

No prior \( \Rightarrow \) posterior = prior

\[ \longleftrightarrow \]

When \( N \to \infty \) effect of \( \beta_h, \beta_T \) is forgotten.
All ideas extend to general BN CPT estimation.

Example: $G = \begin{array}{c} F \\ \downarrow \\ S \\ \downarrow \\ N \end{array}$

Variables: $F, A, S, N$ (all binary)

CPTs:

$P(F) = \begin{bmatrix} 1 - \Theta_F \\ \Theta_F \end{bmatrix}$, $P(A) = \begin{bmatrix} 1 - \Theta_A \\ \Theta_A \end{bmatrix}$

$P(S|F, A) = \begin{array}{cccc|cccc} & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ \hline \Theta_{S|FF} & \Theta_{S|FO} & \Theta_{S|FO} & \Theta_{S|FO} & \Theta_{S|OA} & \Theta_{S|OA} & \Theta_{S|OA} & \Theta_{S|OA} \end{array}$

$P(N|S) = \begin{bmatrix} 1 - \epsilon & 1 - \epsilon \\ \Theta_{ON0} & \Theta_{ON1} \end{bmatrix}$

Parameters to be estimated: $\Theta = \{ \Theta_F, \Theta_A, \Theta_{S|00}, \Theta_{S|01}, \Theta_{S|10}, \Theta_{S|11}, \Theta_{ON0}, \Theta_{ON1} \}$

MLE: $\hat{\Theta}_{MLE} = \arg\max_\Theta \log P(CD|I, \Theta, G)$

$= \arg\max_\Theta \sum_{i=1}^M \log P(F = f^{(i)}, A = a^{(i)}, S = s^{(i)}, N = n^{(i)} | \Theta, G)$

$= \arg\max_\Theta \sum_{i=1}^M \left[ \log P(F = f^{(i)} | \Theta, G) + \log P(A = a^{(i)} | \Theta, G) + \log P(S = s^{(i)} | F = f^{(i)}, A = a^{(i)}, \Theta, G) + \log P(N = n^{(i)} | S = s^{(i)}, \Theta, G) \right]$
Same argument as for a single variable.

In general
\[ \Theta_{X_i=a \mid P_{X_i}=b} = \frac{\text{Count}(X_i=a, P_{X_i}=b)}{\text{Count}(P_{X_i}=b)} \]

For multi-label variables \( X_i, x_i \), and just a bit more math (Lagrange multipliers):
\[ \min_{\Theta_{X_i}} \sum_{x_i} \Theta_{X_i=x_i \mid P_{X_i}=b} = 1 \quad \text{correlated optimization, but same result as binary case.} \]

For MAP estimation
\[ \text{assume } P(\Theta_{X_i \mid P_{X_i}=b}) \sim \text{Dir}(Z_{X_i \mid P_{X_i}=b}) \]

Each CPT column has its own prior given by hyperpriors.