Probabilistic Graphical Models
Formula Sheet

Probability Theory

Chain Rule for Probabilities:
$P(X_1, \ldots, X_n) = P(X_1) \cdots P(X_n | X_1, X_{n-1})$

Bayes Rule:
$P(\alpha \mid \beta) = \frac{P(\beta \mid \alpha)}{P(\alpha)P(\beta)}$

Probability Density Function: $p : \mathbb{R} \to \mathbb{R}$ is a probability density function (PDF) for $\mathcal{X}$ if it is a non-negative integrable function such that:
$\int_{Val(X)} p(x)\, dx = 1$.

Uniform Distribution: $X \sim \text{Unif}[a, b]$ if it has the PDF:
$$p(x) = \begin{cases} \frac{1}{b-a} & b \geq x \geq a \\ 0 & \text{otherwise.} \end{cases}$$

Gaussian Distribution: X has a Gaussian distribution: $X \sim \mathcal{N}(\mu; \sigma^2)$ if it has the PDF:
$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.

Joint Density Function: Let P be a joint distribution over $X_1, \ldots, X_n$. A function $p(x_1, \ldots, x_n)$ is a joint density function of $X_1, \ldots, X_n$ if:

1. $p(x_1, \ldots, x_n) \geq 0 \ \forall x_1, \ldots, x_n \in X_1, \ldots, X_n$.
2. p is integratable.
3. For any choice of $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$:

$$P(a_1 \leq X_1 \leq b_1, \ldots, a_n \leq X_n \leq b_n)$$
$$= \int_{a_1}^{b_1} \ldots \int_{a_n}^{b_n} p(x_1, \ldots, x_n)\, dx_1 \ldots\, dx_n$$

Conditional Density Function: Suppose you would like to condition over the event: $x - \epsilon \leq X \leq x + \epsilon$. Then $P(Y \mid x) = \lim_{\epsilon \to 0} P(Y \mid x - \epsilon \leq X \leq x + \epsilon)$. If there is a continuous joint density function p(x,y) then
$= P(a \leq Y \leq b \mid x - \epsilon \leq X \leq x + \epsilon)$
$= \frac{P(a \leq Y \leq b, x - \epsilon \leq X \leq x + \epsilon)}{P(x - \epsilon \leq X \leq x + \epsilon)} = \frac{\int_a^b \int_{x-\epsilon}^{x+\epsilon} p(x', y)\, dy dx'}{\int_{x-\epsilon}^{x+\epsilon} p(x')\, dx'}$

Expectation of X under P is:
$\mathbb{E}_P[X] = \sum_x x.P(x)$.

Expectation if X is Continuous:
$\mathbb{E}_P[X] = \int x.p(x)\, dx$.

Linearity of Expectation:
$\mathbb{E}_P[X + Y] = \mathbb{E}_P[X] + \mathbb{E}_P[Y]$.

Conditional Expectation:
$\mathbb{E}_P[X \mid \mathbf{y}] = \sum_x x.P(x \mid \mathbf{y})$.

Variance of X:
$\mathbb{V}ar_P[X] = \mathbb{E}_P[(X - \mathbb{E}_P[X])^2]$.

Standard Deviation:
$\sigma_X = \sqrt{\mathbb{V}ar_P[X]}$.

Expectation and Variance of Gaussian distribution $X \sim \mathcal{N}(\mu; \sigma^2)$, then $\mathbb{E}[X] = \mu$ and $\mathbb{V}ar[X] = \sigma^2$.

Graph Theory

A **Graph** is a data structure $\mathcal{K} = (\mathcal{X}, \mathcal{E})$ consisting of a set of nodes, denoted $\mathcal{X} = X_1, \ldots, X_n$, and edges, denoted $\mathcal{E}$.

Induced Subgraph: Let $\mathcal{K} = (\mathcal{X}, \mathcal{E})$, and $\mathbf{X} \in \mathcal{X}$, then an induced subgraph, denoted $\mathcal{K}[\mathbf{X}]$ is a graph $(\mathbf{X}, \mathcal{E}')$ where $\mathcal{E}'$ are all the edges $X \leftrightharpoons Y \in \mathcal{E}'$ such that $X, Y \in \mathbf{X}$.

Complete Graph (Clique): A subgraph over $\mathbf{X}$ is complete if every two nodes in $\mathbf{X}$ are connected by some edge. The set $\mathbf{X}$ is called a clique. A clique $\mathbf{X}$ is maximal if for any superset of nodes $\mathbf{Y} \supset \mathbf{X}$, $\mathbf{Y}$ is not a clique.

Upward Closure: A subset of nodes $\mathbf{X} \in \mathcal{X}$ is upwardly closed in $\mathcal{K}$ if, for any $\mathbf{X} \in \mathcal{X}$, we have that the $\text{Boundary}_X \subset \mathbf{X}$. We define upward closure of $\mathbf{X}$ to be the minimally upward closed subset $\mathbf{Y}$ that contains $\mathbf{X}$.

Topological ordering: An ordering of the nodes $X_1, \ldots, X_n$ is a topological ordering if when we have $(X_i \to X_j) \in \mathcal{E}$, then $i < j$.

Chordal Graph: Let $X_1 - X_2 - \ldots - X_k - X_1$ be a loop in a graph. A chord in a loop is an edge connecting X_i and X_j for two nonconsecutive nodes X_i, X_j. An undirected graph $\mathcal{H}$ is said to be chordal if and loop $X_1 - X_2 - \ldots - X_k - X_1$ for $k > 4$ has a chord. A directed graph $\mathcal{K}$ is said to be chordal if its underlying undirected graph is chordal.

Bayesian Networks

Naïve Bayes:
$P(C, X_1, \ldots, X_n) = P(C) \prod_{i=0}^n P(X_i \mid C)$

Bayesian Network:
$P(X_1, \ldots, X_n) = \prod_{i=1}^n P(X_i \mid \mathtt{Pa}_{X_i}^{\mathcal{G}})$

Deterministic CPD: $f : Val(Pa_X) \mapsto Val(X)$ s.t.:
$$P(x \mid pa_x) = \begin{cases} 1 & \text{if } x = f(pa_X) \\ 0 & \text{if } x \ \mathtt{otherwise} \end{cases}$$

Time Granularity Assumption:
$P(\mathcal{X}^{(0:T)}) = P(\mathcal{X}^{(0)}) \prod_{t=0}^{T-1} P(\mathcal{X}^{(t+1)} \mid \mathcal{X}^{(0:t)})$

Markov Assumption:
$P(\mathcal{X}^{(0:T)}) = P(\mathcal{X}^{(0)}) \prod_{t=0}^{T-1} P(\mathcal{X}^{(t+1)} \mid \mathcal{X}^{(t)})$

Time Invariance Assumption:
$P(\mathcal{X}^{(t+1)} = \xi' \mid \mathcal{X}^{(t)} = \xi) = P(\mathcal{X}' = \xi' \mid \mathcal{X} = \xi)$

Two-TBN:
$P(\mathcal{X}' \mid \mathcal{X}) = P(\mathcal{X}' \mid \mathcal{X}_I) = \prod_{i=1}^n P(X_i' \mid Pa_{X_i'})$

Linear Dynamical Systems:
$P(\mathbf{X}^{(t)} \mid \mathbf{X}^{(t-1)}) = \mathcal{N}(A\mathbf{X}^{(t-1)}; Q)$
$P(O^{(t)} \mid \mathbf{X}^{(t)}) = \mathcal{N}(H\mathbf{X}^{(t)}; R)$

Gibbs Distribution: A distribution P_Φ is a Gibbs distribution parameterised by a set of factors $\Phi = \{\phi_1(\mathbf{D_1}), \ldots, \phi_K(\mathbf{D_k})\}$ if it is defined as:
$P_\Phi(X_1, \ldots, X_n) = \frac{1}{Z} \tilde{P}_\Phi(X_1, \ldots, X_n)$

Inference

Inference:
$P(\mathbf{Y} \mid \mathbf{E} = \mathbf{e}) = \frac{P(\mathbf{Y}, \mathbf{e})}{P(\mathbf{e})} = \frac{\sum_w P(\mathbf{y}, \mathbf{e}, \mathbf{w})}{\sum_{y,w} P(\mathbf{e})}$

Sum-Product Message Passing:
$\delta_{i \to j} = \sum_{\mathbf{C}_i - \mathbf{S}_{i,j}} (\psi_i \times \prod_{k \in (Nb_i - \{j\})} \delta_{k \to i})$

Tree Calibration:
$\sum_{\mathbf{C}_i - \mathbf{S}_{i,j}} \beta_i(\mathbf{C}_i) = \sum_{\mathbf{C}_j - \mathbf{S}_{i,j}} \beta_j(\mathbf{C}_j)$

Graph Calibration:
$\sum_{\mathbf{C}_i - \mathbf{S}_{i,j}} \beta_i = \sum_{\mathbf{C}_j - \mathbf{S}_{i,j}} \beta_j$

MAP:
$\text{MAP}(\mathbf{Y} = \mathbf{y} \mid \mathbf{E} = \mathbf{e})$
$= \text{argmax}_y\, P(\mathbf{Y} = \mathbf{y} \mid \mathbf{E} = \mathbf{e})$

Convergence Bound:
$\hat{\mathbb{E}}_{\mathcal{D}}(f) = \frac{1}{M} \sum_{m=1}^M f(\xi[m])$.

Hoeffding Bound:
$P_{\mathcal{D}}(\hat{P}(\mathbf{y}) \notin [P(\mathbf{y}) - \epsilon, P(\mathbf{y}) + \epsilon]) \leq 2e^{-2M\epsilon^2}$

Chernoff Bound:
$P_{\mathcal{D}}(\hat{P}(\mathbf{y}) \notin [P(\mathbf{y})(\pm\epsilon)]) \leq 2e^{-MP(\mathbf{y})\epsilon^2/3}$
$M \geq 3 \frac{\ln(2/\delta)}{P(\mathbf{y})\epsilon^2}$.

Likelihood Weighting:
$\hat{P}_{\mathcal{D}}(\mathbf{y} \mid \mathbf{e}) = \frac{\sum_{m=1}^{M} w[m] \mathbb{1}\{\mathbf{y}[m]=\mathbf{y}\}}{\sum_{m=1}^{M} w[m]}$.

MCMC Sampling:
$P^{(t+1)}(\mathbf{X}^{(t+1)} = \mathbf{x'}) = \sum_{\mathbf{x} \in Val(\mathbf{X})} P^{(t)}(\mathbf{X}^{(t)} = \mathbf{x}) \mathcal{T}(\mathbf{x} \to \mathbf{x'})$

Stationary Distribution:
$\pi(\mathbf{X} = \mathbf{x'}) = \sum_{\mathbf{x} \in Val(\mathbf{X})} \pi(\mathbf{X} = \mathbf{x}) \mathcal{T}(\mathbf{x} \to \mathbf{x'})$

Detailed Balance Equation:
$\pi(x)\mathcal{T}(x \to x') = \pi(x')\mathcal{T}(x' \to x)$

Acceptance Probability:
$\mathcal{A}(x \to x') = \min[1, \frac{\pi(x')\mathcal{T}^{Q}(x' \to x)}{\pi(x)\mathcal{T}^{Q}(x \to x')}]$.

Metropolis-Hastings Acceptance Probability:
$\mathcal{A}(x_{-i}, x_i \to x_{-i}, x_i') = \min[1, \frac{P_\Phi(x_i', x_{-i})\mathcal{T}_i^{Q}(x_{-i}, x_i' \to x_{-i}, x_i)}{P_\Phi(x_i, x_{-i})\mathcal{T}_i^{Q}(x_{-i}, x_i \to x_{-i}, x_i')}]$.

Learning

Relative Entropy:
$\mathbb{D}(P^* \mid\mid \tilde{P}) = \mathbb{E}_{\xi \sim P^*}[\log(\frac{P^*(\xi)}{\tilde{P}(\xi)})]$,

Negative Empirical Log-loss:
$\log P(\mathcal{D} : \mathcal{M}) = \sum_{m=1}^{M} \log P(\xi[m] : \mathcal{M})$.

Bayesian Parameter Estimation:
$P(\theta \mid x[1], \ldots, x[M]) = \frac{P(x[1],\ldots,x[M] \mid \theta) P(\theta)}{P(x[1],\ldots,x[M])}$

Expected Sufficient Statistics:
$\bar{M}_\theta[\mathbf{y}] = \sum_{m=1}^{M} \sum_{\mathbf{h}[m] \in Val(\mathbf{H}[m])} Q(\mathbf{h}[m]) \mathbb{1}\{\xi[m]\langle \mathbf{Y} \rangle = \mathbf{y}\}$

Maximisation of Expected Parameter:
$\tilde{\theta}_{d^1|c^0} = \frac{\bar{M}_\theta[d^1, c^0]}{\bar{M}_\theta[c^0]}$

Bayesian Clustering:
$\bar{M}_\theta[c] = \frac{\bar{M}_\theta[c]}{M}$
$\bar{M}_\theta[x_i \mid c] = \frac{\bar{M}_\theta[x_i, c]}{\bar{M}_\theta[c]}$

K-means Clustering:
$c[m] = \text{argmax}_c P(c \mid x[m], \theta^t)$

Hypothesis Testing:
$d_{\mathbb{I}}(\mathcal{D}) = \sum_{x,y} \frac{M[x,y]}{M} \log \frac{M[x,y]/M}{M[x]/M \cdot M[y]/M}$
$\mathrm{R}_{d,t}(\mathcal{D}) \begin{cases} \text{Accept} & \text{if } d(\mathcal{D}) \le t \\ \text{Reject} & \text{if } d(\mathcal{D}) > t \end{cases}$
$\text{p-value}(t) = P(\{\mathcal{D} : d(\mathcal{D}) > t\} \mid H_0, M)$

Likelihood:
$\mathbb{I}_{\hat{P}_{\mathcal{D}}}(X_i; Pa_{X_i}^{\mathcal{G}}) = \sum_{\mathbf{u}_i} \sum_{\mathbf{x}_i} \hat{P}(x_i, \mathbf{u}_i) \log \frac{\hat{P}(x_i, \mathbf{u}_i)}{\hat{P}(x_i)\hat{P}(\mathbf{u}_i)}$

Entropy:
$\mathbb{H}_{\hat{P}_{\mathcal{D}}}(X_i) = \sum_{x_i} \hat{P}(x_i) \log \frac{1}{\hat{P}(x_i)}$

Bayesian Structure Learning:
$P(\mathcal{G} \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \mathcal{G}) P(\mathcal{G})}{P(\mathcal{D})}$
$\text{score}_B(\mathcal{G} : \mathcal{D}) = \log P(\mathcal{D} \mid \mathcal{G}) + \log P(\mathcal{G})$
$P(\mathcal{D} \mid \mathcal{G}) = \int_{\Theta_{\mathcal{G}}} P(\mathcal{D} \mid \boldsymbol{\theta}_{\mathcal{G}}, \mathcal{G}) P(\boldsymbol{\theta}_{\mathcal{G}} \mid \mathcal{G}) d\boldsymbol{\theta}_{\mathcal{G}}$

Marginal Likelihood for Binomials:
$P(x[1], \ldots, x[M])$
$= P(x[1]) \cdot \ldots \cdot P(x[m] \mid x[1], \ldots, x[M-1])$

Marginal Likelihood for Multinomials:
$P(x[1], \ldots, x[M]) = \frac{\Gamma(\alpha)}{\Gamma(\alpha+M)} \cdot \prod_{i=1}^{k} \frac{\Gamma(\alpha_i + M[x^i])}{\Gamma(\alpha_i)}$

Bayesian Score:
$P(\mathcal{D} \mid \mathcal{G}) = \prod_i \prod_{\mathbf{u}_i \in Val(Pa_{X_i}^{\mathcal{G}})} \frac{\Gamma(\alpha_{X_i|\mathbf{u}_i}^{\mathcal{G}})}{\Gamma(\alpha_{X_i|\mathbf{u}_i}^{\mathcal{G}} + M[\mathbf{u}_i])} \cdot \prod_{\mathbf{x}_i^j \in Val(X_i)} \left[\frac{\Gamma(\alpha_{x_i^j|\mathbf{u}_i}^{\mathcal{G}} + M[x_i^j, \mathbf{u}_i])}{\Gamma(\alpha_{x_i^j|\mathbf{u}_i}^{\mathcal{G}})} \right]$

BIC Score:
$\text{score}_{BIC}(\mathcal{G} : \mathcal{D}) = M \sum_{i=1}^{n} \mathbb{I}_{\hat{P}_{\mathcal{D}}}(X_i; Pa_{X_i}^{\mathcal{G}}) - \frac{\log M}{2} \dim[\mathcal{G}]$

Decomposibility:
$\text{score}(\mathcal{G} : \mathcal{D}) = \sum_i \text{FamScore}(X_i \mid Pa_{X_i}^{\mathcal{G}} : \mathcal{D})$

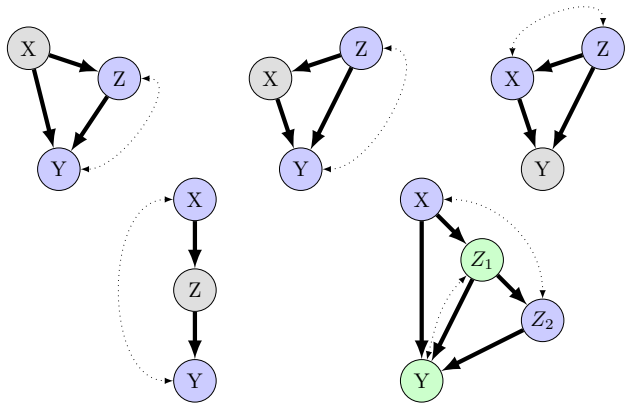
Tree weight:
$w_{i \to j} = \text{FamScore}(X_i \mid X_j : \mathcal{D}) - \text{FamScore}(X_i : \mathcal{D})$

Learning Graphs:
$\mathcal{G}^* = \text{argmax}_{\mathcal{G} \in \boldsymbol{\mathcal{G}}} \text{score}(\mathcal{G} : \mathcal{D})$

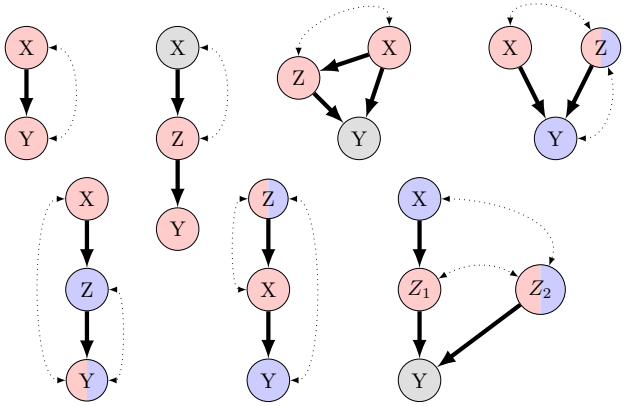
Causality

Intervention Query:
$P_{\mathcal{C}}(\mathbf{Y} \mid do(z), \mathbf{x}) = P_{\mathcal{C}_{\mathbf{Z}=\mathbf{z}}}(\mathbf{Y} \mid \mathbf{x})$

Identifiable when $P(Y \mid do(X))$:



Not Identifiable when $P(Y \mid do(X))$:



Learning with Intervention Data:
$P(\xi \mid do(\mathbf{Z} := \mathbf{z}), \mathcal{C}) = \prod_{X_i \notin \mathbf{Z}} P(x_i \mid \mathbf{u}_i)$

Sufficient Statistics (Intervention Data):
$M[x_i; \mathbf{u}_i] = \sum_{m : X_i \notin \mathbf{Z}[m]} \mathbb{1}\{X_i[m] = x_i, Pa_{X_i}[m] = \mathbf{u}_i\}$

Likelihood of Data (Intervention):
$L(\mathcal{C} : \mathcal{D}) = \prod_{i=1}^{n} \prod_{x_i \in Val(X_i), \mathbf{u}_i \in Val(Pa_{X_i})} \theta_{x_i|\mathbf{u}_i}^{M[x_i;\mathbf{u}_i]}$

Decision Theory

Expected Utility:
$\text{EU}[D[a]] = \sum_{\mathbf{x}} P(\mathbf{x} \mid a) U(\mathbf{x}, a)$

Maximum Expected Utility:
$a^* = \text{argmax}_a \text{EU}[D[a]]$
$= \text{argmax}_a \sum_{\mathbf{x}} P(\mathbf{x} \mid a) U(\mathbf{x}, a)$

Expected Utility with Information:
$\text{EU}[D[\delta_A]] = \sum_{\mathbf{x}, a} P_{\delta_A}(\mathbf{x}, a) U(\mathbf{x}, a)$

Maximal Expected Utility (MEU) Strategy:
$\text{argmax}_{\delta_{D_1}, \ldots, \delta_{D_k}} \text{EU}[\mathcal{I}[\delta_{D_1}, \ldots, \delta_{D_k}]]$

Value of Information:
$\text{VPI}(A \mid X) := \text{MEU}(D_{X \to A}) - \text{MEU}(D)$