with a subset of variables $c_k \in \mathcal{C}$ (clusters), and each edge $(kl) \in \mathcal{E}$ a subset $s_{kl} \in \mathcal{S}$ (separator) satisfying $s_{kl} \subseteq c_k \cap c_l$. We assume that $\mathcal{C}$ subsumes the index set $\mathcal{I}$, that is, for any $\alpha \in \mathcal{I}$, there exists a $c_k \in \mathcal{C}$, denoted $c_\alpha$, such that $\alpha \subseteq c_k$. In this case, we can reparameterize $\theta = \{\theta_\alpha | \alpha \in \mathcal{I}\}$ into $\theta = \{\theta_k | k \in \mathcal{V}\}$ by taking $\theta_k = \sum_{c_\alpha : c_\alpha = c_k} \theta_\alpha$, without changing the distribution. A cluster graph is called a junction graph if it satisfies the running intersection property – for each $i \in \mathcal{V}$, the induced sub-graph consisting of the clusters and separators that include $i$ is a connected tree. A junction graph is a junction tree if $G$ is tree.

To approximate the dual (1), we can replace $M$ with a locally consistent polytope $L$: the set of local marginals $\tau = \{\tau_{c_k}, \tau_{s_{kl}} : k \in \mathcal{V}, (kl) \in \mathcal{E}\}$ satisfying $\sum_{x_{c_k} \setminus c_k} \tau_{c_k}(x_{c_k}) = \tau(x_{s_{kl}})$. Clearly, $M \subseteq L$. We then approximate (1) by

$$\max_{\tau \in L} \{\theta, \tau\} + \sum_{k \in \mathcal{V}} H(x_{c_k}; \tau_{c_k}) - \sum_{(kl) \in \mathcal{E}} H(x_{s_{kl}}; \tau_{s_{kl}}),$$

where the joint entropy is approximated by a linear combination of the entropies of local marginals. The approximate objective can be solved using Lagrange multipliers [Yedidia et al., 2005], leading to a sum-product message passing algorithm that iteratively sends messages between neighboring clusters via

$$m_{k \rightarrow l}(x_{c_k}) \propto \sum_{x_{k} \setminus c_k} \psi_{c_k}(x_{c_k}) m_{\sim k \setminus l}(x_{c_k \setminus N(k)}),$$

where $\psi_{c_k} = \exp(\theta_{c_k})$, and $m_{\sim k \setminus l}$ is the product of messages into $k$ from its neighbors $N(k)$ except $l$. At convergence, the (locally) optimal marginals are

$$\tau_{c_k} \propto \psi_{c_k} m_{\sim k} \quad \text{and} \quad \tau_{s_{kl}} \propto m_{k \rightarrow l} m_{l \rightarrow k},$$

where $m_{\sim k}$ is the product of messages into $k$. Max-product and hybrid methods can be derived analogously for MAP and marginal MAP problems.

### 2.2 Influence Diagrams

Influence diagrams (IDs) or decision networks are extensions of Bayesian networks to represent structured decision problems under uncertainty. Formally, an influence diagram is defined on a directed acyclic graph $G = (V, E)$, where the nodes $V$ are divided into two subsets, $V = R \cup D$, where $R$ and $D$ represent respectively the set of chance nodes and decision nodes. Each chance node $i \in R$ represents a random variable $x_i$ with a conditional probability table $p_i(x_i | x_{pa(i)}).$ Each decision node $i \in D$ represents a controllable decision variable $x_i$, whose value is determined by a decision maker via a decision rule (or policy) $\delta_i : X_{pa(i)} \rightarrow \mathcal{X}_i$, which determines the values of $x_i$ based on the observation on the values of $x_{pa(i)}$; we call the collection of policies $\delta = \{\delta_i | i \in D\}$ a strategy. Finally, a utility function $u : \mathbb{X} \rightarrow \mathbb{R}^+$ measures the reward given an instantiation of $x = [x_R, x_D]$, which the decision maker wants to maximize. It is reasonable to assume some decomposition structure on the utility $u(x)$, either additive, $u(x) = \sum_{j \in U} u_j(x_{\beta_j})$, or multiplicative, $u(x) = \prod_{j \in U} u_j(x_{\beta_j}).$ A decomposable utility function can be visualized by augmenting the DAG with a set of leaf nodes $U$, called utility nodes, each with parent set $\beta_j$. See Fig. 1 for a simple example.

A decision rule $\delta_i$ is alternatively represented as a deterministic conditional “probability” $p_i^\delta(x_i | x_{pa(i)})$, where $p_i^\delta(x_i | x_{pa(i)}) = 1$ for $x_i = \delta_i(x_{pa(i)})$ and zero otherwise. It is helpful to allow soft decision rules where $p_i^\delta(x_i | x_{pa(i)})$ takes fractional values; these define a randomized strategy in which $x_i$ is determined by randomly drawing from $p_i^\delta(x_i | x_{pa(i)})$. We denote by $\Delta^\circ$ the set of deterministic strategies and $\Delta$ the set of randomized strategies. Note that $\Delta^\circ$ is a discrete set, while $\Delta$ is its convex hull.

Given an influence diagram, the optimal strategy should maximize the expected utility function (MEU):

$$\text{MEU} = \max_{\delta \in \Delta} \text{EU}(\delta) = \max_{\delta \in \Delta} E(u(x) | \delta)$$

$$= \max_{\delta \in \Delta} \sum_{x} u(x) \prod_{i \in C} p_i(x_i | x_{pa(i)}) \prod_{i \in D} p_i^\delta(x_i | x_{pa(i)})$$

$$= \max_{\delta \in \Delta} \sum_{x} \exp(\theta(x)) \prod_{i \in D} p_i^\delta(x_i | x_{pa(i)})$$

where $\theta(x) = \log[u(x) \prod_{i \in C} p_i(x_i | x_{pa(i)})]$, we call the distribution $q(x) \propto \exp(\theta(x))$ the augmented distribution [Biela et al., 1999]. The concept of the augmented distribution is critical since it completely specifies a MEU problem without the semantics of the influence diagram; hence one can specify $q(x)$ arbitrarily, e.g., via an undirected MRF, extending the definition of IDs. We can treat MEA as a special sort of “infrence” on the augmented distribution, which as we will show, generalizes more common inference tasks.

In (4) we maximize the expected utility over $\Delta$; this is equivalent to maximizing over $\Delta^\circ$, since

**Lemma 2.1.** For any ID, $\max_{\delta \in \Delta} \text{EU}(\delta) = \max_{\delta \in \Delta^\circ} \text{EU}(\delta)$. 

---

**Figure 1:** A simple influence diagram for deciding vacation activity [Shachter, 2007].

---

525