structure provides an intuitive language for expressing domain-specific knowledge about variable relationships and facilitates the transfer of modeling advances to new applications.

Several different formalisms have been proposed that use graphs to represent probability distributions.[28, 30, 47, 50] *Directed* graphical models, or *Bayesian networks*, are widely used in artificial intelligence to encode causal, generative processes. Such directed graphs provided the basis for the earliest versions of the BP algorithm.[37] Alternatively, *undirected* graphical models, or *Markov random fields* (MRFs), provide popular models for the symmetric dependencies arising in such areas as signal processing, spatial statistics, bioinformatics, and computer vision.

2.1. Pairwise Markov random fields

An undirected graph $\mathcal{G}$ is defined by a set of nodes $\mathcal{V}$ and a corresponding set of undirected edges $\mathcal{E}$ (see Figure 1). Let $\Gamma(i) \triangleq \{j \mid (i,j) \in \mathcal{E}\}$ denote the *neighborhood* of a node $i \in \mathcal{V}$. MRFs associate each node $i \in \mathcal{V}$ with an unobserved, or hidden, random variable $x_i \in \mathcal{X}_i$. Let $x = \{x_i \mid i \in \mathcal{V}\}$ denote the set of all hidden variables. Given evidence or observations y, a *pairwise MRF* represents the posterior distribution $p(x \mid y)$ in factored form:

$$p(x \mid y) \propto p(x, y) \propto \prod_{(i,j)\in\mathcal{E}} \psi_{ij}(x_i, x_j) \prod_{i\in\mathcal{V}} \psi_i(x_i, y). \tag{1}$$

Here, the proportionality sign indicates that $p(x, y)$ may only be known up to an uncertain normalization constant, chosen so that it integrates to one. The positive *potential functions* $\psi_{ij}(x_i, x_j) > 0$ can often be interpreted as soft, local constraints. Note that the local evidence potential $\psi_i(x_i, y)$ does *not* typically equal the marginal distribution $p(x_i \mid y)$, due to interactions with other potentials.

In this paper, we develop algorithms to approximate the conditional marginal distributions $p(x_i \mid y)$ for all $i \in \mathcal{V}$. These densities lead to estimates of x_i, such as the posterior mean $\mathbb{E}[x_i \mid y]$, as well as corresponding measures of uncertainty. We focus on pairwise MRFs due to their simplicity and popularity in practical applications. However, the nonparametric updates we present may also be directly extended to models with higher-order potentials.

2.2. Belief propagation

For graphs that are acyclic or tree-structured, the desired conditional distributions $p(x_i \mid y)$ can be directly calculated by a local message-passing algorithm known as *belief propagation* (BP).[37, 50] At each iteration of the BP algorithm, nodes $j \in \mathcal{V}$ calculate messages $m_{ji}(x_i)$ to be sent to each neighboring node $i \in \Gamma(j)$:

$$m_{ji}(x_i) \propto \int_{\mathcal{X}_j} \psi_{ij}(x_i, x_j)\, \psi_j(x_j, y) \prod_{k\in\Gamma(j)\setminus i} m_{kj}(x_j)dx_j. \tag{2}$$

The outgoing message is a positive function defined on $\mathcal{X}_i$. Intuitively, it is a (possibly approximate) sufficient statistic of the information that node j has collected regarding x_i. At any iteration, each node can produce an approximation $q_i(x_i)$ to the marginal distribution $p(x_i \mid y)$ by combining incoming messages with the local evidence potential:

$$q_i(x_i) \propto \psi_i(x_i, y) \prod_{j\in\Gamma(i)} m_{ji}(x_i). \tag{3}$$

These updates are graphically summarized in Figure 2. For tree-structured graphs, the approximate marginals, or *beliefs*, $q_i(x_i)$ will converge to the true marginals $p(x_i \mid y)$ once messages from each node have propagated across the graph. With an efficient update schedule, the messages for each distinct edge need only be computed once, and BP can be seen as a distributed variant of dynamic programming.

Because each iteration of the BP algorithm involves only local message updates, it can be applied even to graphs with cycles. For such graphs, the statistical dependencies between BP messages are not accounted for, and the sequence of beliefs $q_i(x_i)$ will *not* converge to the true marginals. In many applications, however, the resulting *loopy BP* algorithm[37] exhibits excellent empirical performance.[8, 14, 15, 49] Recently, several theoretical studies have provided insight into the approximations made by loopy BP, establishing connections to other *variational* inference algorithms[47] and partially justifying its application to graphs with cycles.[20, 23, 34, 50, 51]

The BP algorithm implicitly assumes messages $m_{ji}(x_i)$ have a finite parameterization, which can be tractably updated via the integral of Equation 2. Most implementations

Figure 2. Message-passing recursions underlying the BP algorithm. *Left:* Approximate marginal (belief) estimates combine the local observation potential with messages from neighboring nodes. *Right:* A new outgoing message (red) is computed from all other incoming messages (blue).