function and dual block optimization, and demonstrate the effectiveness of the approach.

2 Notations, Problem Setup and Background

Let \( x = (x_1, \ldots, x_n) \) be the realizations of \( n \) discrete random variables where the range of the \( i^{th} \) random variable is \( \{1, \ldots, n_i\} \), i.e., \( x_i \in \{1, \ldots, n_i\} \). We consider a joint distribution \( p(x_1, \ldots, x_n) \) and assume that it factors into a product of non-negative functions \( \psi_r(x_r) \), known as potentials. Usually, the potentials are defined over a small subset of indexes \( r \subseteq \{1, \ldots, n\} \), called regions:

\[
p(x_1, \ldots, x_n) = \frac{1}{Z} \prod_{r \in \mathcal{R}} \psi_r(x_r)
\]

where for singletons, i.e. \( |r| = 1 \), the functions \( \psi_r(x_r) \) represent "local evidence" or prior data on the states of \( x_i \), and for \( |r| > 1 \) the potential functions describe the interactions of their variables \( x_r \subset \{x_1, \ldots, x_n\} \), and \( Z \) is the normalization constant, also called the partition function. For convenience, we adopt the additive form by setting \( \phi_r(x_r) = \ln \psi_r(x_r) \) thereby having the joint probability take the form of a Gibbs distribution:

\[
p(x_1, \ldots, x_n) = e^{\sum_{r \in \mathcal{R}} \phi_r(x_r)} - \ln Z
\]

For example, \( p(x_1, x_2, x_3) \propto \exp(\phi_2(x_2) + \phi_{123}(x_1, x_2, x_3) + \phi_{23}(x_2, x_3)) \) has three factors with regions \( \mathcal{R} = \{2\}, \{1, 2, 3\}, \{2, 3\} \). The factorization structure above defines a hypergraph whose nodes represent the \( n \) random variables, and the regions \( r \in \mathcal{R} \) correspond to its hyperedges. A convenient way to represent this hypergraph is by a region graph. A region graph is a directed graph whose nodes represent the regions and its direct edges correspond to the inclusion relation, i.e., a directed edge from node \( r \) to \( s \) is possible only if \( s \subset r \). We adopt the terminology where \( P(r) \) and \( C(r) \) stand for all nodes that are parents and children of the node \( r \), respectively. Also, we define \( R_i \) to be the set of regions which contains the variable \( i \).

Inference is closely coupled with the ability to evaluate the logarithm of the partition function \( \ln Z \). From a variational perspective, there is a relationship between the (minus) Gibbs-Helmholtz free-energy and \( \ln Z \):

\[
\ln Z = \max_{p(x) \geq 0, \sum_r p(x) = 1} \left\{ \sum_{r \in \mathcal{R}} \sum_{x_r} p(x_r) \phi_r(x_r) + H(p) \right\}
\]

where \( p(x_r) = \sum_{x \setminus x_r} p(x) \) is the marginal probability and \( H(p) = -\sum_x p(x) \ln p(x) \) is the entropy of the distribution. However, the complexity of the variational representation is unwieldy because both the entropy and the simplex constraint require an evaluation over all possible states of the system \( x = (x_1, \ldots, x_n) \), which is exponential in \( n \). Instead one looks for an approximation or bounds. An upper bound is designed with a tractable approximation of the free-energy by (i) upper bounding the entropy term \( H(p) \) by a combination of local entropies over marginal probabilities \( p(x_r) \), and (ii) by outer bounding the probability simplex constraints by the so called "local consistency" constraints.

An upper bound on the entropy function \( H(p) \) proceeds by replacing the entropy by the fractional covering entropy bounds [6, 23]. These upper bounds are defined as sum of local entropies over the marginal probabilities \( H(p(x_r)) = -\sum_x p(x_r) \ln p(x_r) \):

\[
H(p) \leq \sum_{r \in \mathcal{R}} c_r H(p(x_r))
\]

These upper bounds hold whenever \( c_r \geq 0 \) and for every \( i = 1, \ldots, n \) there holds \( \sum_{r \in \mathcal{R}} c_r = 1 \), where \( R_i \) is the set of all regions that contain \( i \). The second step in obtaining an efficient upper bound is replacing the marginal distributions \( p(x_r) \) by "beliefs" \( b_r(x_r) \). The beliefs form "pseudo distributions" in the sense that the beliefs might not necessarily arise as marginal probabilities of some distribution \( p(x_1, \ldots, x_n) \). To maintain local consistency between beliefs which share the same variables, we define the local consistency polytope \( L(G) \) for the region graph \( G \), as follows:

\[
L(G) = \left\{ \{b_r\}_{r \in \mathcal{R}} : \sum_{x \setminus x_r} b_s(x_s) = b_r(x_r) \quad \forall r, s \in P(r) \right\}
\]

For example, assume \( \mathcal{R} \) consists of the three factors \( \{1\}, \{1, 2\}, \{1, 3\} \) then the consistency constraints on the beliefs \( b_1(x_1), b_{1,2}(x_1, x_2), b_{1,3}(x_1, x_3) \) enforce their distribution constraints and marginalization constraints \( \sum_{x_2} b_{1,3}(x_1, x_3) = b_1(x_1) \) and \( \sum_{x_2} b_{1,2}(x_1, x_2) = b_1(x_1) \).

Taken together, the upper bound on the log-partition is defined as follows:

\[
\ln Z \leq \max_{b_r(x_r) \in L(G)} \left\{ \sum_{r \in \mathcal{R}} b_r(x_r) \phi_r(x_r) + \sum_r c_r H(b_r) \right\}
\]

Thus we introduce a family of upper bounds for the partition function, for the set of fractional covering numbers \( c_r \). The computational complexity of these upper bounds is no longer exponential in \( n \), but linear in the number of the regions and the number of assignments in each region.