STAT 535 Lecture 10

# Max Propagation and Sampling in a Junction Tree

©Marina Meilă
mmp@stat.washington.edu

## 1 The MAP inference problem

The so called **Maximum A-posteriori Probability (MAP)** inference problem is the problem of finding the most probably configuration of the variables in $V$ given evidence $E = e_0$, and it probability.

$$(MAP) \qquad \begin{array}{rcl} p^* & = & \max_{x_V \in \Omega_V} P_V(x_V) \\ x^* & = & \text{argmax}_{x_V \in \Omega_V} P_V(x_V) \end{array}$$

In the above, $x^*$ is called the MAP configuration. We will assume for simplicity that $x^*$ is unique.

This problem can be solved by a modification of the Junction Tree algorithm. We will assume for now that the JT potentials contain a valid, normalized and calibrated representation of probability distribution $P_V$.

## 2 The JT Algorithm with Max Propagation – obtaining $p^*$

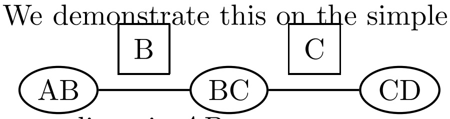
The motivation for Max Propagation is the **distributivity of multiplication w.r.t. the max operation**, for non-negative numbers.

$$\max\{ab_1,\, ab_2\} \;=\; a\max\{b_1, b_2\} \qquad a, b_1, b_2 \geq 0 \tag{1}$$

or, more generally

$$\max_{b \in \Omega_B}\{ab\} \;=\; a \max_{b \in \Omega_B}\{b\} \qquad a, b \geq 0 \tag{2}$$

This property can be applied to a probability distribution represented by a JT. We demonstrate this on the simple example




$P_{ABCD} \;=\; \frac{\phi_{AB}\phi_{BC}\phi_{CD}}{\phi_B \phi_C}$ where we assume the root clique is $AB$.

Finding $p^*$ the maximum value of $P_{ABCD}$ amounts to

$$\begin{aligned}
\max_{abcd} P_V &= \max_{abcd} \frac{\phi_{AB}\phi_{BC}\phi_{CD}}{\phi_B \phi_C} && (3)\\
&= \max_{ab} \frac{\phi_{AB}}{\phi_B} \max_c \left[ \frac{\phi_{BC}}{\phi_C} \underbrace{\max_d \phi_{CD}}_{\phi_C^{new}} \right] && (4)\\
&= \max_{ab} \frac{\phi_{AB}}{\phi_B} \max_c \underbrace{\left[ \phi_{BC} \frac{\phi_C^{new}}{\phi_C} \right]}_{\phi_{BC}^{new}} && (5)\\
&= \max_{ab} \phi_{AB} \frac{\phi_B^{new}}{\phi_B} && (6)\\
&= \max_{ab} \phi_{AB}^{new} && (7)
\end{aligned}$$

The above sequence of algebraic manipulation can be readily seen as a propagation algorithm, where the remotest clique $CD$ passes the"message" $\max_d \phi_{CD}$ to its parent clique through the separator $C$, after which a similar message is passed recursively from $BC$ to its parent $AB$.

The sequence is thus equivalent to a CollectEvidence( $AB$ ) call, where the only modification is in the Absorb function, replaced now with MaxAbsorb.

`MaxAbsorb`$(C \to C')$

1. $\phi_S^{new} \leftarrow \max_{C \setminus S} \phi_C$
2. $\phi_{C'}^{new} \leftarrow \phi_{C'} \frac{\phi_S^{new}}{\phi_S}$
3. $\phi_S \leftarrow \phi_S^{new}$

**Remarks** (The proofs are left as exercise)

1. MaxAbsorb does not change the joint distribution, i.e $P_V = \prod_C \phi_C / \prod_S \phi_S$ is invariant.
2. After CollectEvidence and DistributeEvidence with any root, the JT will be **max-calibrated**, i.e

$$\max_{C \setminus S} \phi_C = \max_{C' \setminus S} \phi_{C'} = \phi_S \qquad (8)$$

for any tree edge $C - S - C'$.

3. After CollectEvidence the root clique $C_0$ contains a potential equal to $\max_{\Omega_{V \setminus C_0}} P_V$, i.e for each configuration $x_{C_0} \in \Omega_{C_0}$, the corresponding $\phi_{C_0}(x_{C_0})$ is the probability of the most likely configuration with the given $x_{C_0}$.

   Hence, the maximum of $\phi_{C_0}$ will be the maximum of $P_V$, $p^*$.
4. After DistributeEvidence, the maximum of $\phi_C$ in any clique $C$ will equal $p^*$. (This is due to the max-calibration property of the JT.)
5. Since $P_V$ was not changed, the JT can be returned to the original calibrated state by performing the standard JT algorithm (without normalization).
6. If evidence $E = e_0$ is entered before the max-propagation steps, then the JT will contain $P_{V,E=e_0}$ and $p^*$ will be the maximum of this new distribution.

Max Propagation is a special case of a more general discrete optimization technique called **Dynamic Programming**. In the special case when the JT represents a Hidden Markov Model, Max Propagation is nothing else than the well known **Viterbi Algorithm**.

## 3 Obtaining the MAP configuration $x^*$

To obtain the (unique) configuration $x^*$ that has probability $p^*$, we need to create a distributed representation for it. Thus for each clique $C$ and separator $S$, we create an additional potential $I_C$, respectively $I_S$ which take values in $\{0, 1\}$. In other words, the $I$ potentials are indicator variables for the maximum configuration in each clique and separator.

If you are familiar with Dynamic Programming, you will recognize in the $I$ variables, the indices for backtracking that a Dynamic Programming uses to recover the optimizing configuration, after it finds the optimal solution.

Max-Propagation with the $I$ potentials proceeds in the following way:

1. At CollectEvidence

   - We set each potential $I_C$ during MaxAbsorb($C \to pa(C)$) as follows: Let $C = D \cup S$, where $S$ is the separator between $C$ and its parent. That is, $D$ contains the variables in $C$ but not in its parent. Now, for $x_S \in \Omega_S$ set $I_C(x_D, x_S) = 1$ if $x_D = \text{argmax}_{\Omega_D} \phi_C(x'_D, x_S)$ and 0 otherwise. The values $I_C(x_D, x_S)$ will be the indicator function of the maximum in $\phi_C(., x_S)$ for each fixed $x_S$.

$$I_C(x_D, x_S) \;=\; \begin{cases} 1 & \text{if} \phi_C(x_D, x_S) = \phi_S^{new}(x_S) \\ 0 & \text{if} \phi_C(x_D, x_S) < \phi_S^{new}(x_S) \end{cases} \tag{9}$$

- Set all $I_S \equiv 1$

2. In stead of normalization set

$$I_{C_0}(x_{C_0}) \;=\; \begin{cases} 1 & \text{if} x_{C_0} = \text{argmax} \phi_{C_0} \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

for the root clique $C_0$. If $x^*$ is unique, then a single value of $I_{C_0}$ will be set to 1.

3. At DistributeEvidence propagate messages for the $I$ potentials by performing a MaxIAbsorb$(C \to C')$ for every MaxAbsorb$(C \to C')$.

`MaxIAbsorb`$(C \to C')$

(a) $I_S^{new} \;\leftarrow\; \max_{C \setminus S} I_C$

(b) $I_{C'}^{new} \;\leftarrow\; I_{C'} \frac{I_S^{new}}{I_S}$

(c) $I_S \;\leftarrow\; I_S^{new}$

One can immediately remark that the divisions by $I_S$ is are superfluous, since these potentials will all be identical to 1. They were included for the sake of unity only. However, the messages $I_S^{new}$ will not be identical to 1. In fact, if $x^*$ is unique, each message $I_S^{new}$ will contain a single 1, indicating the configuration of the parent clique that achieves the maximum. After MaxIAbsorb, the child clique will also contain a unique 1, indicating $x_C^*$, the configuration of its variables that achieves $p^*$. (This fact can be proved easily by induction from the root clique outwards.)

At the end of the Max-Propagation, each $I_C, I_S$ will contain thus a single 1, which will indicate $x_C^*$ the optimal configuration of the variables in $C$. All the $x_C^*$ configurations will be calibrated with the unique $x^*$ – hence we will have obtained a *distributed representation* for $x^*$ by means of the indicator variables $I_C$.

(Note that the algorithm presented here differs slightly from the algorithm described in Cowell (pp. 31). Other variants exist as well.)

If there are more than one most probable configurations, then in each clique we must chose a single one (by setting all other 1's to zeros) before we proceed with DistributeEvidence from that clique. The algorithm will contain at the end *a* most probable configuration only.

## 4 Counting most probable configurations

To find the number of most probable configurations, one can use another JT-like propagation algorithm.

1. Perform a regular Max-Propagation (COLLECT and DISTRIBUTE on the potentials $\phi$ to obtain a max-calibrated JT. (After it, $\max_{\Omega_C} \phi_C = p^*$, $\max_{\Omega_S} \phi_S = p^*$).
2. Create indicator potentials $I_C$ ($I_S$) for all cliques (separators) respectively. Each $I_C$ ($I_S$) contains 1 for the configurations that equal $p^*$ in the respective clique (or separator) and 0 otherwise.
$$I_C(x_C) \;=\; \begin{cases} 1 & \text{if } \phi_C(x_C) = p^* \\ 0 & \text{otherwise} \end{cases} \tag{11}$$
Every $I$ potential will contain at least a one.
3. Perform COLLECTEVIDENCE on the $I$ potentials, with the standard ABSORB function (which sums over $I_{C\setminus S}$ to obtain $I_S^{new}$).
4. "Normalization". Summation over $I_C$ in the root clique will give $Z = |\text{Argmax}_{\Omega_V} P_V|$.
5. **Optional** DISTRIBUTEEVIDENCE will make all the $I_C, I_S$ tables marginally calibrated. Then for any $I_C$ ($I_S$) we will have
$$\sum_{\Omega_C} I_C \;=\; Z \;=\; \#\text{solutions} \tag{12}$$

The $I$ potentials preserve the invariant (Exercise: prove that ABSORB does not change $I_V$).
$$I_V \;=\; \frac{\prod_C I_C}{\prod_S I_S} \tag{13}$$
$I_V(x)$ is an indicator function that is 1 when $x$ is a most probable configuration (i.e $P_V(x) = p^*$) and 0 otherwise.

Note also that the counting algorithm using the $I$ potentials is general, and can be applied to count other types of configurations in the JT.

## 5 Sampling by JT propagation

Taking a sample from a probability table can be thought of as observing evidence. Therefore, sampling by the JT algorithm is a form of DISTRIBUTEEVIDENCE.

Start with a (marginal) calibrated JT (or after ENTEREVIDENCE, COLLECTEVIDENCE if evidence exists).

- DISTRIBUTEEVIDENCE using the following modified Absorb function.

  SAMPLEABSORB($C \to C'$)
  1. ABSORB($C \to C'$)
  2. Sample $x^*_{C'} \sim \phi_{C'}$
  3. Enter evidence $x^*_{C'}$ in $\phi_{C'}$, i.e $\phi_{C'}(x_{C'}) \leftarrow \phi_{C'}(x_{C'})\delta_{x_{C'},x^*_{C'}}$
  4. Normalize $\phi_{C'}$
  5. [Optionally save the normalization constant $Z_{C'}$.]

**Remarks**

- The DISTRIBUTEEVIDENCE part of the algorithm implicitly assumes that the j.t. is rooted at $C$, and sampling is peformed conditionally on the ancestor cliques.
- For the root clique, the absorbtion from the parent is omitted.
- Step 2 is equivalent with sampling $x_{C' \setminus C} \,|\, x_C$, because the variables that are common between $C$ and $C'$ have already been sampled in the parent clique $C$. The ABSORB in step 1 will have set the entries in $\phi_{C'}$ corresponding to $x_C \neq$ sampled value to 0.
- The normalization will produce a $\phi_{C'}$ with a single 1 in location $x^*_{C'}$ and zero elsewhere. Hence, propagating this potential further will zero out all the entries incompatible with $x^*_{C'}$ from the children cliques.
- After ABSORB (i.e before the sampling step) the potential $\phi_{C'}$ will be normalized.
- Multiplying the normalization constants $Z_C$ gives the probability of the sample.

  $$P_V(x^*) = \prod_C Z_C = \frac{\prod_C Z_C}{\prod_S Z_S} \tag{14}$$

  The second equality is true because the $\phi_S$ potentials will contain a single 1, so their normalization constants $Z_S$ will be all 1.