MAP Estimation Via Agreement on Trees: Message-Passing and Linear Programming

M.J. Wainwright, T.S. Jaakkola, A.S. Willsky

Outline

1. Introduction
   - Maximum a Posteriori Probability and Graphical Model
   - Definition and Symbols
2. Motivation
   - Convexity of Phi
   - Upper Bounds
3. Tree-Reweighted Message-Passing Algorithms
   - Max-Marginals
   - Algorithm
4. Conclusion

What Is This Paper About

- MAP derived from a graph is upper bounded by the linear combination of the sub-trees of the graph.
- Although the number of sub-trees may be untraceable, the problem turns out to be solvable using local marginal information that is only related to the nodes and edges in the dual space because of the convexity of the upper bound.
- There is a constraint imposed on the upper bound: the count of edges of the graph and the count of edges in the sub-trees need to be consistent.
- The constraint can be met with a special message passing construction: tree-reweighted belief propagation.

Markov Random Fields

- Advantage of MRF
  - Well structured isotropic behavior
  - Local dependencies
- Disadvantages of MRF
  - Difficult to compute probability
  - Parameter estimation is hard
- Applications: too many to list
  - Machine learning, Imaging, Computer Vision, Economics etc.

Random Fields

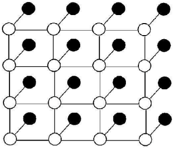
- Random variables Y can be considered a Markov Random Field (MRF) on S if:
  - $P(Y) > 0$
  - $P(y_i | y_{S-\{i\}}) = P(y_i | y_{N_i})$

We can also formulate this based on graphical model. Let $G = (V, E)$ be a graph with vertices V and edges E. Vertices $V = X \cup Y$ with X as the observation and Y (label) as random variables. The factorization is defined by: **Hammersley-Clifford's Theorem.** Assume that $p(y_1, \dots, y_n) > 0$ Then,

$$p(y) = \frac{1}{Z} exp \left(- \sum_{C} \phi_C(x_C) \right)$$

where C is a clique, subset of nodes that are fully connected in the graph.

Graphical Representation

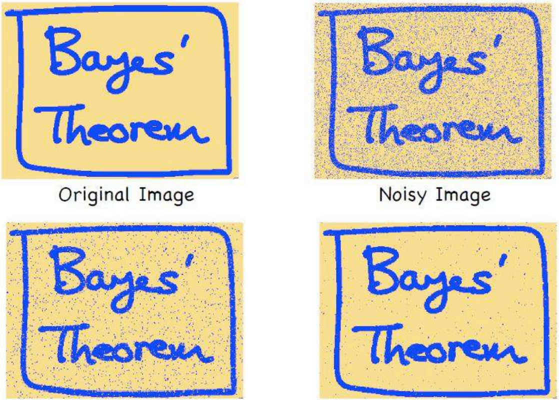


Let's denote black nodes as observed nodes y_i and white nodes as hidden nodes x_i. The overall joint probability of $p(x, y)$ is:

$$p(x,y) = \frac{1}{Z} \prod_{ij} \psi_{ij}(x_i, x_j) \prod_{i} \phi_i(x_i, y_i)$$

where the $\psi_{ij}(x_i, x_j)$ and $\phi_i(x_i, y_i)$ are joint compatibility functions. Then the Maximum A Posteriori (MAP) is given by: $\text{argmax}_x\, p(x|y)$.

A Image Denoise/Segmentation Example



Original Image — Noisy Image

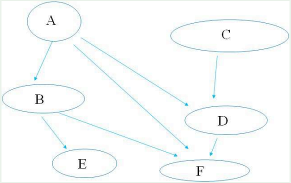
Restored Image (ICM) — Restored Image (Graph cuts)

Pattern Recognition and Machine Learning
(C.M. Bishop)

MAP and Graphical Model

- Message passing (MP) is popular in solving MAP problems (integer programming) in acyclic graph.
- MP exploits the conditional independent properties which is the key to factorize the graph.

Example



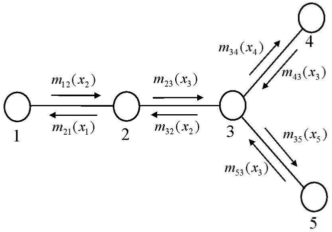
$$P(A,B,C,D,E,F) = P(A)P(C)P(B|A)P(D|C,A)P(E|B)P(F|A,B,D)$$

For the undirected graph, the overall joint probability of p(x,y) is:

$$p(x,y) = \frac{1}{Z}\prod_i \phi_i(x_i)\prod_{i,j}\psi_{ij}(x_i,x_j)$$

MP in Undirected Graph

A typical message passing route is shown as:



The belief at a node i is proportional to the product of the local evidence at that node ($\phi_i(x_i)$), and all messages coming into node i:

$$
\begin{aligned}
b_i(x_i) &= k\phi_i(x_i) \prod_{j \in \mathcal{N}(i)} m_{ji}(x_i) \\
m_{ij}(x_j) &= \sum_{x_i} \phi_i(x_i)\psi_{ij}(x_i, x_j) \prod_{k \in \mathcal{N} \backslash j} m_{ki}(x_i)
\end{aligned}
$$

Preliminaries

An undirected graph is defined as $G(V, E)$. For each $s \in V$, let X_s be a random variable taking values x_s in sample space $\mathcal{X}_s$ and $\mathcal{X}_s := \{0, \ldots, m_s - 1\}$. For $n = |V|$ elements, X takes values x in the Cartesian product space $\mathcal{X}^n := \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_n$. A full collection of potential functions associated with a given clique C is mapping $\phi : \mathcal{X}^n \to \mathbb{R}^d$ with $\{\phi_\alpha | \alpha \in \mathcal{I}\}$ and $d = |\mathcal{I}|$. $\theta = \{\theta_\alpha | \alpha \in \mathcal{I}\}$ is the vector of parameter. Then, strictly positive MRF can be represented as:

$$p(x; \theta) \propto \exp\{\langle \theta, \phi(x) \rangle\} \equiv \exp \sum_{\alpha \in \mathcal{I}} \theta_\alpha \phi_\alpha(x)$$

For easy annotation, we define the following indicator functions:

$$\{\delta_j(x_s) | j \in \mathcal{X}_s\}, \text{ for } s \in V$$
$$\{\delta_j(x_s)\delta_k(x_t) | (j, k) \in (X)_s \times \mathcal{X}_t\}, \text{ for } (s, t) \in E$$

Marginal distribution can be represented by:

$$\mu_{s;j} := \mathbb{E}_p[\delta_j(x_s)] = \sum_{x \in \mathcal{X}^n} p(x)\delta_j(x_s)$$
$$\mu_{st;jk} := \mathbb{E}_p[\delta_j(x_s)\delta_k(x_t)] = \sum_{x \in \mathcal{X}} p(x)[\delta_j(x_s)\delta_k(x_t)]$$

Linear Constraints and MAP Estimation

Linear Constraints

$$\sum_{j\in\mathcal{X}_s} \mu_{s;j} = 1,\ \forall s \in V$$

$$\sum_{(j,k)\in\mathcal{X}_s\times\mathcal{X}_l} \mu_{st;jk} = 1,\ \forall (s,t) \in E, j \in \mathcal{X}_s$$

$$\sum_{k\in\mathcal{X}_l} \mu_{st;jk} = \mu_{s;j}, \forall (s,t) \in E, j \in \mathcal{X}_s$$

MAP Estimation

Let $\overline{\theta}$ be a given vector of parameter of $\mathbb{R}^d$. Let $\overline{\theta}_s(x_s) := \sum_{j\in\mathcal{X}_s} \overline{\theta}_{s;j}\delta_j(x_s)$. Let $\overline{\theta}_{st}(x_s;x_t) := \sum_{(j,k)\in\mathcal{X}_s\times\mathcal{X}_t} \overline{\theta}_{st;jk}\delta_j(x_s)\delta_k(x_t)$. MAP is to maximize:

$$\langle\overline{\theta},\phi(x)\rangle := \sum_{s\in V} \overline{\theta}_s(x_s) + \sum_{(s,t)\in E} \overline{\theta}_{st}(x_s,x_t)$$

For the convenience, we define the MAP as follows:

$$\Phi_\infty(\overline{\theta}) := \max_{x\in(X)^n} \langle\overline{\theta},\phi(x)\rangle \tag{1}$$

Convexity

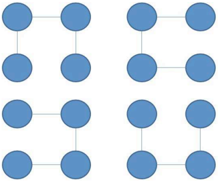
Claim: Φ_∞ is convex in terms of $\overline{\theta}$

Proof

Let's start with a more general log-partition function: $\Phi(\theta) = \log \sum_x \exp\{\theta^T \phi(x)\}$.

$$
\begin{aligned}
\frac{\partial \Phi}{\partial \theta_k} &= \frac{\sum_x \exp\{\theta^T \phi(x)\}\phi_k(x)}{\sum_x \exp\{\theta^T \phi(x)\}} \\
&= \frac{\sum_x \exp\{\theta^T \phi(x)\}\phi_k(x)}{\exp \Phi(\theta)} \\
&= \sum_x \exp\{\theta^T \phi(x) - \Phi(\theta)\}\phi_k(x) \\
&= \sum_x p(x;\theta)\phi_k(x) \\
&= \mathbb{E}\{\phi_k(X)\} \\
\frac{\partial^2 \Phi}{\partial \theta_k \partial \theta_l} &= \frac{\partial}{\partial \theta_l} \sum_x \exp\{\theta^T \phi(x) - \Phi(\theta)\}\phi_k x \\
&= \sum_x \exp\{\theta^T \phi(x) - \Phi(\theta)\}[\phi_l(x) - \frac{\partial \Phi(\theta)}{\partial \theta_l}]\phi_k(x) \\
&= \mathbb{E}\{\phi_k(X)\phi_l(X)\} - \{\phi_k(X)\}\mathbb{E}\{\phi_l(X)\} \\
&= Cov_p\{\phi_k(x), \phi_l(x)\}.
\end{aligned}
$$

Illustration of Edge Appearance in Probabilities



For $p(x, \overline{\theta}) \propto \exp(x_1 x_2 + x_2 x_3 + x_3 x_4 + x_4 x_1)$, $\overline{\theta} = [0\ 0\ 0\ 0\ 1\ 1\ 1\ 1]$. Then $\rho(T_i) = 1/4$ and $\rho_e = 3/4$ for each $e \in E$. The can construct a member θ as follows:

$$\begin{aligned}
\theta(T_1) &= (4/3)[0\ 0\ 0\ 0\ 1\ 1\ 1\ 0] \\
\theta(T_1) &= (4/3)[0\ 0\ 0\ 0\ 1\ 1\ 0\ 1] \\
\theta(T_1) &= (4/3)[0\ 0\ 0\ 0\ 1\ 0\ 1\ 1] \\
\theta(T_1) &= (4/3)[0\ 0\ 0\ 0\ 0\ 1\ 1\ 1]
\end{aligned}$$

Upper Bounds via Convex Combinations

Let ρ^i be a finite collection of nonnegative weights that sum to one, s.t., $\sum_i \rho^i \theta^i = \bar{\theta}$. Applying Jensen's inequality yields the upper bound: $\Phi_\infty \leq \sum_i \rho^i \Phi_\infty(\theta^i)$. Each $\Phi_\infty(\theta^i)$ represents an acyclic subgraph, for which exact computation are tractable. The index i corresponds to a spanning tree of the graph and all parameters are required to respect the structure of the tree. For a convex combination of tree parameters, $\mathcal{E}_\rho[\theta(T)] := \sum_T \rho(T)\theta(T)$:

Tightness of Upper Bounds

$$\begin{aligned} 0 &\leq \left[\sum_T \rho(T)\Phi_\infty(\theta(T))\right] - \Phi(\bar{\theta}) \\ &= \left[\sum_T \rho(T)\Phi_\infty(\theta(T))\right] - \langle \bar{\theta}, \phi(x^*)\rangle \\ &= \sum_T \rho(T)\left[\Phi_\infty(\theta(T)) - \langle \theta(T), \phi(x^*)\rangle\right] \end{aligned}$$

The bound is tight if and only if $x^* \in \cap_T OPT(\theta(T))$ for some $x^* \in OPT(\bar{\theta})$.

Objective

$$\begin{cases} \min_\theta & \sum_T \rho(T)\Phi_\infty(\theta(T)) \\ \text{s.t.} & \sum_T \rho(T)\theta(T) = \overline{\theta} \end{cases}$$

Theorem

The optimal value of the above problem can be obtained by:

$$\max_\mu \sum_{s\in V} \sum_j \mu_{s;j}\overline{\theta}_{s;j} + \sum_{(s,t)\in E} \sum_{j,k} \mu_{st;jk}\overline{\theta}_{st;jk}$$

Proof I

Claim: $\Phi_\infty(\overline{\theta}) = \max \sum_{s \in V} \sum_j u_{s;j} \overline{\theta}_{s;j} + \sum_{(s,j) \in E} \sum_{j,k} \mu_{st;jk} \overline{\theta}_{st;jk}$

Proof

By definition of Φ_∞, $\max_{x \in \mathcal{X}^n} \langle \overline{\theta}, \phi(x) \rangle = \max_{p \in \mathcal{P}} \sum_{x \in \mathcal{X}^n} p(x) \langle \overline{\theta}, \phi(x) \rangle$.

$$\begin{aligned} \sum_{x \in \mathcal{X}^n} p(x) \langle \overline{\theta}, \phi(x) \rangle &= \sum_{x \in \mathcal{X}^n} p(x) \left\{ \sum_{s \in V} \overline{\theta}_s(x_s) + \sum_{(s,t) \in E} \overline{\theta}_{st}(x_s, x_t) \right\} \\ &= \sum_{s \in V} \sum_j \mu_{s;j} \overline{\theta}_{s;j} + \sum_{(s,t) \in E} \sum_{j,k} \mu_{st;jk} \overline{\theta}_{st;jk} \end{aligned}$$

where $\mu_{s;j} := \sum_{x \in \mathcal{X}^n} p(x) \delta_j(x_s)$ and $\mu_{st;jk} := \sum_{x \in \mathcal{X}^n} p(x) \delta_{jk}(x_s, x_t)$

Lagrange Dual of the Objective Function

$$\begin{aligned} \mathcal{L} &= \sum_T \rho(T)\Phi_\infty(\theta(T)) + \tau(\sum_T \rho(T)\theta(T) - \overline{\theta}) \\ &= \sum_T \rho(T)[\Phi_\infty(\theta(T)) - \langle \theta(T), \tau \rangle] + \langle \tau, \overline{\theta} \rangle \end{aligned}$$

The Lagrange dual is then:

$$\sum_T \rho(T) \inf_{\theta(T)} [\Phi_\infty(\theta(T)) - \langle \theta(T), \tau \rangle] + \langle \tau, \overline{\theta} \rangle$$

So, $\frac{\partial \Phi_\infty(\theta(T)) - \langle \theta(T), \tau \rangle}{\partial \theta} = 0$ and $\tau = E[\phi(x)]$.

Max-Marginal Factorization

Any tree-structured distribution can be factorized in terms of its max-marginals as follows:

$$p(x;\theta(T)) \propto \prod_{s\in V} \mu_s(x_s) \prod_{(s,t)\in E(T)} \frac{\mu_{st}(x_s,x_t)}{\mu_s(x_s)\mu_t(t)}$$

The tree-structured parameter $\theta(T)$ is defined in terms of logarithms of μ:

$$\begin{aligned}
\theta_s^n(T)(x_s) &= \log \mu_s(x_s)\ \forall s \in V \\
\theta_{st}^n(x_s,x_t) &= \log \frac{\mu_{st}(x_s,x_t)}{\mu_s(x_s)\mu_t(x_t)}\ (s,t) \in E(T)
\end{aligned}$$

Edge Based Reparameterization Updates

For iterations $n = 0, 1, 2, \ldots$, update the max-marginals as follows:

$$\mu_s^{n+1}(x_s) = k\mu_s^n(x_s) \prod_{t \in \Gamma(s)} \left[\frac{\max_{x_t'} \mu_{st}^n(x_s, x_t')}{\mu_s^n(x_s)} \right]^{\rho_{st}}$$

$$\mu_{st}^{n+1}(x_s, x_t) = k \frac{\mu_{st}^n(x_s, x_t)}{\max_{x_t'} \mu_{st}^n(x_s, x_t') \max_{x_s'} \mu_{st}^n(x_s', x_t)} \mu_s^{n+1}(x_s)\mu_t^{n+1}(x_t)$$

In terms of messages, max-marginals are as follows:

$$\mu_s(x_s) \propto \phi_s(x_s) \prod_{v \in \Gamma(s)} [M_{vs}(x_s)]^{\rho_{vs}}$$

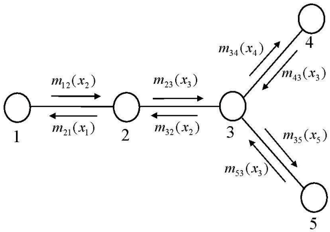
$$\mu_{st}(x_s, x_t) \propto \phi_{st}(x_s, x_t) \frac{\prod_{v \in \Gamma(s) \setminus t} [M_{vs}(x_s)]^{\rho_{vs}}}{[M_{ts}(x_s)]^{(1-\rho_{ts})}} \times \frac{\prod_{v \in \Gamma(t) \setminus s} [M_{vt}(x_t)]^{\rho_{vt}}}{[M_{st}(x_t)]^{(1-\rho_{st})}}$$

Edge Based Reparameterization Updates

The above construction establishes that for all $x \in \mathcal{X}^n$, we have:

$$\sum_{T} \rho(T)\theta(T)(x) = \sum_{s \in V} \overline{\theta}_s(x_s) + \sum_{(s,t) \in E} \overline{\theta}_{st}(x_s, x_t)$$

Parallel Tree-Reweighted Message Passing Algorithm



- Initialize the message $m^0 = m^0_{ij}$ with arbitrary positive numbers.
- for each iteration, update the message as follows:

$$m^{n+1}_{ts}(x_s) = k \sum_{x'_t} \exp\left(\frac{1}{\rho_{ts}}\mu_{st}(x_s, x'_t) + \mu_t(x'_t)\right) \times \frac{\prod_{v \in \Gamma(t)\backslash s}[m^n_{vt}(x'_t)]^{\rho_{vt}}}{[m^n_{st}(x'_t)]^{(1-\rho_{st})}}$$

Local Beliefs

Once the process has converged, the local belief can be calculated as:

$$b_s(x_s) = k \exp(-\mu_s(x_s)) \prod_{t \in \Gamma(s)} [m_{ts}(x_s)]^{\rho_{ts}}$$

Testing Example

Testing Example

Conclusion

- The paper provides an upper bound for the optimal (MAP) configuration.
- The new tree-reweighted free energy is convex with respect to the max-marginals vector
- No sufficient conditions to guarantee convergence on graphs with cycles.
- The demo code can be downloaded from http://code.google.com/p/random-field-blief-propagation/