Advanced Machine Learning
Summer 2019

Part 10 – Graphical Models IV
09.05.2019

Prof. Dr. Bastian Leibe

RWTH Aachen University, Computer Vision Group
http://www.vision.rwth-aachen.de
Course Outline

• Regression Techniques
  – Linear Regression
  – Regularization (Ridge, Lasso)
  – Kernels (Kernel Ridge Regression)

• Deep Reinforcement Learning

• Probabilistic Graphical Models
  – Bayesian Networks
  – Markov Random Fields
  – Inference (exact & approximate)

• Deep Generative Models
  – Generative Adversarial Networks
  – Variational Autoencoders
Topics of This Lecture

- **Recap: Exact inference**
  - Sum-Product algorithm
  - Max-Sum algorithm
  - Junction Tree algorithm

- **Applications of Markov Random Fields**
  - Application examples from computer vision
  - Interpretation of clique potentials
  - Unary potentials
  - Pairwise potentials

- **Solving MRFs with Graph Cuts**
  - Graph cuts for image segmentation
  - s-t mincut algorithm
  - Extension to non-binary case
  - Applications
Recap: Factor Graphs

• Joint probability
  – Can be expressed as product of factors: \( p(x) = \frac{1}{Z} \prod_s f_s(x_s) \)
  – Factor graphs make this explicit through separate factor nodes.

• Converting a directed polytree
  – Conversion to undirected tree creates loops due to moralization!
  – Conversion to a factor graph again results in a tree!
Recap: Sum-Product Algorithm

• Objectives
  – Efficient, exact inference algorithm for finding marginals.

• Procedure:
  – Pick an arbitrary node as root.
  – Compute and propagate messages from the leaf nodes to the root, storing received messages at every node.
  – Compute and propagate messages from the root to the leaf nodes, storing received messages at every node.
  – Compute the product of received messages at each node for which the marginal is required, and normalize if necessary.

\[ p(x) \propto \prod_{s \in \text{ne}(x)} \mu_{f_s \rightarrow x}(x) \]

• Computational effort
  – Total number of messages = 2 \cdot \text{number of graph edges}. 

Slide adapted from Chris Bishop
Recap: Sum-Product Algorithm

- Two kinds of messages
  - Message from factor node to variable nodes:
    - **Sum** of factor contributions
      \[
      \mu_{f_s \rightarrow x}(x) \equiv \sum_{X_s} F_s(x, X_s)
      \]
      \[
      = \sum_{X_s} f_s(x_s) \prod_{m \in \text{ne}(f_s) \setminus x} \mu_{x_m \rightarrow f_s}(x_m)
      \]
  - Message from variable node to factor node:
    - **Product** of incoming messages
      \[
      \mu_{x_m \rightarrow f_s}(x_m) \equiv \prod_{l \in \text{ne}(x_m) \setminus f_s} \mu_{f_l \rightarrow x_m}(x_m)
      \]

⇒ Simple propagation scheme.
Recap: Sum-Product from Leaves to Root

Message definitions:

\[
\mu_{f_s \rightarrow x}(x) \equiv \sum_{X_s} f_s(x_s) \prod_{m \in \text{ne}(f_s) \setminus x} \mu_{x_m \rightarrow f_s}(x_m)
\]

\[
\mu_{x_m \rightarrow f_s}(x_m) \equiv \prod_{l \in \text{ne}(x_m) \setminus f_s} \mu_{f_l \rightarrow x_m}(x_m)
\]

\[
\mu_x \rightarrow f(x) = 1 \quad \mu_f \rightarrow x(x) = f(x)
\]
Recap: Sum-Product from Root to Leaves

Message definitions:

\[ \mu_{f_s \rightarrow x}(x) \equiv \sum_{X_s} f_s(x_s) \prod_{m \in \text{ne}(f_s) \setminus x} \mu_{x_m \rightarrow f_s}(x_m) \]

\[ \mu_{x_m \rightarrow f_s}(x_m) \equiv \prod_{l \in \text{ne}(x_m) \setminus f_s} \mu_{f_l \rightarrow x_m}(x_m) \]

\[ \mu_{x \rightarrow f}(x) = 1 \]

\[ \mu_{f \rightarrow x}(x) = f(x) \]
Max-Sum Algorithm

• Objective: an efficient algorithm for finding
  – Value $x^{\text{max}}$ that maximises $p(x)$;
  – Value of $p(x^{\text{max}})$.
\[ \Rightarrow \] Application of dynamic programming in graphical models.

• In general, maximum marginals $\neq$ joint maximum.
  – Example:

<table>
<thead>
<tr>
<th></th>
<th>$x = 0$</th>
<th>$x = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y = 0$</td>
<td>0.3</td>
<td>0.4</td>
</tr>
<tr>
<td>$y = 1$</td>
<td>0.3</td>
<td>0.0</td>
</tr>
</tbody>
</table>

\[
\arg \max_x p(x, y) = 1 \quad \arg \max_x p(x) = 0
\]
Max-Sum Algorithm – Key Ideas

• Key idea 1: Distributive Law (again)

  \[
  \max(ab, ac) = a \max(b, c) \\
  \max(a + b, a + c) = a + \max(b, c)
  \]

  \(\Rightarrow\) Exchange products/summations and max operations exploiting the tree structure of the factor graph.

• Key idea 2: Max-Product \(\rightarrow\) Max-Sum

  – We are interested in the maximum value of the joint distribution

  \[
  p(x^{\text{max}}) = \max_x p(x)
  \]

  \(\Rightarrow\) Maximize the product \(p(x)\).

  – For numerical reasons, use the logarithm.

  \[
  \ln\left(\max_x p(x)\right) = \max_x \ln p(x).
  \]

  \(\Rightarrow\) Maximize the sum (of log-probabilities).
Max-Sum Algorithm

- Maximizing over a chain (max-product)

\[ p(x^{\text{max}}) = \max_x p(x) = \max_{x_1} \ldots \max_{x_M} p(x) \]

\[ = \frac{1}{Z} \max_{x_1} \ldots \max_{x_N} [\psi_{1,2}(x_1, x_2) \cdot \ldots \cdot \psi_{N-1,N}(x_{N-1}, x_N)] \]

\[ = \frac{1}{Z} \max_{x_1} \left[ \max_{x_2} \left[ \psi_{1,2}(x_1, x_2) \left[ \ldots \max_{x_N} \psi_{N-1,N}(x_{N-1}, x_N) \right] \ldots \right] \right] \]

- Exchange max and product operators

- Generalizes to tree-structured factor graph

\[ \max_x p(x) = \max_{x_n} \prod_{f_s \in \text{ne}(x_n)} \max_{X_s} f_s(x_n, X_s) \]
Max-Sum Algorithm

• Initialization (leaf nodes)

\[ \mu_{x \rightarrow f}(x) = 0 \quad \mu_{f \rightarrow x}(x) = \ln f(x) \]

• Recursion
  – Messages

\[ \mu_{f \rightarrow x}(x) = \max_{x_1, \ldots, x_M} \left[ \ln f(x, x_1, \ldots, x_M) + \sum_{m \in \text{ne}(f_s) \setminus x} \mu_{x_m \rightarrow f}(x_m) \right] \]

\[ \mu_{x \rightarrow f}(x) = \sum_{l \in \text{ne}(x) \setminus f} \mu_{f_l \rightarrow x}(x) \]

  – For each node, keep a record of which values of the variables gave rise to the maximum state:

\[ \phi(x) = \arg \max_{x_1, \ldots, x_M} \left[ \ln f(x, x_1, \ldots, x_M) + \sum_{m \in \text{ne}(f_s) \setminus x} \mu_{x_m \rightarrow f}(x_m) \right] \]

Slide adapted from Chris Bishop
Max-Sum Algorithm

• Termination (root node)
  – Score of maximal configuration
  \[ p_{\text{max}} = \max_x \left[ \sum_{s \in \text{ne}(x)} \mu_{f_s \rightarrow x}(x) \right] \]
  – Value of root node variable giving rise to that maximum
  \[ x_{\text{max}} = \arg \max_x \left[ \sum_{s \in \text{ne}(x)} \mu_{f_s \rightarrow x}(x) \right] \]
  – Back-track to get the remaining variable values
  \[ x_{n-1}^{\text{max}} = \phi(x_n^{\text{max}}) \]
Visualization of the Back-Tracking Procedure

- Example: Markov chain

\[ k = 1 \]
\[ k = 2 \]
\[ k = 3 \]

\[ n - 2 \quad n - 1 \quad n \quad n + 1 \]

⇒ Same idea as in Viterbi algorithm for HMMs…

Image source: C. Bishop, 2006
Topics of This Lecture

• Recap: Exact inference
  – Sum-Product algorithm
  – Max-Sum algorithm
  – Junction Tree algorithm

• Applications of Markov Random Fields
  – Motivation
  – Unary potentials
  – Pairwise potentials

• Solving MRFs with Graph Cuts
  – Graph cuts for image segmentation
  – s-t mincut algorithm
  – Extension to non-binary case
  – Applications
Junction Tree Algorithm

- **Motivation**
  - **Exact** inference on general graphs.
  - Works by turning the initial graph into a *junction tree* with one node per clique and then running a sum-product-like algorithm.
  - **Intractable** on graphs with large cliques.

![Diagram](image-url)
Junction Tree Algorithm

• Motivation
  – **Exact** inference on general graphs.
  – Works by turning the initial graph into a *junction tree* and then running a sum-product-like algorithm.
  – **Intractable** on graphs with large cliques.

• Main steps
  1. If starting from directed graph, first convert it to an undirected graph by **moralization**.
  2. Introduce additional links by **triangulation** in order to reduce the size of cycles.
  3. **Find cliques** of the moralized, triangulated graph.
  4. Construct a new graph from the **maximal cliques**.
  5. Remove minimal links to **break cycles** and get a Junction Tree.

⇒ Apply regular **message passing** to perform inference.
Junction Tree Algorithm

• Starting from an undirected graph…
1. Convert to an undirected graph through **moralization**.
   - Marry the parents of each node.
   - Remove edge directions.
2. **Triangulate**
   - Such that there is no loop of length > 3 without a chord.
   - This is necessary so that the final junction tree satisfies the “running intersection” property (explained later).
3. **Find cliques** of the moralized, triangulated graph.
4. Construct a new junction graph from maximal cliques.
   – Create a node from each clique.
   – Each link carries a list of all variables in the intersection.
     ▪ Drawn in a “separator” box.
5. **Remove links to break cycles** ⇒ junction tree.
   - For each cycle, remove the link(s) with the minimal number of shared nodes until all cycles are broken.
   - Result is a maximal spanning tree, the junction tree.
**Junction Tree – Properties**

- **Running intersection property**
  - “If a variable appears in more than one clique, it also appears in all intermediate cliques in the tree”.
  - This ensures that neighboring cliques have consistent probability distributions.
  - Local consistency $\rightarrow$ global consistency
Interpretation of the Junction Tree

• Undirected graphical model

- Junction tree

\[ P(U) = \prod P(\text{Clique}) / \prod P(\text{Separator}) \]

\[ P(A,B,C) = P(A,B) P(B,C) / P(B) \]

Slide adapted from Pawan Kumar
Junction Tree: Example 1

- Algorithm
  1. Moralization
  2. Triangulation (not necessary here)
Junction Tree: Example 1

(b) Moral graph

(c) Junction graph

• Algorithm
  1. Moralization
  2. Triangulation (not necessary here)
  3. Find cliques
  4. Construct junction graph

Image source: J. Pearl, 1988
Junction Tree: Example 1

Algorithm
1. Moralization
2. Triangulation (not necessary here)
3. Find cliques
4. Construct junction graph
5. Break links to get junction tree

Image source: J. Pearl, 1988
Junction Tree: Example 2

- Without triangulation step
  - The final graph will contain cycles that we cannot break without losing the running intersection property!
• When applying the triangulation
  – Only small cycles remain that are easy to break.
  – Running intersection property is maintained.
Junction Tree Algorithm

• Good news
  – The junction tree algorithm is efficient in the sense that for a given graph there does not exist a computationally cheaper approach.

• Bad news
  – This may still be too costly.
  – Effort determined by number of variables in the largest clique.
  – Grows exponentially with this number (for discrete variables).
  ⇒ Algorithm becomes impractical if the graph contains large cliques!
Loopy Belief Propagation

• Alternative algorithm for loopy graphs
  – Sum-Product on general graphs.
  – Strategy: simply ignore the problem.
  – Initial unit messages passed across all links, after which messages are passed around until convergence
    ▪ Convergence is not guaranteed!
    ▪ Typically break off after fixed number of iterations.
  – Approximate but tractable for large graphs.
  – Sometime works well, sometimes not at all.
Topics of This Lecture

• Recap: Exact inference
  – Sum-Product algorithm
  – Max-Sum algorithm
  – Junction Tree algorithm

• Applications of Markov Random Fields
  – Motivation
  – Unary potentials
  – Pairwise potentials

• Solving MRFs with Graph Cuts
  – Graph cuts for image segmentation
  – s-t mincut algorithm
  – Extension to non-binary case
  – Applications
Markov Random Fields (MRFs)

• What we’ve learned so far…
  – We know they are undirected graphical models.
  – Their joint probability factorizes into clique potentials,
    \[ p(x) = \frac{1}{Z} \prod_C \psi_C(x_C) \]
    which are conveniently expressed as
    \[ \psi_C(x_C) = \exp\{-E(x_C)\} \]
  – We know how to perform inference for them.
    ▪ Sum/Max-Product BP for exact inference in tree-shaped MRFs.
    ▪ Loopy BP for approximate inference in arbitrary MRFs.
    ▪ Junction Tree algorithm for converting arbitrary MRFs into trees.

• But what are they actually good for?
  – And how do we apply them in practice?
Markov Random Fields

• Allow rich probabilistic models.
  – But built in a local, modular way.
  – Learn local effects, get global effects out.

• Very powerful when applied to regular structures.
  – Such as images…

Observed evidence

Hidden “true states”

Neighborhood relations

Slide adapted from William Freeman
Applications of MRFs

• Movie “No Way Out” (1987)
Applications of MRFs

- Many applications for low-level vision tasks
  - Image denoising

Results by [Roth & Black, CVPR'05]
Applications of MRFs

• Many applications for low-level vision tasks
  – Image denoising

Results by [Roth & Black, CVPR’05]
Applications of MRFs

• Many applications for low-level vision tasks
  – Image denoising
  – Inpainting

Since 1699, when French explorers landed at the great bend of the Mississippi River and celebrated the first Mardi Gras in North America, New Orleans has brewed a fascinating melange of cultures. It was French, then Spanish, then French again, then sold to the United States. Through all these years, and even into the 1900s, others arrived from everywhere: Acadians (Cajuns), Africans, indige-
Applications of MRFs

- Many applications for low-level vision tasks
  - Image denoising
  - Inpainting
  - Image restoration

Results by [Roth & Black, CVPR’05]
Applications of MRFs

• Many applications for low-level vision tasks
  – Image denoising
  – Inpainting
  – Image restoration
  – Image segmentation

Image source: Pawan M. Kumar
Applications of MRFs

- Many applications for low-level vision tasks
  - Image denoising
  - Inpainting
  - Image restoration
  - Image segmentation
  - Super-resolution

Convert a low-res image into a high-res image!

Image source: [Freeman et al., CG&A'03]
Applications of MRFs

- Many applications for low-level vision tasks
  - Image denoising
  - Inpainting
  - Image restoration
  - Image segmentation
  - Super-resolution
Applications of MRFs

• Many applications for low-level vision tasks
  – Image denoising
  – Inpainting
  – Image restoration
  – Image segmentation
  – Super-resolution
  – Optical flow

Image patches

Scene patches

Image pair

Scene

Image source: William Freeman
Applications of MRFs

• Many applications for low-level vision tasks
  – Image denoising
  – Inpainting
  – Image restoration
  – Image segmentation
  – Super-resolution
  – Optical flow
  – Stereo depth estimation
Applications of MRFs

• Many applications for low-level vision tasks
  – Image denoising
  – Inpainting
  – Image restoration
  – Image segmentation
  – Super-resolution
  – Optical flow
  – Stereo depth estimation

• MRFs have become a standard tool for such tasks.
  – Let’s look at how they are applied in detail…
MRF Structure for Images

• Basic structure

  ![Graphical model diagram]

Noisy observations

“True” image content

• Two components
  – Observation model
    ▪ How likely is it that node $x_i$ has label $L_i$ given observation $y_i$?
    ▪ This relationship is usually learned from training data.
  
  – Neighborhood relations
    ▪ Simplest case: 4-neighborhood
    ▪ Serve as smoothing terms.
    ⇒ Discourage neighboring pixels to have different labels.
    ▪ This can either be learned or be set to fixed “penalties”.
MRF Nodes as Pixels

These neighborhood statistics can be learned from training data!

Slide adapted from William Freeman
MRF Nodes as Patches

More general relationships expressed by potential functions $\Phi$ and $\Psi$. 

Slide credit: William Freeman
Network Joint Probability

- Interpretation of the factorized joint probability

\[ P(x, y) = \prod_i \Phi(x_i, y_i) \prod_{i,j} \Psi(x_i, x_j) \]
### Energy Formulation

- **Energy function**

  \[ E(x, y) = \sum_i \phi(x_i, y_i) + \sum_{i,j} \psi(x_i, x_j) \]

  - **Single-node potentials** \( \phi \)
    - Encode local information about the given pixel/patch.
    - How likely is a pixel/patch to belong to a certain class (e.g. foreground/background)?

  - **Pairwise potentials** \( \psi \)
    - Encode neighborhood information.
    - How different is a pixel/patch’s label from that of its neighbor? (e.g. based on intensity/color/texture difference, edges)
How to Set the Potentials? Some Examples

• Unary potentials
  – E.g., color model, modeled with a Mixture of Gaussians

\[
\phi(x_i, y_i; \theta_\phi) = \log \sum_k \theta_\phi(x_i, k)p(k|x_i)\mathcal{N}(y_i; \bar{y}_k, \Sigma_k)
\]

⇒ Learn color distributions for each label
How to Set the Potentials? Some Examples

• Pairwise potentials
  – Potts Model
    \[ \psi(x_i, x_j; \theta_\psi) = \theta_\psi \delta(x_i \neq x_j) \]
    - Simplest discontinuity preserving model.
    - Discontinuities between any pair of labels are penalized equally.
    - Useful when labels are unordered or number of labels is small.

  – Extension: “contrast sensitive Potts model”
    
    \[ \psi(x_i, x_j, g_{ij}(y); \theta_\psi) = \theta_\psi g_{ij}(y) \delta(x_i \neq x_j) \]
    where
    
    \[ g_{ij}(y) = e^{-\beta \|y_i - y_j\|^2} \]
    \[ \beta = 2 \cdot \text{avg} \left( \|y_i - y_j\|^2 \right) \]
    - Discourages label changes except in places where there is also a large change in the observations.
Extension: Conditional Random Fields (CRF)

- **Idea:** Model conditional instead of joint probability

\[
E(x) = \sum_{i \in S} \left( \phi(D|x_i) + \sum_{j \in N_i} (\phi(D|x_i, x_j) + \psi(x_i, x_j)) \right) + \text{const}
\]

- **Energy formulation**

  - Unary potential
  \[
  \phi(D|x_i)
  \]

  - Pairwise potential
  \[
  \phi(D|x_i, x_j)
  \]

  - Uniform Prior (Potts Model)

  - Contrast Term

Slide credit: Phil Torr
Example: MRF for Image Segmentation

- MRF structure
  - Unary potential
    \( \phi(D | x_i) \)
  - Pairwise potential
    \( \phi(D | x_i, x_j) \)

Data (D)  Unary likelihood  Pair-wise Terms  MAP Solution

Pixels  Labels  Prior Potts model

Slide credit: Phil Torr
Energy Minimization

- **Goal:**
  - Infer the optimal labeling of the MRF.

- **Many inference algorithms are available, e.g.:**
  - Simulated annealing
  - Iterated conditional modes (ICM)
  - Belief propagation
  - Graph cuts
  - Variational methods
  - Monte Carlo sampling

- **Recently, Graph Cuts have become a popular tool:**
  - Only suitable for a certain class of energy functions.
  - But the solution can be obtained very fast for typical vision problems (~1MPixel/sec).

*What you saw in the movie.*

*Too simple.*

*Last lecture*

*Use this one!*

*For more complex problems*
Topics of This Lecture

• Recap: Exact inference
  – Factor Graphs
  – Sum-Product/Max-Sum Belief Propagation
  – Junction Tree algorithm

• Applications of Markov Random Fields
  – Motivation
  – Unary potentials
  – Pairwise potentials

• Solving MRFs with Graph Cuts
  – Graph cuts for image segmentation
  – s-t mincut algorithm
  – Extension to non-binary case
  – Applications
Graph Cuts for Binary Problems

- Idea: convert MRF into source-sink graph

Minimum cost cut can be computed in polynomial time
(max-flow/min-cut algorithms)

\[ w_{pq} = \exp \left\{ -\frac{\Delta I_{pq}}{2\sigma^2} \right\} \]

Boykov & Jolly, ICCV'01
Simple Example of Energy

\[ E(L) = \sum_{p} D_p(L_p) + \sum_{pq \in N} w_{pq} \cdot \delta(L_p \neq L_q) \]

- unary potentials
- pairwise potentials

\( \delta(L_p \neq L_q) \)

**t-links**

**n-links**

\( \sum_{p} D_p(t) \)

\( \sum_{p} D_p(s) \)

\( L_p \in \{s, t\} \) (binary object segmentation)

Slide adapted from Yuri Boykov
Adding Regional Properties

Regional bias example

Suppose $I^s$ and $I^t$ are given “expected” intensities of object and background

\[
D_p(s) \propto \exp \left( -\frac{\| I_p - I^s \|^2}{2\sigma^2} \right)
\]

\[
D_p(t) \propto \exp \left( -\frac{\| I_p - I^t \|^2}{2\sigma^2} \right)
\]

NOTE: hard constrains are not required, in general.
Adding Regional Properties

“expected” intensities of object and background $I^s$ and $I^t$ can be re-estimated

$D_p(t) \propto \exp \left( -\| I_p - I^t \|^2 / 2\sigma^2 \right)$

$D_p(s) \propto \exp \left( -\| I_p - I^s \|^2 / 2\sigma^2 \right)$

EM-style optimization

[Boykov & Jolly, ICCV’01]
References and Further Reading

• A gentle introduction to Graph Cuts can be found in the following paper:

• Try the GraphCut implementation at https://pub.ist.ac.at/~vnk/software.html