EE12A – Advanced Inference in Graphical Models
Fall 2011

Prof. Jeff Bilmes

University of Washington, Seattle
Department of Electrical Engineering
Fall Quarter, 2011
http://j.ee.washington.edu/~bilmes/classes/ee512a_fall_2011/

Lecture 1 - September 28th, 2011
Announcements

- Welcome to the class!
Class information

- WF 10:30-12:30 in MEB-242 (this room).
Class information

- WF 10:30-12:30 in MEB-242 (this room).
- Office hours: Wednesdays 12:30pm-1:30pm, or stop by when I’m available.
Class information

- WFi 10:30-12:30 in MEB-242 (this room).
- Office hours: Wednesdays 12:30pm-1:30pm, or stop by when I’m available.
- Class web page: http://j.ee.washington.edu/~bilmes/classes/ee512a_fall_2011 will have announcements, readings, homework assignments, copies of these slides, bboard, and so on.
Class information

- WFi 10:30-12:30 in MEB-242 (this room).
- Office hours: Wednesdays 12:30pm-1:30pm, or stop by when I’m available.
- Class web page: http://j.ee.washington.edu/~bilmes/classes/ee512a_fall_2011 will have announcements, readings, homework assignments, copies of these slides, bboard, and so on.
- We’ll have 2-5 homeworks this quarter. You’ll have about a week to turn them in. You can either print them out, or email me pdf files.
Class information

- WFi 10:30-12:30 in MEB-242 (this room).
- Office hours: Wednesdays 12:30pm-1:30pm, or stop by when I’m available.
- Class web page: http://j.ee.washington.edu/~bilmes/classes/ee512a_fall_2011 will have announcements, readings, homework assignments, copies of these slides, bboard, and so on.
- We’ll have 2-5 homeworks this quarter. You’ll have about a week to turn them in. You can either print them out, or email me pdf files.
- Copies of lecture slides available on the web (but only just before lecture or very soon thereafter).
Class information

- WFi 10:30-12:30 in MEB-242 (this room).
- Office hours: Wednesdays 12:30pm-1:30pm, or stop by when I’m available.
- Class web page: http://j.ee.washington.edu/~bilmes/classes/ee512a_fall_2011 will have announcements, readings, homework assignments, copies of these slides, bboard, and so on.
- We’ll have 2-5 homeworks this quarter. You’ll have about a week to turn them in. You can either print them out, or email me pdf files.
- Copies of lecture slides available on the web (but only just before lecture or very soon thereafter).
- Copies of (most) readings available on the web
Homework

- Again, about 2-5 this quarter.
Homework

- Again, about 2-5 this quarter.
- Problem sets, prove a theorem, etc.
Homework

- Again, about 2-5 this quarter.
- Problem sets, prove a theorem, etc.
- Alternatively programming projects, so you should be extremely familiar with at least one programming language (e.g., C, C++, or matlab).
Homework

- Again, about 2-5 this quarter.
- Problem sets, prove a theorem, etc.
- Alternatively programming projects, so you should be extremely familiar with at least one programming language (e.g., C, C++, or matlab).
- All homework must be turned in electronically via our class web page.
Final Project

- There will be a final project, consisting of a 4-page conference-style paper, and a 10-minute final presentation.
**Final Project**

- There will be a final project, consisting of a 4-page conference-style paper, and a 10-minute final presentation.
- There will be a few milestones (1-page project proposal, 1-page progress summaries) during the quarter. These are graded.
Final Project

- There will be a final project, consisting of a 4-page conference-style paper, and a 10-minute final presentation.
- There will be a few milestones (1-page project proposal, 1-page progress summaries) during the quarter. These are graded.
- The final project should be regarding graphical models - either as a user in an application, or as a researcher (i.e., new inference method, new proof, etc.).

Final project reports due Saturday, Dec 10th, at 11:45pm. All final project reports must be turned in electronically via our class web page.
Final Project

- There will be a final project, consisting of a 4-page conference-style paper, and a 10-minute final presentation.
- There will be a few milestones (1-page project proposal, 1-page progress summaries) during the quarter. These are graded.
- The final project should be regarding graphical models - either as a user in an application, or as a researcher (i.e., new inference method, new proof, etc.).
- The date of the final project is Friday, December 9th. Time and location TBD.
Final Project

- There will be a final project, consisting of a 4-page conference-style paper, and a 10-minute final presentation.

- There will be a few milestones (1-page project proposal, 1-page progress summaries) during the quarter. These are graded.

- The final project should be regarding graphical models - either as a user in an application, or as a researcher (i.e., new inference method, new proof, etc.).

- The date of the final project is Friday, December 9th. Time and location TBD.

- Final project reports due Saturday, Dec 10th, at 11:45pm.
There will be a final project, consisting of a 4-page conference-style paper, and a 10-minute final presentation.

There will be a few milestones (1-page project proposal, 1-page progress summaries) during the quarter. These are graded.

The final project should be regarding graphical models - either as a user in an application, or as a researcher (i.e., new inference method, new proof, etc.).

The date of the final project is Friday, December 9th. Time and location TBD.

Final project reports due Saturday, Dec 10th, at 11:45pm.

All final project relate assignment must be turned in electronically via our class web page.
Our texts

- There will be three sources of reading material we’ll use this term.
Our texts

There will be three sources of reading material we’ll use this term.

- Handouts written by me (these are being prepared now, and are not quite ready, but chapters should be done soon). Material here will be mostly on GM semantics and exact inference methods.
There will be three sources of reading material we’ll use this term.

- Handouts written by me (these are being prepared now, and are not quite ready, but chapters should be done soon). Material here will be mostly on GM semantics and exact inference methods.
- Two text books (see below). One is available for free electronically.
There will be three sources of reading material we’ll use this term.

- Handouts written by me (these are being prepared now, and are not quite ready, but chapters should be done soon). Material here will be mostly on GM semantics and exact inference methods.
- Two text books (see below). One is available for free electronically.
- Research papers (links will be given in the class slides and on the web).
Our texts

- There will be three sources of reading material we’ll use this term.
  - Handouts written by me (these are being prepared now, and are not quite ready, but chapters should be done soon). Material here will be mostly on GM semantics and exact inference methods.
  - Two text books (see below). One is available for free electronically.
  - Research papers (links will be given in the class slides and on the web).
- Also might pick up a copy of the recent book by Koller and Friedman.
Our texts

- There will be three sources of reading material we’ll use this term.
  - Handouts written by me (these are being prepared now, and are not quite ready, but chapters should be done soon). Material here will be mostly on GM semantics and exact inference methods.
  - Two text books (see below). One is available for free electronically.
  - Research papers (links will be given in the class slides and on the web).
- Also might pick up a copy of the recent book by Koller and Friedman.
- Lauritzen 1996 is a classic book on GMs.
Our texts

- There will be three sources of reading material we’ll use this term.
  - Handouts written by me (these are being prepared now, and are not quite ready, but chapters should be done soon). Material here will be mostly on GM semantics and exact inference methods.
  - Two text books (see below). One is available for free electronically.
  - Research papers (links will be given in the class slides and on the web).

- Also might pick up a copy of the recent book by Koller and Friedman.
- Lauritzen 1996 is a classic book on GMs.
- Two other books on Bayesian networks include Jensen 1996 and 2001.
There will be three sources of reading material we’ll use this term.

- Handouts written by me (these are being prepared now, and are not quite ready, but chapters should be done soon). Material here will be mostly on GM semantics and exact inference methods.
- Two text books (see below). One is available for free electronically.
- Research papers (links will be given in the class slides and on the web).

Also might pick up a copy of the recent book by Koller and Friedman.

- Lauritzen 1996 is a classic book on GMs.
- Two other books on Bayesian networks include Jensen 1996 and 2001.
Our two main texts

Announcements

- Reading assignment: Read the “trees.pdf” chapter soon to be posted on the web page.
- Slides from previous time this course was offered are at http://ssli.ee.washington.edu/~bilmes/ee512fa09/ but this time we’ll be different.
We need to find one makeup lecture this term.

<table>
<thead>
<tr>
<th>Lecture</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>9/28</td>
</tr>
<tr>
<td>L2</td>
<td>10/5</td>
</tr>
<tr>
<td>L3</td>
<td>10/7</td>
</tr>
<tr>
<td>L4</td>
<td>10/12</td>
</tr>
<tr>
<td>L5</td>
<td>10/14</td>
</tr>
<tr>
<td>L6</td>
<td>10/19</td>
</tr>
<tr>
<td>L7</td>
<td>10/21</td>
</tr>
<tr>
<td>L8</td>
<td>10/26</td>
</tr>
<tr>
<td>L9</td>
<td>10/28</td>
</tr>
<tr>
<td>L10</td>
<td>11/2</td>
</tr>
<tr>
<td>L11</td>
<td>11/4</td>
</tr>
<tr>
<td>L12</td>
<td>11/9</td>
</tr>
<tr>
<td>LXX</td>
<td>11/11</td>
</tr>
<tr>
<td>L13</td>
<td>11/16</td>
</tr>
<tr>
<td>L14</td>
<td>11/18</td>
</tr>
<tr>
<td>L15</td>
<td>11/23</td>
</tr>
<tr>
<td>LXX</td>
<td>11/25</td>
</tr>
<tr>
<td>L16</td>
<td>11/30</td>
</tr>
<tr>
<td>L17</td>
<td>12/2</td>
</tr>
<tr>
<td>L18</td>
<td>12/7</td>
</tr>
<tr>
<td>L19</td>
<td>12/9</td>
</tr>
</tbody>
</table>
Graphical models inference is hard,
Graphical model inference is hard,
Graphical model inference is hard, NP-hard,
Graphical model inference is hard, NP-hard, inapproximable.

Best of cases, exact inference is infeasible.
Graphical model inference is hard, NP-hard, inapproximable.
Best of cases, exact inference is infeasible.
Approximation is necessary
Graphical model inference is hard, NP-hard, inapproximable.

Best of cases, exact inference is infeasible.

Approximation is necessary

Plethora of approximation methods
Graphical models are hard, NP-hard, inapproximable. Best of cases, exact inference is infeasible. Approximation is necessary. Plethora of approximation methods. The course this term will concentrate on two broad approximation methods.
The variational approach encompasses many standard approximate inference methods, including:
The variational approach encompasses many standard approximate inference methods, including:

- sum-product
- cluster variational methods
- expectation-propagation
- mean field methods
- max-product
- linear programming relaxations
- conic programming relaxations

and is therefore worthy of study. Of particular interest is for the class of exponential models (which have strong relationships to convexity).
Many inference methods from computer vision have appeared recently.
Approximation Method: Move making

- Many inference methods from computer vision have appeared recently.
- Simplest of the ideas: use an efficient graph-cut approach to find the minimum energy configuration in a pairwise binary Markov random field.
Approximation Method: Move making

- Many inference methods from computer vision have appeared recently.
- Simplest of the ideas: use an efficient graph-cut approach to find the minimum energy configuration in a pairwise binary Markov random field.
- When is this optimal? When is this an approximation? How can we generalize this to non-binary variables, non-pairwise potentials, and richer potentials?
Approximation Method: Move making

• Many inference methods from computer vision have appeared recently.
• Simplest of the ideas: use an efficient graph-cut approach to find the minimum energy configuration in a pairwise binary Markov random field.
• When is this optimal? When is this an approximation? How can we generalize this to non-binary variables, non-pairwise potentials, and richer potentials?
• Many generalizations, including move making algorithms such as alpha-beta swaps, alpha expansions, fusion moves, and other recent more sophisticated and energy aware “move making” algorithms.
Approximation Method: Move making

- Many inference methods from computer vision have appeared recently.
- Simplest of the ideas: use an efficient graph-cut approach to find the minimum energy configuration in a pairwise binary Markov random field.
- When is this optimal? When is this an approximation? How can we generalize this to non-binary variables, non-pairwise potentials, and richer potentials?
- Many generalizations, including move making algorithms such as alpha-beta swaps, alpha expansions, fusion moves, and other recent more sophisticated and energy aware “move making” algorithms.
- Computer vision only?
Sampling, Monte Carlo, MCMC methods, importance sampling
Not covering this term

- Sampling, Monte Carlo, MCMC methods, importance sampling
- Search based methods, cut condition, value elimination, as done in CSP/SAT communities. This includes AND/OR search trees, and other modern search based methods.
Not covering this term

- Sampling, Monte Carlo, MCMC methods, importance sampling
- Search based methods, cut condition, value elimination, as done in CSP/SAT communities. This includes AND/OR search trees, and other modern search based methods.
- Beam pruning methods often go hand-in-hand with search based methods.
Some notation

- **Distributions**
  
  \[ p(x) \equiv p(x_1:N) \equiv p(x_1, \ldots, x_N) \equiv P_{X_1,\ldots,X_N}(X_1 = x_1, \ldots, X_N = x_N) \]

- **Subsets**

  \[ V \triangleq \{1, 2, \ldots, N\} \quad A, B \subseteq V \quad A = \{a_1, \ldots, a_{|A|}\} \quad (1) \]

  \[ X_A \triangleq \{X_{a_1}, X_{a_2}, \ldots, X_{a_{|A|}}\} \quad (2) \]

- **Example:** If \( A = \{1, 3, 7\} \) then \( X_A = \{X_1, X_3, X_7\} \) and

  \[ p(X_A = x_A | X_B = x_B) \equiv p(x_1, x_2 | x_3, x_4) \]

  if \( A = \{1, 2\}, B = \{3, 4\} \)

- **\( p(x_A) \) requires table of size** \( r^{|A|}, \ r = |D_X| \) where \( \forall i, x_i \in D_X \)

- \( \bar{x}^{(i)} \) and \( \bar{x}^{(j)} \) are different vector samples for \( i \neq j \).
What do we want to do with $p(x)$?

- **Marginal quantities**
  - Given $\bar{x}$ compute $p(\bar{x})$
  - Given $E \subseteq V$, $H = V \setminus E$ compute $p(\bar{x}_E) = \sum_{x_H} p(x_H, \bar{x}_E)$.

- Model relationship between two signals $x_1$ and $x_2$ (e.g., $x_1$ a feature vector, $x_2$ is a class or regression variable).
  - compute $p(\bar{x}_1, \bar{x}_2)$.
  - Given $\bar{x}_1$ compute $x^*_2 \in \arg\max_{x_2} p(\bar{x}_1, x_2)$ or equivalently $x^*_2 \in \arg\max_{x_2} p(x_2 | \bar{x}_1)$ \hspace{1cm} (3)

- **Learning**
  - Given $D$ find $\theta^* \in \arg\max_{\theta} J(\theta)$.
  - Given $\theta^*$, how can we interpret its values?
  - Given $\theta^*$, use $p_{\theta^*}(x)$ or $p_{\theta^*}(x_1, x_2)$ as the truth.
Learning depends on the loss functions

- Generative learning if \( L(D, \theta) = \sum_{j=1}^{|D|} \log p_\theta(x^{(j)}) \) (maximum likelihood).

With \( x^{(j)} \in \arg\max \ p_\theta(x_2, x_{(j)}) \), where \( \Delta(x^{(j)}, x_2) \) is a normalizing label loss, and \( \ell(\cdot) \) is a local point-wise margin-based loss.
Learning depends on the loss functions

- Generative learning if $L(D, \theta) = \sum_{j=1}^{D} \log p_{\theta}(x^{(j)})$ (maximum likelihood).
- Discriminative learning if $L(D, \theta) = \sum_{j=1}^{D} \log p_{\theta}(x_2^{(j)} | x_1^{(j)})$. 

Each requires a different form of optimization, but each form of optimization invariably entails computing quantities over $p(x)$ like those mentioned earlier. The need to efficiently compute with $p(x)$ over restricted sub-families of distributions.
Learning depends on the loss functions

- Generative learning if $L(D, \theta) = \sum_{j=1}^{\left| D \right|} \log p_\theta(x^{(j)})$ (maximum likelihood).

- Discriminative learning if $L(D, \theta) = \sum_{j=1}^{\left| D \right|} \log p_\theta(x_2^{(j)}|x_1^{(j)})$.

- Max-margin learning if

$$L(D, \theta) = \sum_{i=1}^{\left| D \right|} \ell \left( \Delta(x_2^{(j)}, x_2') - (\log p_\theta(x_2^{(j)}, x_1^{(j)}) - \log p_\theta(x_2', x_1^{(j)})) \right)$$

(4)

with $x_2' \in \text{argmax}_{x_2} p_\theta(x_2, x_1^{(j)})$, where $\Delta(x_2^{(j)}, x_2')$ is a normalizing labeling loss, and $\ell(.)$ is a local point-wise margin-based loss.
Learning depends on the loss functions

- **Generative learning** if $L(D, \theta) = \sum_{j=1}^{\lvert D \rvert} \log p_\theta(x^{(j)})$ (maximum likelihood).

- **Discriminative learning** if $L(D, \theta) = \sum_{j=1}^{\lvert D \rvert} \log p_\theta(x_2^{(j)} | x_1^{(j)})$.

- **Max-margin learning** if

$$L(D, \theta) = \sum_{i=1}^{\lvert D \rvert} \ell \left( \Delta(x_2^{(j)}, x'_2) - (\log p_\theta(x_2^{(j)}, x_1^{(j)}) - \log p_\theta(x'_2, x_1^{(j)})) \right)$$

with $x'_2 \in \text{argmax}_{x_2} p_\theta(x_2, x_1^{(j)})$, where $\Delta(x_2^{(j)}, x'_2)$ is a normalizing labeling loss, and $\ell(\cdot)$ is a local point-wise margin-based loss.

- Each requires a different form of optimization, but each form of optimization invariably entails computing quantities over $p(x)$ like those mentioned earlier.
Learning depends on the loss functions

- Generative learning if \( L(D, \theta) = \sum_{j=1}^{\lvert D \rvert} \log p_\theta(x^{(j)}) \) (maximum likelihood).
- Discriminative learning if \( L(D, \theta) = \sum_{j=1}^{\lvert D \rvert} \log p_\theta(x_2^{(j)}|x_1^{(j)}) \).
- Max-margin learning if

\[
L(D, \theta) = \sum_{i=1}^{\lvert D \rvert} \ell \left( \Delta(x_2^{(j)}, x_2') - (\log p_\theta(x_2^{(j)}, x_1^{(j)}) - \log p_\theta(x_2', x_1^{(j)})) \right)
\]

with \( x_2' \in \operatorname{arg\max}_{x_2} p_\theta(x_2, x_1^{(j)}) \), where \( \Delta(x_2^{(j)}, x_2') \) is a normalizing labeling loss, and \( \ell() \) is a local point-wise margin-based loss.

- Each requires a different form of optimization, but each form of optimization invariably entails computing quantities over \( p(x) \) like those mentioned earlier.
- The need to efficiently compute with \( p(x) \) over restricted sub-families of distributions.
Machine learning within restricted families

- Sample data, along with domain knowledge, used to select resulting $p(x)$ from $\mathcal{U}$ that is “close enough” to $p_{\text{true}}(x_1, \ldots, x_N)$.
- Searching within $\mathcal{U}$ is infeasible/impossible.
- Desire a restricted family $\mathcal{F} \subset \mathcal{U}$.

- Size of $\mathcal{U}$ too large
- $\mathcal{U}$ complex, local optima
- Actual solution in $\mathcal{F}$ better than possible solution in $\mathcal{U}$
- Framework for $\mathcal{F}$ but not $\mathcal{U}$
A graphical model is a visual, abstract, and mathematically formal description of properties of families of probability distributions (densities, mass functions).

There are many types of graphical models, for example:

$G_1$

$G_2$

$G_3$
Graphical models are encodings of families of probability distributions. For the most part, the encodings are done via a graph that formally specifies either a set (conditional) independence properties, or more fundamentally, a set of factorization properties.

This is a crucial idea to understand: a graphical model is a set of constraints that all family members must obey.

Graphical Models encode constraints by factorization requirements that all members of the family must obey.

Factorization requirements are often identical to conditional independence requirements.

Factorization, in general, allows sums to be distributed into products thereby making (exact) inference quantities more efficient than if factorization properties did not exist.
Graph Theory

- We’ll define what we need as we go along.
- Graph $G = (V, E)$ where $V$ is set of nodes (or vertices) and $E \subseteq V \times V$ is a set of edges. If $i, j \in V$ then $(i, j) \in E$ means that nodes $i$ and $j$ are connected.
- Nodes are in one-to-one correspondence to a set of random variables. For each $v \in V$ we have that $X_v$ is a random variable (r.v.). $X_V$ is the complete set of r.v.’s.
- A graphical model describes a family of distributions $p(x_V)$ over $X_V$. 

Each type of graphical model requires a certain type of graph (e.g., undirected, or DAG) and a set of rules (or properties) to define the GM.

A graphical model is a pair $\mathcal{F}(G, R) = ((V, E), R)$, a graph $G$ and a set of rules $R$ that define what the graphical model means.

a rule $r \in R$ is a predicate on a graph and a distribution, so $r(p, G) \in \{\text{true}, \text{false}\}$.

$(G, R)$ consists of a family of distributions over $x_V$ such that for all rules hold. That is

$$\mathcal{F}(G, R) = \{p : p \text{ is a distribution over } X_V \text{ and } , r(p, G) = \text{true}, \forall r \in R\}$$

$$\mathcal{F}(G, R) \subseteq \mathcal{U}$$
Rules

- Rules are what are required of every family member. Any $p \in \mathcal{F}(G, R)$ satisfies all rules $r \in R$ for $G$. Any $p \in \mathcal{U} \setminus \mathcal{F}(G, R)$ violates at least one rule for $G$.

- A $p \in \mathcal{U}$ might have more properties. $R$ is like a filter, lets in those $p$ that satisfy, but will let in those that satisfy more.

- Example $r \in R$ might be “if there are two nodes $u, v \in V$ that are neither directly nor indirectly connected in $G$ (i.e., there no path leading from $u$ to $v$ in $G$) then the corresponding random variables in $p$ are independent”
Questions about Rules
Needing to be mathematically proven

- For a given type of graphical model, can the rule set $R$ be listed in finite space and computed efficiently? (answer, yes).

- For a given type of graphical model, are there more than one set of rules that define a family? In other words, are there rule sets $R_1$ and $R_2$ such that $\mathcal{F}(G, R_1) = \mathcal{F}(G, R_2)$ for all $G$? Answer, yes.

- Much of the Lauritzen 1996 book studies graphs, rules (or Markov properties) and proves when the corresponding families are either identical, or subsets of each other.
Questions about Rules (cont.)

Needing to be mathematically proven

- Is there a smallest rule set? In other words, are there rules sets $R_1 \subset R_2$ such that $\mathcal{F}(G, R_1) = \mathcal{F}(G, R_2)$, and can we compute the smallest rule set $R'$ so that $\mathcal{F}(G, R') = \mathcal{F}(G, R)$ where $|R'|$ is minimal?

- Are there rule sets that are non-equivalent? I.e. $R_1$ and $R_2$ such that $\mathcal{F}(G, R_1) \neq \mathcal{F}(G, R_2)$ for some $G$? Answer, yes.

In general, much of graphical model theory is regarding deducing properties of rules and corresponding properties of graphs and the distributions they represent. This allows us to reason about graphs as a means of reasoning about families of distributions.
A society of rules

- $\mathcal{G}_N = \text{set of all undirected graphs over } N \text{ nodes.}$
- Consider
  \[ \mathcal{F}_N(R) = \bigcup_{G \in \mathcal{G}_N} \mathcal{F}(G, R) \tag{6} \]
- and
  \[ \mathcal{F}(R) = \bigcup_{N=1}^{\infty} \mathcal{F}_N(R) \tag{7} \]
  family of all distributions over any number of random variables that obeys rules $R$ for some undirected graph $G$.
- $R$ determines the type of graphical model.
- $R^{(\text{mrf})}$ rules for MRF, then $\mathcal{F}(R^{(\text{mrf})})$ are the distributions representable by MRF.
- $R^{(\text{bn})}$ rules for Bayesian network, then $\mathcal{F}(R^{(\text{bn})})$ are the distributions representable by BN.
Different families

- Families may be different.
- For a given graph $G$, we might have neither $\mathcal{F}(G, R^{(mrf)}) \subset \mathcal{F}(G, R^{(bn)})$ nor $\mathcal{F}(G, R^{(bn)}) \subset \mathcal{F}(G, R^{(mrf)})$.
- The relationship for the family in its entirety might be different. I.e., when we compare the set of all MRFs vs. the set of all BNs, i.e., $\mathcal{F}(R^{(mrf)})$ vs. $\mathcal{F}(R^{(bn)})$.
- Large part of GM research is understanding these relationships.
Markov random fields

- One type of graphical model (we’ll study in this course).
- Has its origin in statistical physics (Boltzmann distributions, Ising models of atomic spin) and image processing (grid-based models).

Example Ising model: Let $W = [w_{ij}]_{ij}$ be a matrix of weights. Note that many of these weights might be zero. Let $S = [s_i]_i$ be a vector of binary random variables, $s_i \in \{-1, +1\}$. Define the “energy” as

$$E = -\sum_{ij} s_i s_j w_{ij}$$

(8)

Then define a distribution over $S$ as

$$p(s) = \frac{1}{Z} \exp(-E/T)$$

(9)

where $T$ is the temperature of the model.
Most often $S$ corresponds to a grid (i.e., $S$ is really a matrix or 3D-matrix).

Ising model: $w_{ij}$ determines the interaction style of variables: if $w_{ij} = 0$ the no interaction. If $w_{ij} > 0$ then more probable for $s_i = s_j = \pm 1$. If $w_{ij} < 0$ then more probable for $s_i \neq s_j$.

We can think of matrix $W$ and vector $S$ as a graph, $G = (V, E)$ where $S$ corresponds to $V$ and $W$ corresponds to $E$ ($(i, j) = e \in E$ only when $w_{ij} \neq 0$).

We might expect that any Ising model $p \in \mathcal{F}(G, R^{(mrf)})$ for appropriately defined MRF rules.
The “Cliques” of a graph $G = (V, E)$, or $C(G)$, in a graph are the set of fully connected nodes.

If $C \in C(G)$ and $u, v \in C$ then $(u, v) \in E(G)$

In the following graph

cliques are $C = \{1, 2, 3, 4, 5\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}, \{3, 4\}, \{2, 4\}, \{2, 3, 4\}, \{2, 5\}$. 
Clique Factorization

- Given graph $G$ with cliques $C(G)$, consider a probability distribution that can be represented as follows:

$$p(x_V) = \frac{1}{Z} \prod_{C \in C(G)} \phi_C(x_C)$$  \hspace{1cm} (10)

$$Z = \sum_{x_V} \prod_{C \in C} \phi_C(x_C)$$  \hspace{1cm} (11)

- Actually, we don’t even need $Z$ since it is a constant and can be distributed into the factors in a variety of ways.
More formally, consider the following family:

\[ \mathcal{F}(G, R^{(cf)}) = \{ p : \forall C \in \mathcal{C}(G), \exists \psi_C(x_C) \geq 0 \} \]

and

\[ p(x_V) = \prod_{C \in \mathcal{C}(G)} \psi_C(x_C) \]  \hspace{1cm} (12)

are factors unique?
MaxClique Factorization

- The “MaxCliques” of a graph $G = (V, E)$, or $C^{(mc)}(G)$, in a graph are the set of fully connected nodes that can’t be enlarged.
- MaxCliques of previous graph are $\{\{1, 2, 3\}, \{2, 3, 4\}, \{2, 5\}\}$.
- New rule $R^{(mcf)}$ based on maxcliques define family
  \[
  \mathcal{F}(G, R^{(mcf)}) = \left\{ p : \forall C \in C^{(mc)}(G), \exists \psi_C(x_C) \geq 0 \right\}
  \]
  and $p(x_V) = \prod_{C \in C^{(mc)}} \psi_C(x_C)$ \hspace{1cm} (13)
Comparisons of families

How do $\mathcal{F}(G, R^{(cf)})$ and $\mathcal{F}(G, R^{(mcf)})$ compare?
Comparisons of families

How do $\mathcal{F}(G, R^{(cf)})$ and $\mathcal{F}(G, R^{(mcf)})$ compare?

Lemma 5.1

$\mathcal{F}(G, R^{(cf)}) \subseteq \mathcal{F}(G, R^{(mcf)})$
Comparisons of families

- How do $\mathcal{F}(G, R^{(cf)})$ and $\mathcal{F}(G, R^{(mcf)})$ compare?

**Lemma 5.1**

$\mathcal{F}(G, R^{(cf)}) \subseteq \mathcal{F}(G, R^{(mcf)})$

**Lemma 5.2**

$\mathcal{F}(G, R^{(cf)}) \supseteq \mathcal{F}(G, R^{(mcf)})$
Comparisons of families

- How do $\mathcal{F}(G, R^{(cf)})$ and $\mathcal{F}(G, R^{(mcf)})$ compare?

**Lemma 5.1**

$\mathcal{F}(G, R^{(cf)}) \subseteq \mathcal{F}(G, R^{(mcf)})$

**Lemma 5.2**

$\mathcal{F}(G, R^{(cf)}) \supseteq \mathcal{F}(G, R^{(mcf)})$

Therefore

**Corollary 5.3**

$\mathcal{F}(G, R^{(cf)}) = \mathcal{F}(G, R^{(mcf)})$

- Since rules are identical, we use rule $R^{(f)}$ and family $\mathcal{F}(G, R^{(f)})$.
- Often, it is not so obvious that different families are identical.
- Equally often, different families are indeed different.
Most of this material comes from the reading handouts that will soon be made available.