EE512A – Advanced Inference in Graphical Models

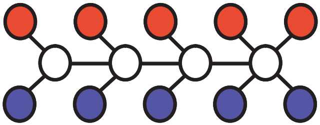
— Fall Quarter, Lecture 15 —

http://j.ee.washington.edu/~bilmes/classes/ee512a_fall_2014/

Prof. Jeff Bilmes

University of Washington, Seattle
Department of Electrical Engineering
http://melodi.ee.washington.edu/~bilmes

Nov 19th, 2014



Announcements

- Wainwright and Jordan *Graphical Models, Exponential Families, and Variational Inference* http://www.nowpublishers.com/product.aspx?product=MAL&doi=2200000001
- Read chapters 1,2, and 3 in this book. Start reading chapter 4.
- Assignment due Friday (Nov 21st) morning, 9:am. Final project proposals (one page max).

Class Road Map - EE512a

- L1 (9/29): Introduction, Families, Semantics
- L2 (10/1): MRFs, elimination, Inference on Trees
- L3 (10/6): Tree inference, message passing, more general queries, non-tree)
- L4 (10/8): Non-trees, perfect elimination, triangulated graphs
- L5 (10/13): triangulated graphs, k-trees, the triangulation process/heuristics
- L6 (10/15): multiple queries, decomposable models, junction trees
- L7 (10/20): junction trees, begin intersection graphs
- L8 (10/22): intersection graphs, inference on junction trees
- L9 (10/27): inference on junction trees, semirings,
- L10 (11/3): conditioning, hardness, LBP
- L11 (11/5): LBP, exponential models,
- L12 (11/10): exponential models, mean params and polytopes,
- L13 (11/12): polytopes, tree outer bound, Bethe entropy approx.
- L14 (11/17): Bethe entropy approx, loop series correction
- L15 (11/19): Hypergraphs, posets, Mobius, Kikuchi
- L16 (11/24):
- L17 (11/26):
- L18 (12/1):
- L19 (12/3):
- Final Presentations: (12/10):

Finals Week: Dec 8th-12th, 2014.

Bethe Variational Problem and LBP

Original variational representation of log partition function

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{15.14}$$

Approximate variational representation of log partition function

$$A_{\text{Bethe}}(\theta) = \sup_{\tau \in \mathbb{L}} \{\langle \theta, \tau \rangle + H_{\text{Bethe}}(\tau)\} \tag{15.15}$$

$$= \sup_{\tau \in \mathbb{L}} \left\{ \langle \theta, \tau \rangle + \sum_{v \in V(G)} H_v(\tau_v) - \sum_{(s,t) \in E(G)} I_{st}(\tau_{st}) \right\} \tag{15.16}$$

- Exact when $G = T$ but we do this for any G, still commutable
- we get an approximate log partition function, and approximate (pseudo) marginals (in $\mathbb{L}$), but this is perhaps much easier to compute.
- We can optimize this directly using a Lagrangian formulation.

Comparison of A and A_{Bethe}: loop series expansion

Proposition 15.2.2

Consider a pairwise MRF with binary variables, with $A_{Bethe}(\theta)$ being the optimized free energy evaluated at a LBP fixed point $\tau = (\tau_s, s \in V; \tau_{st}, (s,t) \in E(G))$. Then we have the following relationship with the cumulant function $A(\theta)$.

$$A(\theta) = A_{Bethe}(\theta) + \log \left\{ 1 + \sum_{\emptyset \neq \tilde{E} \subseteq E} \beta_{\tilde{E}} \prod_{s \in V} \mathbb{E}_{\tau_s} \left[(X_s - \tau_s)^{d_s(\tilde{E})} \right] \right\} \quad (15.6)$$

- For any $\tilde{E}$ such that $\exists s$ with $d_s(\tilde{E}) = 1$, inner term is zero and vanishes. why? Since $E_{\tau_s}\left[(X_s - \tau_s)^d\right]$ is the d^{th} central moment. Thus, terms in the sum only exists for generalized loops.
- The generalized loops give the correction!
- For trees, there are no generalized loops, and so if G is a tree then we have an equality between $A(\theta)$ and $A_{\text{Bethe}}(\theta)$ (recall both defs here).

General idea of Kikuchi

- Variational representation of log partition function

$$A(\theta) = \sup_{\mu \in \mathbb{M}(G)} \{\langle \theta, \mu \rangle - A^*(\mu)\} \quad (15.14)$$

- So far, we used a replacement for $-A^*(\mu)$ and $\mathbb{M}(G)$ inspired by trees.
- A tree is just a 1-tree, so one simple generalization would be to use a k-tree, for constant k, where k is not too large.
- More generally still, why not some other structure, like junction tree (embedable into a k-tree for k not too large).
- Junction trees are hypertrees (to be defined) that satisfy r.i.p. (special case of hypergraphs). Every clique need not be of size $k + 1$.
- So approach is the following: 1) derive expression for $-A^*(\mu)$ associated with a hypertree/junction tree; 2) generalize this expression for any hypergraph; 3) consider local consistency properties of hypertrees/junction tree; 4) use hypertrees local consistency property for generalized polytope associated with any hypergraph.
- $\Rightarrow$ Kikuchi variational approach ("clustered variational approximation")

Hypergraphs

- Recall, a graph $G = (V, E)$ is a set of nodes V and edges E where every $(s, t) = e \in E$ is only two nodes.

Hypergraphs

- Recall, a graph $G = (V, E)$ is a set of nodes V and edges E where every $(s, t) = e \in E$ is only two nodes.
- A hypergraph is a generalization of a graph.

Hypergraphs

- Recall, a graph $G = (V, E)$ is a set of nodes V and edges E where every $(s, t) = e \in E$ is only two nodes.
- A hypergraph is a generalization of a graph.

Definition 15.3.1 (hypergraph)

A *hypergraph* $H = (V, E)$ is a set of vertices V and a collection of hyperedges E, where each element $e \in E$ is a subset of V, so $\forall e \in E, e \subseteq V$.

Hypergraphs

- Recall, a graph $G = (V, E)$ is a set of nodes V and edges E where every $(s, t) = e \in E$ is only two nodes.
- A hypergraph is a generalization of a graph.

Definition 15.3.1 (hypergraph)

A *hypergraph* $H = (V, E)$ is a set of vertices V and a collection of hyperedges E, where each element $e \in E$ is a subset of V, so $\forall e \in E, e \subseteq V$.

- Thus, a hypergraph is a set system (V, E) where every $e \in E$ can consist of any number of nodes. I.e., we might have $\{v_1, v_2, \dots, v_{k_e}\} = e \in E(G)$ for a hypergraph.

Hypergraphs

- Recall, a graph $G = (V, E)$ is a set of nodes V and edges E where every $(s,t) = e \in E$ is only two nodes.
- A hypergraph is a generalization of a graph.

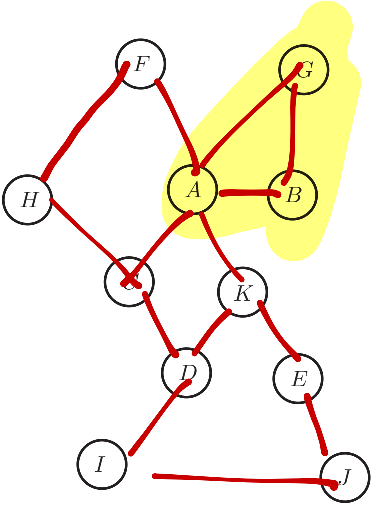
Definition 15.3.1 (hypergraph)

A *hypergraph* $H = (V, E)$ is a set of vertices V and a collection of hyperedges E, where each element $e \in E$ is a subset of V, so $\forall e \in E, e \subseteq V$.

- Thus, a hypergraph is a set system (V, E) where every $e \in E$ can consist of any number of nodes. I.e., we might have $\{v_1, v_2, \dots, v_{k_e}\} = e \in E(G)$ for a hypergraph.
- In a graph, $|e| = 2$. Thus, a graph is a (restricted) hypergraph, but not vice verse.

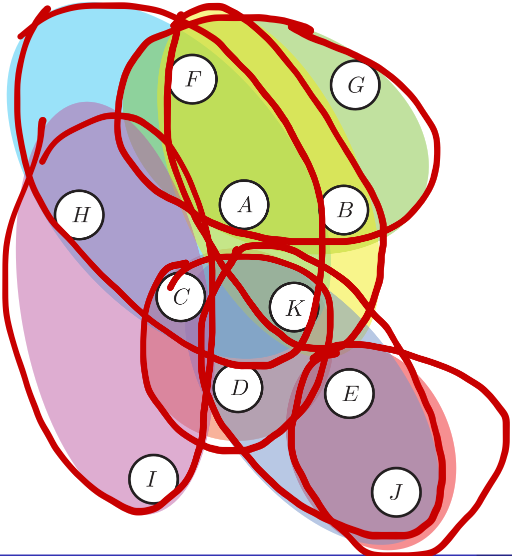
Drawing/Visualizing Hypergraphs

- A set of vertices, normally edges connect two nodes.



Drawing/Visualizing Hypergraphs

- Hypergraph: hyperedges are shaded regions, each region a vertex cluster



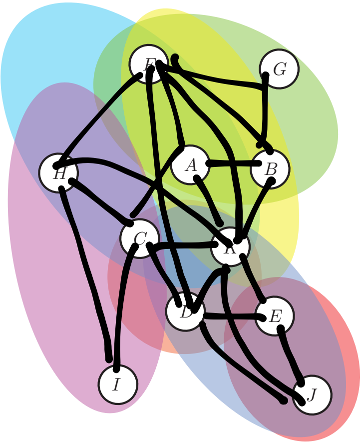
Hypergraphs and bipartite graphs

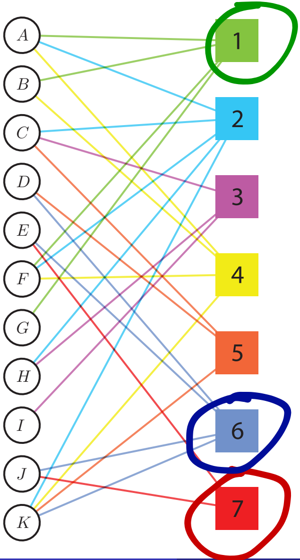
Hypergraphs can be represented by a bipartite $G = (V, F, E)$ graphs where V is a set of left-nodes, F is a set of right nodes, and E is a set of size-two edges. Right nodes are hyperedges in the hypergraphs.

Next slide shows an example.

Drawing/Visualizing Hypergraphs as Bipartite Graphs

- Hypergraph (shaded regions) on left, while bipartite graph representation on the right.





Graph of a hypergraph, conformal, and acyclic

- Let $H = (V, E)$ be a hypergraph with vertex set V and edge set E.

Graph of a hypergraph, conformal, and acyclic

- Let $H = (V, E)$ be a hypergraph with vertex set V and edge set E.
- The graph of a hypergraph $G(H)$ is a graph $G(H) = (V, E')$ where E' is a set of vertex pairs, and where there is an edge in E' for every pair of nodes that are in the same hyperedge.

Graph of a hypergraph, conformal, and acyclic

- Let $H = (V, E)$ be a hypergraph with vertex set V and edge set E.
- The graph of a hypergraph $G(H)$ is a graph $G(H) = (V, E')$ where E' is a set of vertex pairs, and where there is an edge in E' for every pair of nodes that are in the same hyperedge.
- A hypergraph H is conformal (to the graph $G(H)$) if every clique of $G(H)$ is contained in an edge of H.

Graph of a hypergraph, conformal, and acyclic

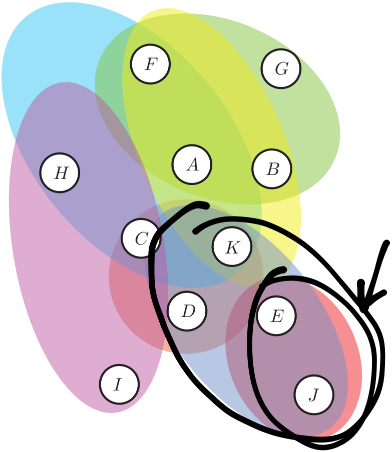
- Let $H = (V, E)$ be a hypergraph with vertex set V and edge set E.
- The graph of a hypergraph $G(H)$ is a graph $G(H) = (V, E')$ where E' is a set of vertex pairs, and where there is an edge in E' for every pair of nodes that are in the same hyperedge.
- A hypergraph H is conformal (to the graph $G(H)$) if every clique of $G(H)$ is contained in an edge of H.
- A hypergraph H is acyclic if H is conformal and $G(H)$ is chordal/triangulated.

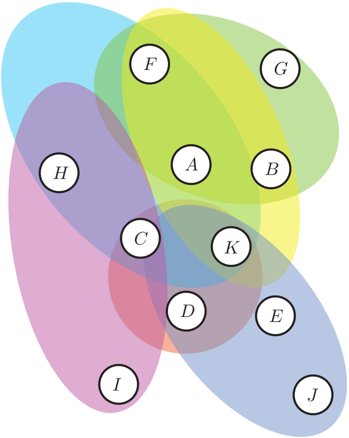
Hypergraph vs. Reduced Hypergraph

- A hypergraph is reduced if no edge is a subset of another edge.

Hypergraph vs. Reduced Hypergraph

- A hypergraph is reduced if no edge is a subset of another edge.
- Hypergraph (as shaded regions) on left, reduced hypergraph on the right (i.e., hyper edge $\{E, J\} \subset \{E, J, D, K\}$ is removed).



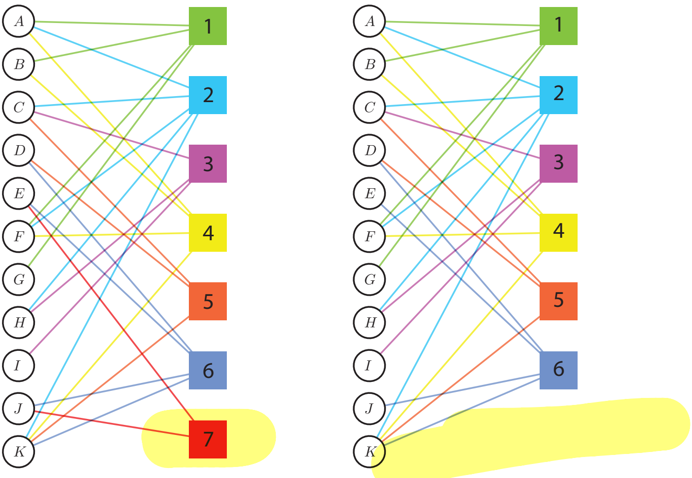


Hypergraph vs. Reduced Hypergraph

- A hypergraph is reduced if no edge is a subset of another edge.

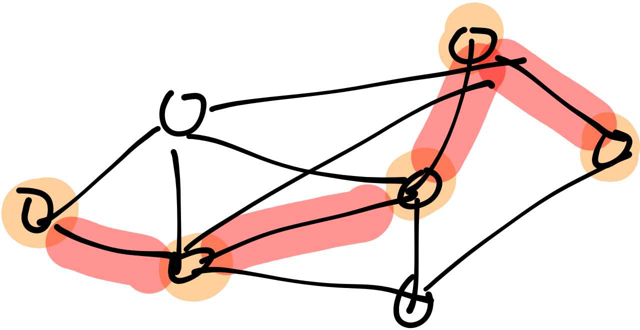
Hypergraph vs. Reduced Hypergraph

- A hypergraph is reduced if no edge is a subset of another edge.
- Hypergraph (as bipartite graph) on left, reduced hypergraph on the right (edge {E,J} removed).



Hypergraphs and Junction Trees

- A hypergraph path from $s \in V$ to $t \in V$ is a sequence of $k \geq 1$ edges $(e_1, e_2, \ldots, e_k)$ such that $s \in e_1$, $t \in e_k$, and $e_i \cap e_{i+1} \neq \emptyset$ for $1 \leq i < k$.



Hypergraphs and Junction Trees

- A hypergraph path from $s \in V$ to $t \in V$ is a sequence of $k \geq 1$ edges $(e_1, e_2, \ldots, e_k)$ such that $s \in e_1$, $t \in e_k$, and $e_i \cap e_{i+1} \neq \emptyset$ for $1 \leq i < k$.
- Recall, a junction tree is a tree of clusters of vertices of a graph, where the tree satisfies r.i.p. (i.e., induced sub-tree property).

Hypergraphs and Junction Trees

- A hypergraph path from $s \in V$ to $t \in V$ is a sequence of $k \geq 1$ edges $(e_1, e_2, \ldots, e_k)$ such that $s \in e_1$, $t \in e_k$, and $e_i \cap e_{i+1} \neq \emptyset$ for $1 \leq i < k$.
- Recall, a junction tree is a tree of clusters of vertices of a graph, where the tree satisfies r.i.p. (i.e., induced sub-tree propert).
- A hypertree is a hypergraph that can be transformed to a tree in a particular way, we've already seen them in the forms of junction trees.

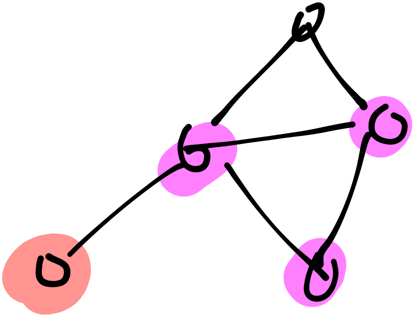
Hypergraphs and Junction Trees

- A hypergraph path from $s \in V$ to $t \in V$ is a sequence of $k \geq 1$ edges $(e_1, e_2, \ldots, e_k)$ such that $s \in e_1$, $t \in e_k$, and $e_i \cap e_{i+1} \neq \emptyset$ for $1 \leq i < k$.
- Recall, a junction tree is a tree of clusters of vertices of a graph, where the tree satisfies r.i.p. (i.e., induced sub-tree propert).
- A hypertree is a hypergraph that can be transformed to a tree in a particular way, we've already seen them in the forms of junction trees.
- In fact, a junction tree is a hypertree where the cliques (which are clusters of original graph nodes) in the junction tree are the edges of the hypertree.

Hypergraphs and Hypertrees

Definition 15.3.2 (leaf)

A vertex $v \in V(H)$ of H is called a *leaf* if it appears only in a single maximal hyper-edge $h \in H$.



Hypergraphs and Hypertrees

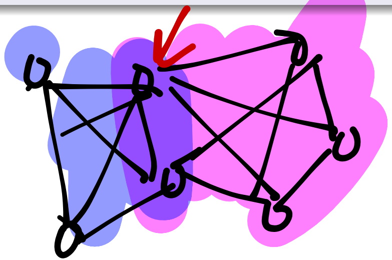
Definition 15.3.2 (leaf)

$\Leftarrow$ simplicial in $G(H)$

A vertex $v \in V(H)$ of H is called a *leaf* if it appears only in a single maximal hyper-edge $h \in H$.

Definition 15.3.3 (acyclic)

A hypergraph H is called *acyclic* if it is empty, or if it contains **a leaf** v such that induced hypergraph $H(V - \{v\})$ is acyclic (note, generalization of perfect elimination order in a triangulated graph, junction tree).



Hypergraphs and Hypertrees

Definition 15.3.2 (leaf)

A vertex $v \in V(H)$ of H is called a *leaf* if it appears only in a single maximal hyper-edge $h \in H$.

Definition 15.3.3 (acyclic)

A hypergraph H is called *acyclic* if it is empty, or if it contains **a leaf** v such that induced hypergraph $H(V - \{v\})$ is acyclic (note, generalization of perfect elimination order in a triangulated graph, junction tree).

Definition 15.3.4 (acyclic)

A hypergraph H is called *acyclic* if it is conformal to a graph that that is chordal. I.e., H is acyclic if $G(H)$ is triangulated.

Hypergraphs and Hypertrees

Definition 15.3.2 (leaf)

A vertex $v \in V(H)$ of H is called a *leaf* if it appears only in a single maximal hyper-edge $h \in H$.

Definition 15.3.3 (acyclic)

A hypergraph H is called *acyclic* if it is empty, or if it contains **a leaf** v such that induced hypergraph $H(V - \{v\})$ is acyclic (note, generalization of perfect elimination order in a triangulated graph, junction tree).

Definition 15.3.4 (acyclic)

A hypergraph H is called *acyclic* if it is conformal to a graph that that is chordal. I.e., H is acyclic if $G(H)$ is triangulated.

Definition 15.3.5 (hypertree)

A hypergraph H that is acyclic is called a *hypertree*.

Partially ordered set (poset)

- A partially ordered set (poset) is a set $\mathcal{P}$ of objects with an order.

Partially ordered set (poset)

- A partially ordered set (poset) is a set $\mathcal{P}$ of objects with an order.
- Set of objects $\mathcal{P}$ and a binary relation $\preceq$ which can be read as "is contained in" or "is part of" or "is less than or equal to".

$A = \{1, 2, 3, 4\}$ $\quad C = \{4, 5)$

$B = \{1, 4\}$ $\quad C \npreceq A$

$A \preceq B$

Partially ordered set (poset)

- A partially ordered set (poset) is a set $\mathcal{P}$ of objects with an order.
- Set of objects $\mathcal{P}$ and a binary relation $\preceq$ which can be read as "is contained in" or "is part of" or "is less than or equal to".
- For any $x, y \in \mathcal{P}$, we may ask is $x \preceq y$ which is either true or false.

Partially ordered set (poset)

- A partially ordered set (poset) is a set $\mathcal{P}$ of objects with an order.
- Set of objects $\mathcal{P}$ and a binary relation $\preceq$ which can be read as "is contained in" or "is part of" or "is less than or equal to".
- For any $x, y \in \mathcal{P}$, we may ask is $x \preceq y$ which is either true or false.
- In a poset, for any $x, y, z \in \mathcal{P}$ the following conditions hold (by definition):

For all $x, x \preceq x$.	(Reflexive)	(P1.)
If $x \preceq y$ and $y \preceq x$, then $x = y$	(Antisymmetriy)	(P2.)
If $x \preceq y$ and $y \preceq z$, then $x \preceq z$.	(Transitivity)	(P3.)

Partially ordered set (poset)

- A partially ordered set (poset) is a set $\mathcal{P}$ of objects with an order.
- Set of objects $\mathcal{P}$ and a binary relation $\preceq$ which can be read as "is contained in" or "is part of" or "is less than or equal to".
- For any $x, y \in \mathcal{P}$, we may ask is $x \preceq y$ which is either true or false.
- In a poset, for any $x, y, z \in \mathcal{P}$ the following conditions hold (by definition):

For all $x, x \preceq x$.	(Reflexive)	(P1.)
If $x \preceq y$ and $y \preceq x$, then $x = y$	(Antisymmetriy)	(P2.)
If $x \preceq y$ and $y \preceq z$, then $x \preceq z$.	(Transitivity)	(P3.)

- We can use the above to get other operators as well such as "less than" via $x \preceq y$ and $x \neq y$ implies $x \prec y$. And $x \succeq y$ is read "x contains y". And so on.

Partially ordered set

- Given two elements, we need not have either $x \preceq y$ or $y \preceq x$ be true, i.e., these elements might not be comparable. If for all $x, y \in V$ we have $x \preceq y$ or $y \preceq x$ then the poset is **totally ordered**.

Partially ordered set

- Given two elements, we need not have either $x \preceq y$ or $y \preceq x$ be true, i.e., these elements might not be comparable. If for all $x, y \in V$ we have $x \preceq y$ or $y \preceq x$ then the poset is **totally ordered**.
- If total order exists, then $x \succ y$ is identical to not $x \preceq y$.

Partially ordered set

- Given two elements, we need not have either $x \preceq y$ or $y \preceq x$ be true, i.e., these elements might not be comparable. If for all $x, y \in V$ we have $x \preceq y$ or $y \preceq x$ then the poset is **totally ordered**.
- If total order exists, then $x \succ y$ is identical to not $x \preceq y$.
- There exists only one element x which satisfies $x \preceq y$ for all y:

Partially ordered set

- Given two elements, we need not have either $x \preceq y$ or $y \preceq x$ be true, i.e., these elements might not be comparable. If for all $x, y \in V$ we have $x \preceq y$ or $y \preceq x$ then the poset is **totally ordered**.
- If total order exists, then $x \succ y$ is identical to not $x \preceq y$.
- There can exists only one element x which satisfies $x \preceq y$ for all y: If $x \preceq y$ for all y, and $z \preceq y$ for all y, then $z \preceq x$ and $x \preceq z$ implying $x = z$.

Partially ordered set

- Given two elements, we need not have either $x \preceq y$ or $y \preceq x$ be true, i.e., these elements might not be comparable. If for all $x, y \in V$ we have $x \preceq y$ or $y \preceq x$ then the poset is **totally ordered**.
- If total order exists, then $x \succ y$ is identical to not $x \preceq y$.
- There exists only one element x which satisfies $x \preceq y$ for all y: If $x \preceq y$ for all y, and $z \preceq y$ for all y, then $z \preceq x$ and $x \preceq z$ implying $x = z$. If it exists, we can name this element 0 (zero). The dual maximal element is called 1.

Partially ordered set

- Given two elements, we need not have either $x \preceq y$ or $y \preceq x$ be true, i.e., these elements might not be comparable. If for all $x, y \in V$ we have $x \preceq y$ or $y \preceq x$ then the poset is **totally ordered**.
- If total order exists, then $x \succ y$ is identical to not $x \preceq y$.
- There exists only one element x which satisfies $x \preceq y$ for all y: If $x \preceq y$ for all y, and $z \preceq y$ for all y, then $z \preceq x$ and $x \preceq z$ implying $x = z$.If it exists, we can name this element 0 (zero). The dual maximal element is called 1.
- We define a set of elements $x_1, x_2, \ldots, x_n$ as a chain if $x_1 \preceq x_2 \preceq \cdots \preceq x_n$, which means $x_1 \preceq x_2$ and $x_2 \preceq x_3$ and $\ldots x_{n-1} \preceq x_n$. Normally think of chain elements as distinct, but they need not be in general.

Partially ordered set

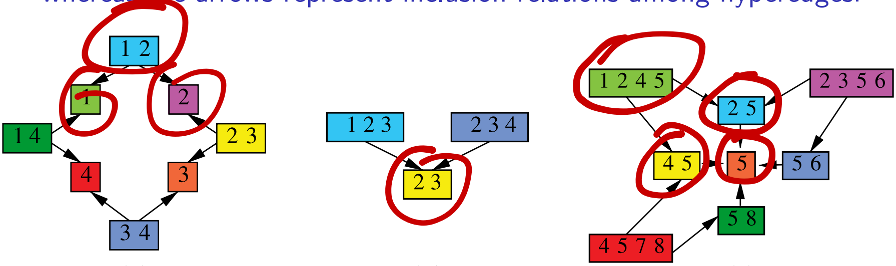
- Given two elements, we need not have either $x \preceq y$ or $y \preceq x$ be true, i.e., these elements might not be comparable. If for all $x, y \in V$ we have $x \preceq y$ or $y \preceq x$ then the poset is **totally ordered**.
- If total order exists, then $x \succ y$ is identical to not $x \preceq y$.
- There exists only one element x which satisfies $x \preceq y$ for all y: If $x \preceq y$ for all y, and $z \preceq y$ for all y, then $z \preceq x$ and $x \preceq z$ implying $x = z$.If it exists, we can name this element 0 (zero). The dual maximal element is called 1.
- We define a set of elements $x_1, x_2, \ldots, x_n$ as a chain if $x_1 \preceq x_2 \preceq \cdots \preceq x_n$, which means $x_1 \preceq x_2$ and $x_2 \preceq x_3$ and $\ldots x_{n-1} \preceq x_n$. Normally think of chain elements as distinct, but they need not be in general.
- The length of a chain of n elements is $n - 1$.

Hypergraph, edge representations

- It is possible to represent hypergraphs by only showing their hyperedges.

Hypergraph, edge representations

- It is possible to represent hypergraphs by only showing their hyperedges.
- Here, we see graphical representations of three hypergraphs. Subsets of nodes corresponding to hyperedges are shown in rectangles, whereas the arrows represent inclusion relations among hyperedges.

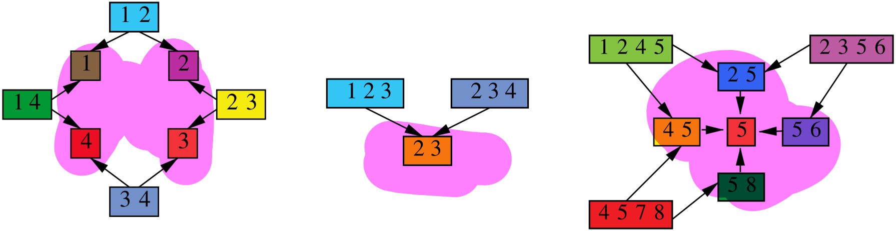


(a) (b) (c)

$\{1,2,3\} \longrightarrow \{1\}$

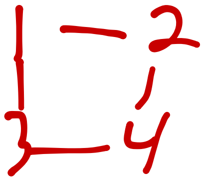
Hypergraph, edge representations

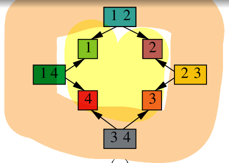
- It is possible to represent hypergraphs by only showing their hyperedges.
- Here, we see graphical representations of three hypergraphs. Subsets of nodes corresponding to hyperedges are shown in rectangles, whereas the arrows represent inclusion relations among hyperedges.



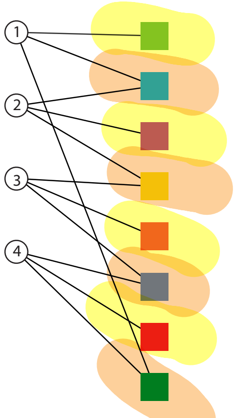
- Which ones, if any, are in reduced representation?

Hypergraph, edge representations, bipartite graphs



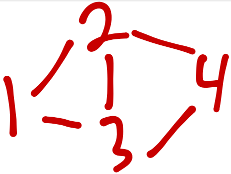


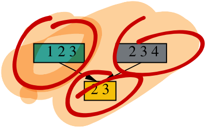
(a)



Edge-representations of hypergraphs and their corresponding bipartite graph representation.
An ordinary single 4-cycle graph represented as a hypergraph.

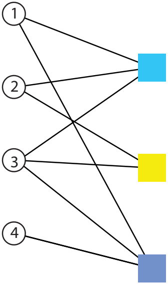
Hypergraph, edge representations, bipartite graphs





(b)

Edge-representations of hypergraphs and their corresponding bipartite graph representation.
A simple hypertree of "width" two.

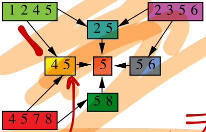


Hypergraph, edge representations, bipartite graphs

1 2 4 5

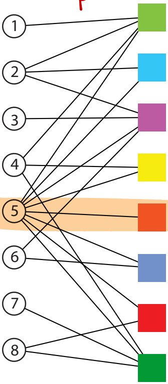
4 5 7 8

2 3 5 6



(c)

$\phi(x_4,x_5) = \sum_{x_1,x_2} \phi(x_1,x_2,x_4,x_5)$

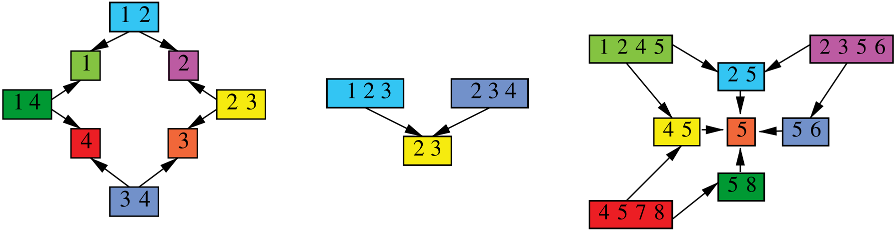


Edge-representations of hypergraphs and their corresponding bipartite graph representation.

A more complex hypertree of "width" three.

Hypergraph, edge representations, and posets

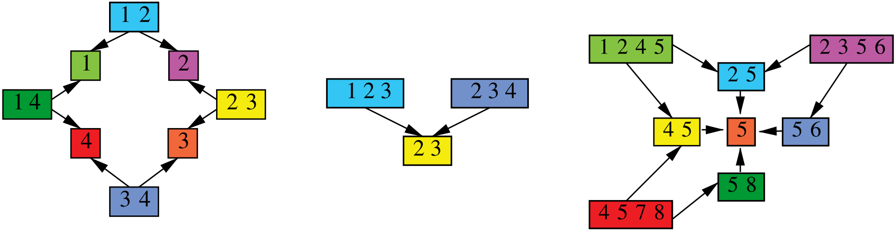
- Hypergraphs and edge representations.



(a) (b) (c)

Hypergraph, edge representations, and posets

- Hypergraphs and edge representations.

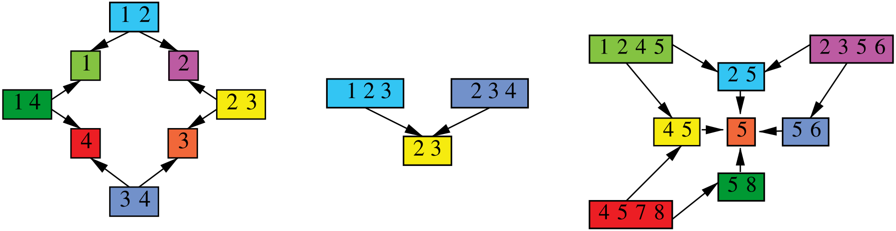


(a) (b) (c)

- Here, $a \to b$ if it is the case that $b \preceq a$ and there does not exist a c such that $b \preceq c \preceq a$, similar to a Hasse lattice diagram.

Hypergraph, edge representations, and posets

- Hypergraphs and edge representations.



(a) (b) (c)

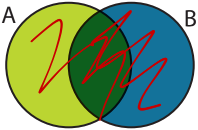
- Here, $a \to b$ if it is the case that $b \preceq a$ and there does not exist a c such that $b \preceq c \preceq a$, similar to a Hasse lattice diagram.
- Hence, the edges of a hypergraph form a partially ordered set.

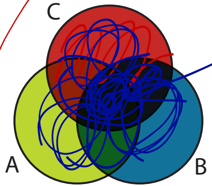
Inclusion-Exclusion

- Given ground set U and $A, B \subseteq U$, we may express the size of $A \cup B$ as: $|A \cup B| = |A| + |B| - |A \cap B|$.

Inclusion-Exclusion

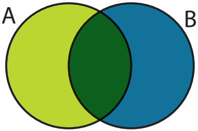
- Given ground set U and $A, B \subseteq U$, we may express the size of $A \cup B$ as: $|A \cup B| = |A| + |B| - |A \cap B|$.
- More generally, given $A, B, C \subseteq U$, then $|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C|$. Start by including, then excluding, and then including again.

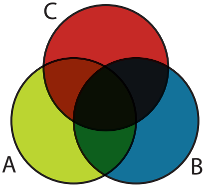




Inclusion-Exclusion

- Given ground set U and $A, B \subseteq U$, we may express the size of $A \cup B$ as: $|A \cup B| = |A| + |B| - |A \cap B|$.
- More generally, given $A, B, C \subseteq U$, then $|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C|$. Start by including, then excluding, and then including again.

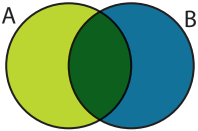


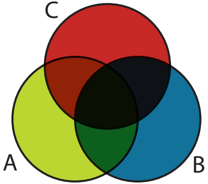


- In general, inclusion-exclusion refers to measuring a quantity by,

Inclusion-Exclusion

- Given ground set U and $A, B \subseteq U$, we may express the size of $A \cup B$ as: $|A \cup B| = |A| + |B| - |A \cap B|$.
- More generally, given $A, B, C \subseteq U$, then $|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C|$. Start by including, then excluding, and then including again.

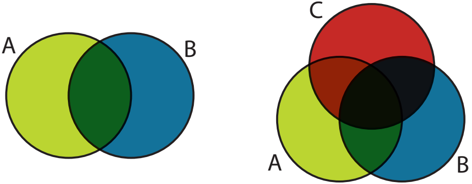




- In general, inclusion-exclusion refers to measuring a quantity by, first "adding" some other quantities and overshooting,

Inclusion-Exclusion

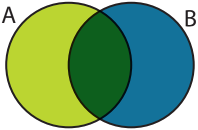
- Given ground set U and $A, B \subseteq U$, we may express the size of $A \cup B$ as: $|A \cup B| = |A| + |B| - |A \cap B|$.
- More generally, given $A, B, C \subseteq U$, then $|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C|$. Start by including, then excluding, and then including again.

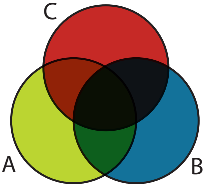


- In general, inclusion-exclusion refers to measuring a quantity by, first "adding" some other quantities and overshooting, then "subtracting" off some more quantities and undershooting,

Inclusion-Exclusion

- Given ground set U and $A, B \subseteq U$, we may express the size of $A \cup B$ as: $|A \cup B| = |A| + |B| - |A \cap B|$.
- More generally, given $A, B, C \subseteq U$, then $|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C|$. Start by including, then excluding, and then including again.

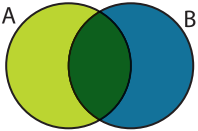


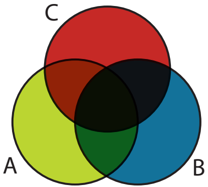


- In general, inclusion-exclusion refers to measuring a quantity by, first "adding" some other quantities and overshooting, then "subtracting" off some more quantities and undershooting,then "adding" some still other quantities and again overshooting,

Inclusion-Exclusion

- Given ground set U and $A, B \subseteq U$, we may express the size of $A \cup B$ as: $|A \cup B| = |A| + |B| - |A \cap B|$.
- More generally, given $A, B, C \subseteq U$, then $|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C|$. Start by including, then excluding, and then including again.

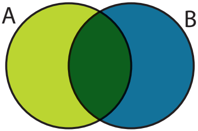


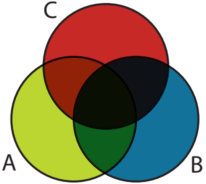


- In general, inclusion-exclusion refers to measuring a quantity by, first "adding" some other quantities and overshooting, then "subtracting" off some more quantities and undershooting,then "adding" some still other quantities and again overshooting,then "subtracting" off some still more quantities and again undershooting,

Inclusion-Exclusion

- Given ground set U and $A, B \subseteq U$, we may express the size of $A \cup B$ as: $|A \cup B| = |A| + |B| - |A \cap B|$.
- More generally, given $A, B, C \subseteq U$, then $|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C|$. Start by including, then excluding, and then including again.





- In general, inclusion-exclusion refers to measuring a quantity by, first "adding" some other quantities and overshooting, then "subtracting" off some more quantities and undershooting,then "adding" some still other quantities and again overshooting,then "subtracting" off some still more quantities and again undershooting, and so on, until we reach the right answer. "adding" might mean "multiplying", etc.

Entropic Inclusion-Exclusion

- Inclusion/exclusion also applies to entropy.

Entropic Inclusion-Exclusion

- Inclusion/exclusion also applies to entropy.
- That is, we have

$$
\begin{aligned}
H(X,Y) &= H(X) + H(Y) - I(X;Y) && (15.1)\\
H(X,Y,Z) &= H(X) + H(Y) + H(Z) && (15.2)\\
&\quad - I(X;Y) - I(X;Z) - I(Y;Z) && (15.3)\\
&\quad + I(X;Y;Z). && (15.4)
\end{aligned}
$$

and so on (see Yeung's book on information theory, the chapter on information measures).

Inclusion/Exclusion, general form for set measure

- Given $X_1, X_2, \ldots, X_n \subseteq U$,

Inclusion/Exclusion, general form for set measure

- Given $X_1, X_2, \dots, X_n \subseteq U$,
- Exclusion/exclusion formula for cardinality set measure $\mu(X) = |X|$:

$$\mu(\cap_{i=1}^n X_i) = \sum_{i=1}^n \mu(X_i) - \sum_{1 \le i < j \le n} \mu(X_i \cup X_j) \tag{15.5}$$

$$+ \sum_{1 \le i < j < k \le n} \mu(X_i \cup X_j \cup X_k) + \dots \tag{15.6}$$

$$+ (-1)^{n-1} \mu(X_1 \cup X_2 \cup \dots \cup X_n) \tag{15.7}$$

Inclusion/Exclusion, general form for set measure

- Given $X_1, X_2, \ldots, X_n \subseteq U$,
- Exclusion/exclusion formula for cardinality set measure $\mu(X) = |X|$:

$$\mu(\cap_{i=1}^n X_i) = \sum_{i=1}^n \mu(X_i) - \sum_{1 \le i < j \le n} \mu(X_i \cup X_j) \tag{15.5}$$

$$+ \sum_{1 \le i < j < k \le n} \mu(X_i \cup X_j \cup X_k) + \ldots \tag{15.6}$$

$$+ (-1)^{n-1} \mu(X_1 \cup X_2 \cup \ldots \cup X_n) \tag{15.7}$$

- A "dual" form has the form:

$$\mu(\cup_{i=1}^n X_i) = \sum_{i=1}^n \mu(X_i) - \sum_{1 \le i < j \le n} \mu(X_i \cap X_j) \tag{15.8}$$

$$+ \sum_{1 \le i < j < k \le n} \mu(X_i \cap X_j \cap X_k) + \ldots \tag{15.9}$$

$$+ (-1)^{n-1} \mu(X_1 \cap X_2 \cap \ldots \cap X_n) \tag{15.10}$$

Inclusion/Exclusion, general form for set measure

- Another (easier, shorter) way of writing these is as:

$$\mu(\cap_{i=1}^n X_i) = \sum_{k=1}^n (-1)^{k+1} \left(\sum_{1 \leq i_1 < i_2 < \cdots < i_k \leq n} \mu(X_{i_1} \cup \cdots \cup X_{i_k}) \right) \quad (15.11)$$

and

$$\mu(\cup_{i=1}^n X_i) = \sum_{k=1}^n (-1)^{k+1} \left(\sum_{1 \leq i_1 < i_2 < \cdots < i_k \leq n} \mu(X_{i_1} \cap \cdots \cap X_{i_k}) \right) \quad (15.12)$$

Möbius Inversion Lemma and Inclusion-Exclusion

- For any $A \subseteq V$, define two function $\Omega : 2^V \to \mathbb{R}_+$ and $\Upsilon : 2^V \to \mathbb{R}_+$.

Möbius Inversion Lemma and Inclusion-Exclusion

- For any $A \subseteq V$, define two function $\Omega : 2^V \to \mathbb{R}_+$ and $\Upsilon : 2^V \to \mathbb{R}_+$.
- Then the above inclusion-exclusion principle is one instance of the more general Möbius Inversion lemma, namely that each of the below two equations implies the other.

$$\forall A \subseteq V : \Upsilon(A) = \sum_{B : B \subseteq A} \Omega(B) \tag{15.13}$$

$$\forall A \subseteq V : \Omega(A) = \sum_{B : B \subseteq A} (-1)^{|A \setminus B|} \Upsilon(B) \tag{15.14}$$

Möbius Inversion Lemma and Inclusion-Exclusion

- For any $A \subseteq V$, define two function $\Omega : 2^V \to \mathbb{R}$ and $\Upsilon : 2^V \to \mathbb{R}$.
- Then the above inclusion-exclusion principle is one instance of the more general Möbius Inversion lemma, namely that each of the below two equations implies the other.

$$\forall A \subseteq V : \Upsilon(A) = \sum_{B : B \subseteq A} \Omega(B) \tag{15.13}$$

$$\forall A \subseteq V : \Omega(A) = \sum_{B : B \subseteq A} (-1)^{|A \setminus B|} \Upsilon(B) \tag{15.14}$$

- Möbius Inversion lemma is also used to prove the Hammersley-Clifford theorem (that factorization and Markov property definitions of families are identical for positive distributions).

Möbius Inversion Lemma and Inclusion-Exclusion

- For any $A \subseteq V$, define two function $\Omega : 2^V \to \mathbb{R}_+$ and $\Upsilon : 2^V \to \mathbb{R}_+$.
- Then the above inclusion-exclusion principle is one instance of the more general Möbius Inversion lemma, namely that each of the below two equations implies the other.

$$\forall A \subseteq V : \Upsilon(A) = \sum_{B : B \subseteq A} \Omega(B) \tag{15.13}$$

$$\forall A \subseteq V : \Omega(A) = \sum_{B : B \subseteq A} (-1)^{|A \setminus B|} \Upsilon(B) \tag{15.14}$$

- Möbius Inversion lemma is also used to prove the Hammersley-Clifford theorem (that factorization and Markov property definitions of families are identical for positive distributions).
- We use it here to come up with alternative expressions for the entropy and for the marginal polytope.

Hammersley Clifford Theorem

Theorem 15.3.6

(Hammersley and Clifford) Let $\mathcal{F}^+$ be the family of distributions with positive (and continuous in the continuous case) density (i.e., $p(x) > 0$ for all $p \in \mathcal{F}^+$). Then $\mathcal{F}^+ \cap \mathcal{F}(G, \mathcal{M}^{(f)}) = \mathcal{F}^+ \cap \mathcal{F}(G, \mathcal{M}^{(p)})$. ~~*In other words, the four above rules define the same family of distributions over a graph G when the distributions are restricted to be everywhere positive.*~~

- $\mathcal{F}(G, \mathcal{M}^{(f)})$ is the family we've seen before in this class, namely those distributions that factorize w.r.t. the cliques of graph G.

Hammersley Clifford Theorem

Theorem 15.3.6

(Hammersley and Clifford) Let $\mathcal{F}^+$ be the family of distributions with positive (and continuous in the continuous case) density (i.e., $p(x) > 0$ for all $p \in \mathcal{F}^+$). Then $\mathcal{F}^+ \cap \mathcal{F}(G, \mathcal{M}^{(f)}) = \mathcal{F}^+ \cap \mathcal{F}(G, \mathcal{M}^{(p)})$. In other words, the four above rules define the same family of distributions over a graph G when the distributions are restricted to be everywhere positive.

- $\mathcal{F}(G, \mathcal{M}^{(f)})$ is the family we've seen before in this class, namely those distributions that factorize w.r.t. the cliques of graph G.
- $\mathcal{F}(G, \mathcal{M}^{(p)})$ refers to the pairwise Markov property which states that if $u, v \in V(G)$ are not connected by an edge, then $X_u \perp\!\!\!\perp X_v \perp\!\!\!\perp X_{V \setminus \{u,v\}}$.

Hammersley Clifford Theorem

Theorem 15.3.6

(Hammersley and Clifford) Let $\mathcal{F}^+$ be the family of distributions with positive (and continuous in the continuous case) density (i.e., $p(x) > 0$ for all $p \in \mathcal{F}^+$). Then $\mathcal{F}^+ \cap \mathcal{F}(G, \mathcal{M}^{(f)}) = \mathcal{F}^+ \cap \mathcal{F}(G, \mathcal{M}^{(p)})$. In other words, the four above rules define the same family of distributions over a graph G when the distributions are restricted to be everywhere positive.

- $\mathcal{F}(G, \mathcal{M}^{(f)})$ is the family we've seen before in this class, namely those distributions that factorize w.r.t. the cliques of graph G.
- $\mathcal{F}(G, \mathcal{M}^{(p)})$ refers to the pairwise Markov property which states that if $u, v \in V(G)$ are not connected by an edge, then $X_u \perp\!\!\!\perp X_v \perp\!\!\!\perp X_{V \setminus \{u,v\}}$.
- In fact, $\mathcal{F}(G, \mathcal{M}^{(p)}) \supseteq \mathcal{F}(G, \mathcal{M}^{(f)})$. always holds. Hammersley and Clifford theorem (which uses Möbius inversion lemma) shows that $\mathcal{F}^+(G, R^{(p)}) \subseteq \mathcal{F}^+(G, R^{(f)})$, where $\mathcal{F}^+(G, \mathcal{M}) = \mathcal{F}^+ \cap \mathcal{F}(G, \mathcal{M})$.

Möbius Inversion Lemma

Lemma 15.3.7 (Möbius Inversion Lemma (for sets))

Let Υ and Ω be functions defined on the set of all subsets of a finite set V, taking values in an Abelian group (i.e., a set and an operator having properties closure, associativity, identity, and inverse, and for which the elements also commute, the real numbers being just one example). The following two equations imply each other.

$$\forall A \subseteq V : \Upsilon(A) = \sum_{B:B\subseteq A} \Omega(B) \tag{15.15}$$

$$\forall B \subseteq V \quad \Omega(B) = \sum_{C:C\subseteq B} (-1)^{|B\setminus C|}\Upsilon(C) \tag{15.16}$$

Proof of Möbius Inversion Lemma

Proof.

$$\sum_{B:B\subseteq A} \Omega(B) = \sum_{B:B\subseteq A} \sum_{C:C\subseteq B} (-1)^{|B\setminus C|} \Upsilon(C) \tag{15.17}$$

$$= \sum_{C:C\subseteq A} \sum_{B:C\subseteq B \& B\subseteq A} \Upsilon(C)(-1)^{|B\setminus C|} \tag{15.18}$$

$$= \sum_{C:C\subseteq A} \Upsilon(C) \sum_{B:C\subseteq B \& B\subseteq A} (-1)^{|B\setminus C|} \tag{15.19}$$

$$= \sum_{C:C\subseteq A} \Upsilon(C) \sum_{H:H\subseteq A\setminus C} (-1)^{|H|} \tag{15.20}$$

A

Proof of Möbius Inversion Lemma

Proof Cont.

Also, note that for any set D,

$$\sum_{H:H\subseteq D}(-1)^{|H|} = \sum_{i=0}^{|D|}\binom{|D|}{i}(-1)^i = \sum_{i=0}^{|D|}\binom{|D|}{i}(-1)^i(1)^{|D|-i} \tag{15.21}$$

$$= (1-1)^{|D|} = \begin{cases} 1 & \text{if } |D| = 0 \\ 0 & \text{otherwise} \end{cases} \tag{15.22}$$

which means that when we take $D = A \setminus C$, with $C \subseteq A$, we get

$$\sum_{H:H\subseteq A\setminus C}(-1)^{|H|} = \begin{cases} 1 & \text{if } A = C \\ 0 & \text{otherwise} \end{cases} \tag{15.23}$$

Proof of Möbius Inversion Lemma

Proof Cont.

This gives

$$\sum_{B:B\subseteq A} \Omega(B) = \sum_{C:C\subseteq A} \Upsilon(C)\mathbf{1}\{A = C\} = \Upsilon(A) \tag{15.24}$$

thus proving one direction. The other direction is very similar. $\square$

Möbius Inversion Lemma for posets

- Let $\mathcal{P}$ be a partially ordered set with binary relation $\preceq$.

Möbius Inversion Lemma for posets

- Let $\mathcal{P}$ be a partially ordered set with binary relation $\preceq$.
- A zeta function of a poset is a mapping $\zeta : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ defined by

$$\zeta(g,h) = \begin{cases} 1 & \text{if } g \preceq h, \\ 0 & \text{otherwise.} \end{cases} \tag{15.25}$$

Möbius Inversion Lemma for posets

- Let $\mathcal{P}$ be a partially ordered set with binary relation $\preceq$.
- A zeta function of a poset is a mapping $\zeta : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ defined by

$$\zeta(g,h) = \begin{cases} 1 & \text{if } g \preceq h, \\ 0 & \text{otherwise.} \end{cases} \tag{15.25}$$

- The Möbius function $\omega : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ is the multiplicative inverse of this function. It is defined recursively:

Möbius Inversion Lemma for posets

- Let $\mathcal{P}$ be a partially ordered set with binary relation $\preceq$.
- A zeta function of a poset is a mapping $\zeta : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ defined by

$$\zeta(g,h) = \begin{cases} 1 & \text{if } g \preceq h, \\ 0 & \text{otherwise.} \end{cases} \tag{15.25}$$

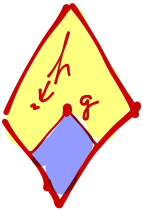
- The Möbius function $\omega : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ is the multiplicative inverse of this function. It is defined recursively:
- $\omega(g,g) = 1$ for all $g \in \mathcal{P}$

Möbius Inversion Lemma for posets

- Let $\mathcal{P}$ be a partially ordered set with binary relation $\preceq$.
- A zeta function of a poset is a mapping $\zeta : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ defined by

$$\zeta(g, h) = \begin{cases} 1 & \text{if } g \preceq h, \\ 0 & \text{otherwise.} \end{cases} \tag{15.25}$$

- The Möbius function $\omega : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ is the multiplicative inverse of this function. It is defined recursively:
- $\omega(g, g) = 1$ for all $g \in \mathcal{P}$
- $\omega(g, h) = 0$ for all $h : h \not\preceq g$.



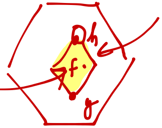
Möbius Inversion Lemma for posets

- Let $\mathcal{P}$ be a partially ordered set with binary relation $\preceq$.
- A zeta function of a poset is a mapping $\zeta : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ defined by

$$\zeta(g,h) = \begin{cases} 1 & \text{if } g \preceq h, \\ 0 & \text{otherwise.} \end{cases} \tag{15.25}$$

- The Möbius function $\omega : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ is the multiplicative inverse of this function. It is defined recursively:
- $\omega(g,g) = 1$ for all $g \in \mathcal{P}$
- $\omega(g,h) = 0$ for all $h : h \npreceq g$.
- Given $\omega(g,f)$ defined for f such that $g \preceq f \prec h$, we define

$$\omega(g,h) = - \sum_{\{f | g \preceq f \prec h\}} \omega(g,f) \tag{15.26}$$



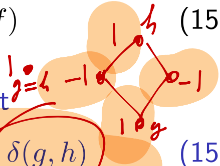
Möbius Inversion Lemma for posets

- Let $\mathcal{P}$ be a partially ordered set with binary relation $\preceq$.
- A zeta function of a poset is a mapping $\zeta : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ defined by

$$\zeta(g,h) = \begin{cases} 1 & \text{if } g \preceq h, \\ 0 & \text{otherwise.} \end{cases} \tag{15.25}$$

- The Möbius function $\omega : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ is the multiplicative inverse of this function. It is defined recursively:
- $\omega(g,g) = 1$ for all $g \in \mathcal{P}$
- $\omega(g,h) = 0$ for all $h : h \not\preceq g$.
- Given $\omega(g,f)$ defined for f such that $g \preceq f \prec h$, we define

$$\omega(g,h) = - \sum_{\{f | g \preceq f \prec h\}} \omega(g,f) \tag{15.26}$$



- Then, ω and ζ are multiplicative inverses, in that

$$\sum_{f \in \mathcal{P}} \omega(g,f)\zeta(f,h) = \sum_{\{f | g \preceq f \preceq h\}} \omega(g,f) = \delta(g,h) \tag{15.27}$$

General Möbius Inversion Lemma

Lemma 15.3.8 (General Möbius Inversion Lemma)

Given real valued functions Υ *and* Ω *defined on poset* $\mathcal{P}$*, then* $\Omega(h)$ *may be expressed via* $\Upsilon(\cdot)$ *via*

$$\Omega(h) = \sum_{g \preceq h} \Upsilon(g) \quad \text{for all } h \in \mathcal{P} \tag{15.28}$$

iff $\Upsilon(h)$ *may be expressed via* $\Omega(\cdot)$ *via*

$$\Upsilon(h) = \sum_{g \preceq h} \Omega(g)\omega(g,h) \quad \text{for all } h \in \mathcal{P} \tag{15.29}$$

When $\mathcal{P} = 2^V$ for some set V (so this means that the poset consists of sets and all subsets of an underlying set V) this can be simplified, where $\preceq$ becomes $\subseteq$; and $\succeq$ becomes $\supseteq$, like we saw above.
(see Stanley, "Enumerative Combinatorics" for more info.)

Back to Kikuchi: Möbius and expressions of factorization

- Suppose we are given marginals that factor w.r.t. a hypergraph $G = (V, E)$, so we have $\mu = (\mu_h, h \in E)$, then we can define new functions $\varphi = (\varphi_h, h \in E)$ via Möbius inversion lemma as follows

$$\log \varphi_h(x_h) \triangleq \sum_{g \preceq h} \omega(g, h) \log \mu_g(x_g) \tag{15.30}$$

Back to Kikuchi: Möbius and expressions of factorization

- Suppose we are given marginals that factor w.r.t. a hypergraph $G = (V, E)$, so we have $\mu = (\mu_h, h \in E)$, then we can define new functions $\varphi = (\varphi_h, h \in E)$ via Möbius inversion lemma as follows

$$\log \varphi_h(x_h) \triangleq \sum_{g \preceq h} \omega(g,h) \log \mu_g(x_g) \tag{15.30}$$

- From Möbius inversion lemma, this then gives us a new way to write the log marginals, i.e., as

$$\log \mu_h(x_h) = \sum_{g \preceq h} \log \varphi_g(x_g) \tag{15.31}$$

Back to Kikuchi: Möbius and expressions of factorization

- Suppose we are given marginals that factor w.r.t. a hypergraph $G = (V, E)$, so we have $\mu = (\mu_h, h \in E)$, then we can define new functions $\varphi = (\varphi_h, h \in E)$ via Möbius inversion lemma as follows

$$\log \varphi_h(x_h) \triangleq \sum_{g \preceq h} \omega(g,h) \log \mu_g(x_g) \qquad (15.30)$$

- From Möbius inversion lemma, this then gives us a new way to write the log marginals, i.e., as

$$\log \mu_h(x_h) = \sum_{g \preceq h} \log \varphi_g(x_g) \qquad (15.31)$$

- Key, when φ_h is defined as above, and G is a hypertree we have

$$p_\mu(x) = \prod_{h \in E} \varphi_h(x_h) \qquad (15.32)$$

$\Rightarrow$ general way to factorize a distribution that factors w.r.t. a hypergraph.

1-Tree factorization and Möbius

- When a 1-tree, we recover factorization we already know.

1-Tree factorization and Möbius

- When a 1-tree, we recover factorization we already know.
- That is, the hypergraph is just a tree (a 1-tree), then the hyperedges E consist of a poset consising of both standard-graph edges and standard graph nodes, where if $(u, v) = e \in E$ with $u, v \in V$ then $u \prec e$ and $v \prec e$.

1-Tree factorization and Möbius

- When a 1-tree, we recover factorization we already know.
- That is, the hypergraph is just a tree (a 1-tree), then the hyperedges E consist of a poset consising of both standard-graph edges and standard graph nodes, where if $(u, v) = e \in E$ with $u, v \in V$ then $u \prec e$ and $v \prec e$.
- In such case, from Equation (15.31), we have that for all $s \in V$, $\varphi_s(x_s) = \mu_s(x_s)$ and for all $(s, v) = e \in E$, we have:

$$\varphi_{st}(x_s, x_t) = \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s)\mu_t(x_t)} \tag{15.33}$$

1-Tree factorization and Möbius

- When a 1-tree, we recover factorization we already know.
- That is, the hypergraph is just a tree (a 1-tree), then the hyperedges E consist of a poset consising of both standard-graph edges and standard graph nodes, where if $(u, v) = e \in E$ with $u, v \in V$ then $u \prec e$ and $v \prec e$.
- In such case, from Equation (15.31), we have that for all $s \in V$, $\varphi_s(x_s) = \mu_s(x_s)$ and for all $(s, v) = e \in E$, we have:

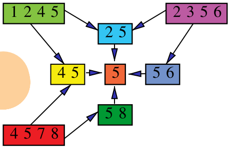
$$\varphi_{st}(x_s, x_t) = \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s)\mu_t(x_t)} \tag{15.33}$$

- Gives us the tree factorization we saw early in this course, namely:

$$p(x) = \prod_{h \in E} \varphi_h(x_h) = \prod_{s \in V} \mu_s(x_s) \prod_{(s,t)=e \in E} \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s)\mu_t(x_t)} \tag{15.34}$$

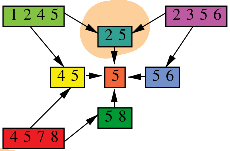
HyperTree factorization and Möbius

- For the more general hypertree, consider edge set $E = \{(12345), (2356), (4578), (25), (45), (56), (58), (5)\}$.
  Check: is this a junction tree of cliques?



HyperTree factorization and Möbius

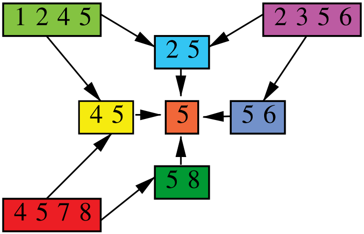
- For the more general hyper-tree, consider edge set $E = \{(12345),(2356),(4578),(25),(45),(56),(58),(5)\}$. Check: is this a junction tree of cliques?



- Then, from Eqn. (15.31), we get unaries $\varphi_s(x_s) = \mu_s(x_s)$ and pairwise (e.g., $\varphi_{25} = \mu_{25}/\mu_5$, etc.) and

$$\varphi_{1245} = \frac{\mu_{1245}}{\varphi_{25}\varphi_{45}\varphi_5} = \frac{\mu_{1245}}{\frac{\mu_{25}}{\mu_5}\frac{\mu_{45}}{\mu_5}\mu_5} = \frac{\mu_{1245}\mu_5}{\mu_{25}\mu_{45}} \tag{15.35}$$

expressions of factorization and Möbius

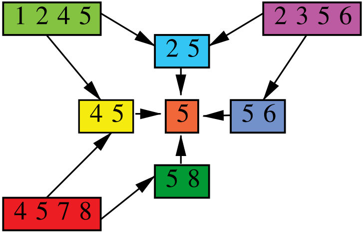


- Doing this for all maxcliques of the figure, we get a factorization of the form:

$$p_\mu = \frac{\mu_{1245}\mu_5}{\mu_{25}\mu_{45}} \frac{\mu_{2356}\mu_5}{\mu_{25}\mu_{56}} \frac{\mu_{4578}\mu_5}{\mu_{45}\mu_{58}} \frac{\mu_{25}}{\mu_5} \frac{\mu_{45}}{\mu_5} \frac{\mu_{56}}{\mu_5} \frac{\mu_{58}}{\mu_5} \mu_5 \tag{15.36}$$

$$= \frac{\mu_{1245}\mu_{2356}\mu_{4578}}{\mu_{25}\mu_{45}} \tag{15.37}$$

expressions of factorization and Möbius



- Doing this for all maxcliques of the figure, we get a factorization of the form:

$$p_\mu = \frac{\mu_{1245}\mu_5}{\mu_{25}\mu_{45}} \frac{\mu_{2356}\mu_5}{\mu_{25}\mu_{56}} \frac{\mu_{4578}\mu_5}{\mu_{45}\mu_{58}} \frac{\mu_{25}}{\mu_5} \frac{\mu_{45}}{\mu_5} \frac{\mu_{56}}{\mu_5} \frac{\mu_{58}}{\mu_5} \mu_5 \tag{15.36}$$

$$= \frac{\mu_{1245}\mu_{2356}\mu_{4578}}{\mu_{25}\mu_{45}} \tag{15.37}$$

- This is the same as the junction tree factorization with max cliques $\{1245\}$, $\{4578\}$, and $\{2356\}$ and separators $\{25\}$ and $\{45\}$.

New expressions of entropy

- Can express entropic quantities as well, such as the hyperedge entropy

$$H_h(\mu_h) = -\sum_{x_h} \mu_h(x_h) \log \mu_h(x_h) \tag{15.38}$$

and the "multi-information" function

$$I_h(\mu_h) = \sum_{x_h} \mu_h(x_h) \log \varphi_h(x_h) \tag{15.39}$$

New expressions of entropy

- Can express entropic quantities as well, such as the hyperedge entropy

$$H_h(\mu_h) = -\sum_{x_h} \mu_h(x_h) \log \mu_h(x_h) \tag{15.38}$$

and the "multi-information" function

$$I_h(\mu_h) = \sum_{x_h} \mu_h(x_h) \log \varphi_h(x_h) \tag{15.39}$$

- E.g., singletons $I_s(\mu_s) = -H(X_s)$ and pairs (in above hypergraph) are $I_{25}(\mu_{2,5}) = H(X_5) - H(X_2, X_5)$.

New expressions of entropy

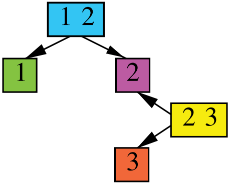
- Can express entropic quantities as well, such as the hyperedge entropy

$$H_h(\mu_h) = -\sum_{x_h} \mu_h(x_h) \log \mu_h(x_h) \tag{15.38}$$

and the "multi-information" function

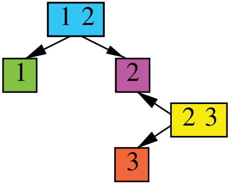
$$I_h(\mu_h) = \sum_{x_h} \mu_h(x_h) \log \varphi_h(x_h) \tag{15.39}$$

- E.g., singletons $I_s(\mu_s) = -H(X_s)$ and pairs (in above hypergraph) are $I_{25}(\mu_{2,5}) = H(X_5) - H(X_2, X_5)$.
- In the case of a single tree edge $h = (s,t)$, then $I_h(\mu_h) = I(X_s; X_t)$ the standard mutual information $(= H(X_s) + H(X_t) - H(X_s, X_t))$.



New expressions of entropy

- Can express entropic quantities as well, such as the hyperedge entropy
$$H_h(\mu_h) = -\sum_{x_h} \mu_h(x_h) \log \mu_h(x_h) \tag{15.38}$$
and the "multi-information" function
$$I_h(\mu_h) = \sum_{x_h} \mu_h(x_h) \log \varphi_h(x_h) \tag{15.39}$$
- E.g., singletons $I_s(\mu_s) = -H(X_s)$ and pairs (in above hypergraph) are $I_{25}(\mu_{2,5}) = H(X_5) - H(X_2, X_5)$.
- In the case of a single tree edge $h = (s,t)$, then $I_h(\mu_h) = I(X_s; X_t)$ the standard mutual information $(= H(X_s) + H(X_t) - H(X_s, X_t))$.



- By Eqn (15.32), overall entropy of any hypertree distribution becomes
$$H_{\text{hyper}}(\mu) = -\sum_{h \in E} I_h(\mu_h) \tag{15.40}$$

multi-information decomposition

- Using Möbius, and Eqn. (15.30) we can write

$$I_h(\mu_h) = \sum_{x_h} \mu_h(x_h) \log \varphi_h(x_h) = \sum_{x_h} \mu_h(x_h) \left(\sum_{g \preceq h} \omega(g,h) \log \mu_g(x_g) \right)$$

$$= \sum_{g \preceq h} \omega(g,h) \left\{ \sum_{x_h} \mu_h(x_h) \log \mu_g(x_g) \right\}$$

$$= \sum_{f \preceq h} \sum_{e \succeq f} \omega(f,e) \left\{ \sum_{x_f} \mu_f(x_f) \log \mu_f(x_f) \right\} = -\sum_{f \preceq h} c(f) H_f(\mu_f)$$

multi-information decomposition

- Using Möbius, and Eqn. (15.30) we can write

$$I_h(\mu_h) = \sum_{x_h} \mu_h(x_h) \log \varphi_h(x_h) = \sum_{x_h} \mu_h(x_h) \left(\sum_{g \preceq h} \omega(g,h) \log \mu_g(x_g) \right)$$

$$= \sum_{g \preceq h} \omega(g,h) \left\{ \sum_{x_h} \mu_h(x_h) \log \mu_g(x_g) \right\}$$

$$= \sum_{f \preceq h} \sum_{e \succeq f} \omega(f,e) \left\{ \sum_{x_f} \mu_f(x_f) \log \mu_f(x_f) \right\} = -\sum_{f \preceq h} c(f) H_f(\mu_f)$$

where we define overcounting numbers ($\sim$ shattering coefficient)

$$c(f) \triangleq \sum_{e \succeq f} \omega(f,e) \tag{15.41}$$

multi-information decomposition

- Using Möbius, and Eqn. (15.30) we can write

$$I_h(\mu_h) = \sum_{x_h} \mu_h(x_h) \log \varphi_h(x_h) = \sum_{x_h} \mu_h(x_h) \left(\sum_{g \preceq h} \omega(g,h) \log \mu_g(x_g) \right)$$

$$= \sum_{g \preceq h} \omega(g,h) \left\{ \sum_{x_h} \mu_h(x_h) \log \mu_g(x_g) \right\}$$

$$= \sum_{f \preceq h} \sum_{e \succeq f} \omega(f,e) \left\{ \sum_{x_f} \mu_f(x_f) \log \mu_f(x_f) \right\} = - \sum_{f \preceq h} c(f) H_f(\mu_f)$$

where we define overcounting numbers ($\sim$ shattering coefficient)

$$c(f) \triangleq \sum_{e \succeq f} \omega(f,e) \tag{15.41}$$

- This gives us a new expression for the hypertree entropy

$$H_{\text{hyper}}(\mu) = \sum_{h \in E} c(h) H_h(\mu_h) \tag{15.42}$$

Usable to get Kikuchi variational approximation

- Given arbitrary hypergraph now, we can generalize the hypertree-specific expressions above to this arbitrary hypergraph. This will give us a variational expression that approximates cumulant.

Usable to get Kikuchi variational approximation

- Given arbitrary hypergraph now, we can generalize the hypertree-specific expressions above to this arbitrary hypergraph. This will give us a variational expression that approximates cumulant.
- Given hypergraph $G = (V, E)$, we have

$$p_\theta(x) \propto \exp\left\{\sum_{h \in E} \sigma_h(x_h)\right\} \tag{15.43}$$

using same form of parameterization.

Usable to get Kikuchi variational approximation

- Given arbitrary hypergraph now, we can generalize the hypertree-specific expressions above to this arbitrary hypergraph. This will give us a variational expression that approximates cumulant.
- Given hypergraph $G = (V, E)$, we have

$$p_\theta(x) \propto \exp\left\{\sum_{h \in E} \sigma_h(x_h)\right\} \tag{15.43}$$

  using same form of parameterization.
- Hypergraph will give us local marginal constraints on hypergraph pseudo marginals, i.e., for each $h \in E$, we form marginal $\tau_h(x_h)$ and define constraints, non-negative, and

$$\sum_{x_h} \tau_h(x_h) = 1 \tag{15.44}$$

Usable to get Kikuchi variational approximation

- Sum to one constraint:

$$\sum_{x_h} \tau_h(x_h) = 1 \tag{15.45}$$

Usable to get Kikuchi variational approximation

- Sum to one constraint:

$$\sum_{x_h} \tau_h(x_h) = 1 \tag{15.45}$$

- Local agreement via the hypergraph constraint. For any $g \preceq h$ must have marginalization condition

$$\sum_{x_{h \setminus g}} \tau_h(x_h) = \tau_g(x_g) \tag{15.46}$$

Usable to get Kikuchi variational approximation

- Sum to one constraint:

$$\sum_{x_h} \tau_h(x_h) = 1 \tag{15.45}$$

- Local agreement via the hypergraph constraint. For any $g \preceq h$ must have marginalization condition

$$\sum_{x_{h \setminus g}} \tau_h(x_h) = \tau_g(x_g) \tag{15.46}$$

- Define new polyhedral constraint set $\mathbb{L}_t(G)$

$$\mathbb{L}_t(G) = \{\tau \geq 0 | \text{ Equations (15.45) } \forall h, \text{ and (15.46) } \forall g \preceq h \text{ hold}\} \tag{15.47}$$

Kikuchi variational approximation

- Generalized approximate (app) entropy for the hypergraph:

$$H_{\text{app}} = \sum_{g \in E} c(g) H_g(\tau_g) \tag{15.48}$$

where H_g is hyperedge entropy and overcounting number defined by:

$$c(g) = \sum_{f \succeq g} \omega(g, f) \tag{15.49}$$

Kikuchi variational approximation

- This at last gets the Kikuchi variational approximation

$$A_{\mathsf{Kikuchi}}(\theta) = \max_{\tau \in \mathbb{L}_t(G)} \{\langle \theta, \tau \rangle + H_{\mathsf{app}}(\tau)\} \tag{15.50}$$

Kikuchi variational approximation

- This at last gets the Kikuchi variational approximation

$$A_{\text{Kikuchi}}(\theta) = \max_{\tau \in \mathbb{L}_t(G)} \{\langle \theta, \tau \rangle + H_{\text{app}}(\tau)\} \tag{15.50}$$

- For a graph, this is exactly $A_{\text{Bethe}}(\theta)$.

Kikuchi variational approximation

- This at last gets the Kikuchi variational approximation

$$A_{\mathsf{Kikuchi}}(\theta) = \max_{\tau \in \mathbb{L}_t(G)} \{\langle \theta, \tau \rangle + H_{\mathsf{app}}(\tau)\} \tag{15.50}$$

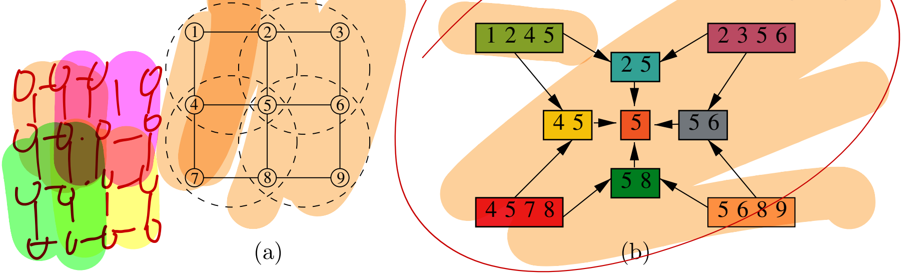
- For a graph, this is exactly $A_{\mathsf{Bethe}}(\theta)$.
- Also, if hypergraph is junction tree (r.i.p. holds, tree-local consistency implies global consistency), then also exact (although expensive, exponential in the tree-width to compute H_{app}).

Kikuchi variational approximation

- This at last gets the Kikuchi variational approximation

$$A_{\text{Kikuchi}}(\theta) = \max_{\tau \in \mathbb{L}_t(G)} \{\langle \theta, \tau \rangle + H_{\text{app}}(\tau)\} \tag{15.50}$$

- For a graph, this is exactly $A_{\text{Bethe}}(\theta)$.
- Also, if hypergraph is junction tree (r.i.p. holds, tree-local consistency implies global consistency), then also exact (although expensive, exponential in the tree-width to compute H_{app}).
- We can define message passing algorithms on the hypertree, and show that if it converges, it is a fixed point of the associated Lagrangian.



(a)   (b)

Sources for Today's Lecture

- Wainwright and Jordan *Graphical Models, Exponential Families, and Variational Inference* http://www.nowpublishers.com/product.aspx?product=MAL&doi=2200000001