

Exponential Families and Kernels

Lecture 4

Alexander J. Smola
Alex.Smola@nicta.com.au

Machine Learning Program
National ICT Australia
RSISE, The Australian National University

Outline

Exponential Families

- Maximum likelihood and Fisher information
- Priors (conjugate and normal)

Conditioning and Feature Spaces

- Conditional distributions and inner products
- Clifford Hammersley Decomposition

Applications

- Classification and novelty detection
- Regression

Applications

- Conditional random fields
- Intractable models and semidefinite approximations

Lecture 4

Conditional Random Fields

- Structured random variables
- Subspace representer theorem and decomposition
- Derivatives and conditional expectations

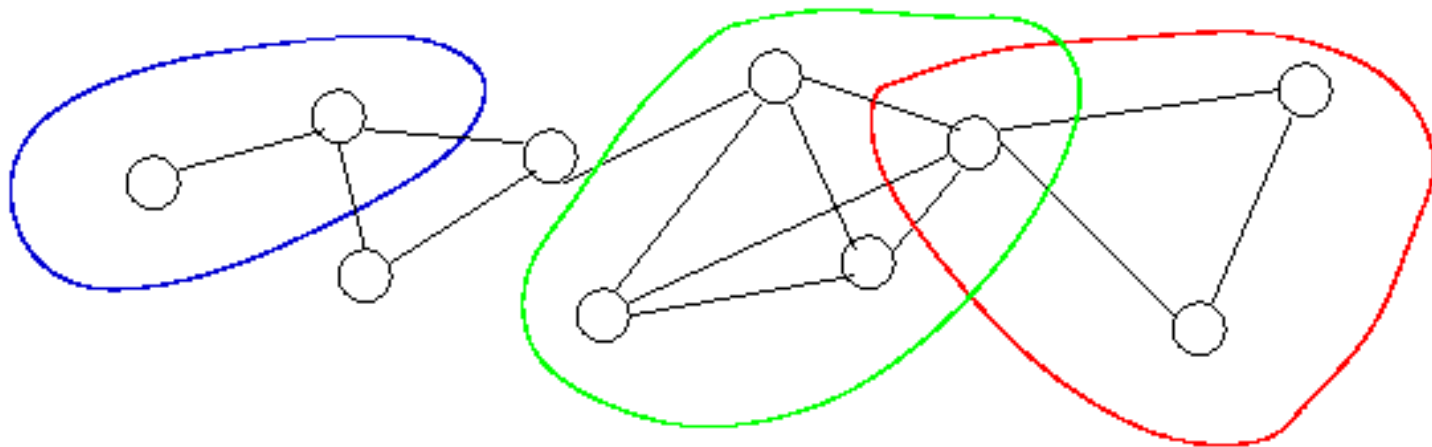
Inference and Message Passing

- Dynamic programming
- Message passing and junction trees
- Intractable cases

Semidefinite Relaxations

- Marginal polytopes
- Fenchel duality and entropy
- Relaxations for conditional random fields

Hammersley Clifford Corollary



Decomposition

The sufficient statistics $\phi(x)$ decompose according to

$$\phi(x) = (\dots, \phi_c(x_c), \dots)$$

Consequently we can write the kernel via

$$k(x, x') = \langle \phi(x), \phi(x') \rangle = \sum_c \langle \phi_c(x_c), \phi_c(x'_c) \rangle = \sum_c k_c(x_c, x'_c)$$

Computational Issues

Key Points

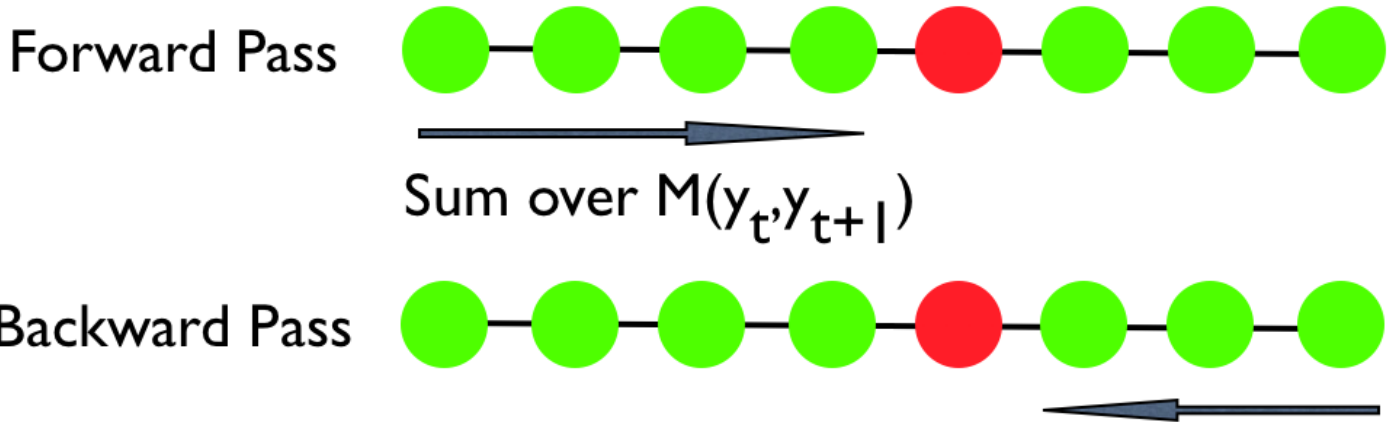
- Compute $g(\theta|x)$ via dynamic
- Assume stationarity of the model, that is θ_c does not depend on the position of the

Dynamic Programming

$$\begin{aligned} & g(\theta|x) \\ &= \log \sum_{y_1, \dots, y_T} \prod_{t=1}^T \underbrace{\exp(\langle \phi_{xy}(x_t, y_t), \theta_{xy} \rangle + \langle \phi_{yy}(y_t, y_{t+1}), \theta_{yy} \rangle)}_{M_t(y_t, y_{t+1})} \\ &= \log \sum_{y_1} \sum_{y_2} M_1(y_1, y_2) \sum_{y_3} M_2(y_2, y_3) \dots \sum_{y_T} M_T(y_{T-1}, y_T) \end{aligned}$$

So we can compute $g(\theta|x)$, $p(y_t|x, \theta)$ and $p(y_t, y_{t+1}|x, \theta)$ via dynamic programming.

Forward Backward Algorithm



Key Idea

- Store sum over all y_1, \dots, y_{t-1} (forward pass) and over all y_{t+1}, \dots, y_T as intermediate values
- We get those values for all positions t in one sweep.
- Extend this to message passing (when we have trees).

Minimization

Objective Function

$$-\log p(\theta|X, Y) = \sum_{i=1}^m -\langle \phi(x_i, y_i), \theta \rangle + g(\theta|x_i) + \frac{1}{2\sigma^2} \|\theta\|^2 + c$$

$$\partial_{\theta} -\log p(\theta|X, Y) = \sum_{i=1}^m -\phi(x_i, y_i) + \mathbf{E} [\phi(x_i, y_i)|x_i] + \frac{1}{\sigma^2}\theta$$

We only need $\mathbf{E} [\phi_{xy}(x_{it}, y_{it})|x_i]$ and $\mathbf{E} [\phi_{yy}(y_{it}, y_{i(t+1)})|x_i]$.

Kernel Trick

- Conditional expectations of $\Phi(x_{it}, y_{it})$ cannot be computed explicitly **but** inner products can.

$$\langle \phi_{xy}(x'_t, y'_t), \mathbf{E} [\phi_{xy}(x_t, y_t)|x] = \mathbf{E} [k((x'_t, y'_t), (x_t, y_t)|x]$$

- Only need marginals $p(y_t|x, \theta)$ and $p(y_t, y_{t+1}|x, \theta)$, which we get via dynamic programming.

Subspace Representer Theorem

Representer Theorem

Solutions of the MAP problem are given by

$$\theta \in \text{span}\{\phi(x_i, y) \text{ for all } y \in \mathcal{Y} \text{ and } 1 \leq i \leq n\}$$

Big Problem

$|\mathcal{Y}|$ could be huge, e.g. for sequence annotation 2^n .

Solution

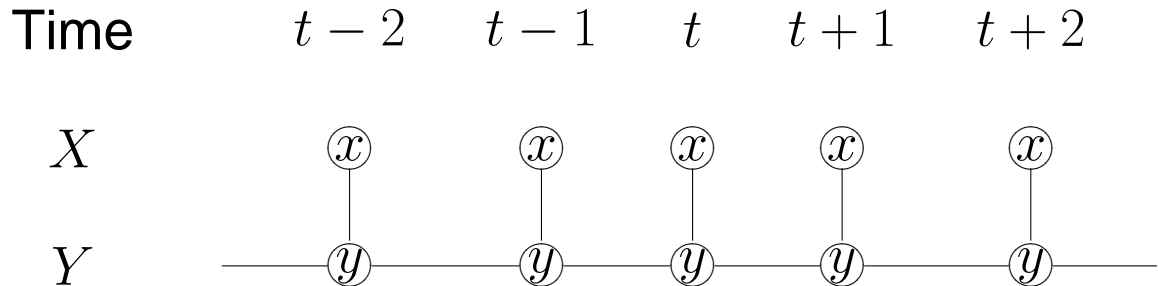
- Exploit decomposition of $\phi(x, y)$ into sufficient statistics on cliques.
- Restriction of \mathcal{Y} to cliques is much smaller.

$$\theta_c \in \text{span}\{\phi_c(x_{ci}, y_c) \text{ for all } y_c \in \mathcal{Y}_c \text{ and } 1 \leq i \leq n\}$$

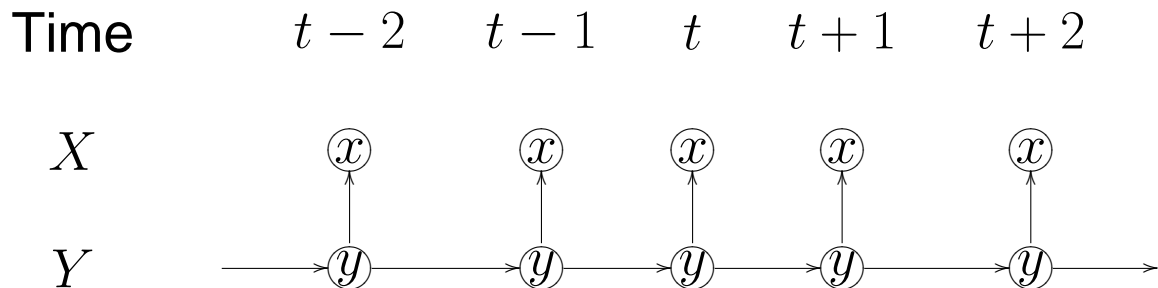
Rather than 2^n we now get $2^{|c|}$.

CRFs and HMMs

Conditional Random Field: maximize $p(y|x, \theta)$



Hidden Markov Model: maximize $p(x, y|\theta)$



Equivalence Theorem

Theorem

CRFs and HMMs yield **identical** probability estimates for $p(y|x, \theta)$, if the set of functions is equally expressive.

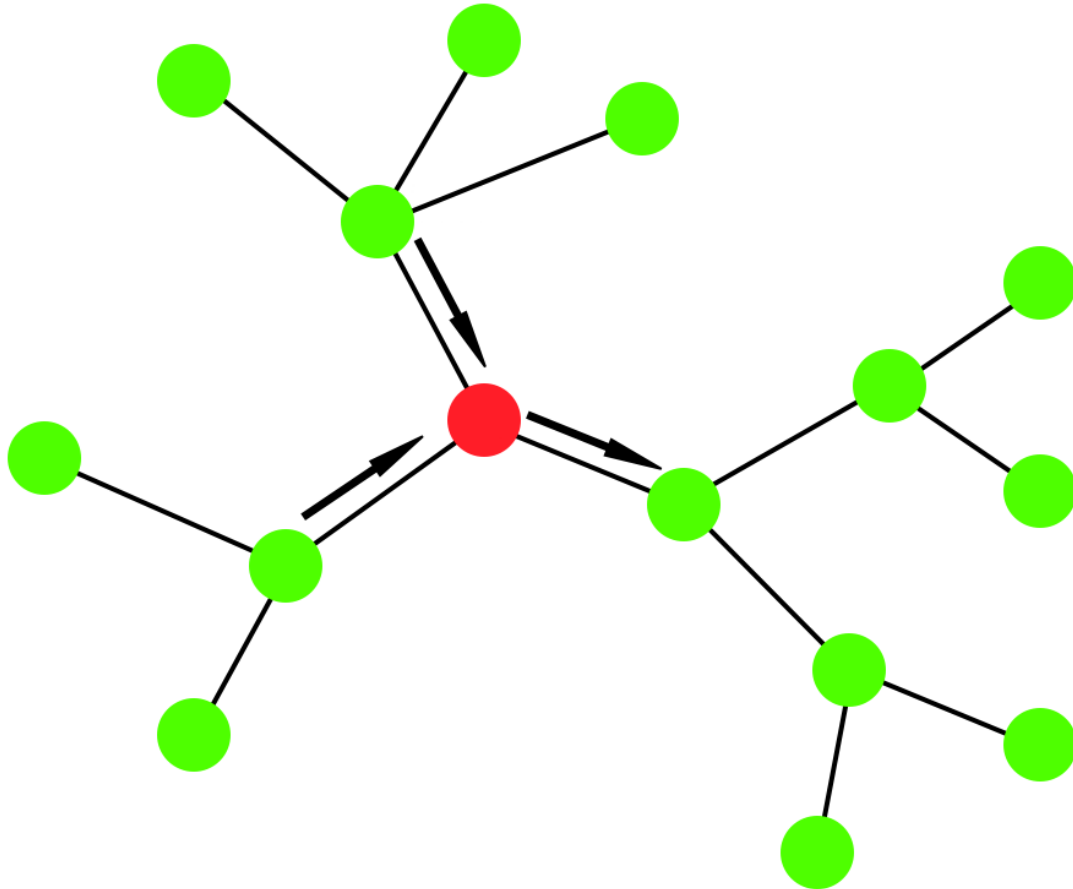
Proof

- Write out $p_{\text{CRF}}(y|x, \theta)$ and $p_{\text{HMM}}(x, y|\theta)$, and show that they only differ in the normalization.
- This disappears when computing $p_{\text{HMM}}(y|x, \theta)$.

Consequence

Differential training for current HMM implementations.

Message Passing



Message Passing

Idea

Extend the forward-backward idea to trees.

Algorithm

- Given clique potentials $M(y_i, y_j)$
- Initialize messages $\mu_{ij}(y_j) = 1$
- Update outgoing messages by

$$\mu_{ij}(y_j) = \sum_{y_i \in \mathcal{Y}_i} \prod_{k \neq j} \mu_{ki}(y_i) M_{ij}(y_i, y_j)$$

Here (i, k) is an edge in the graph.

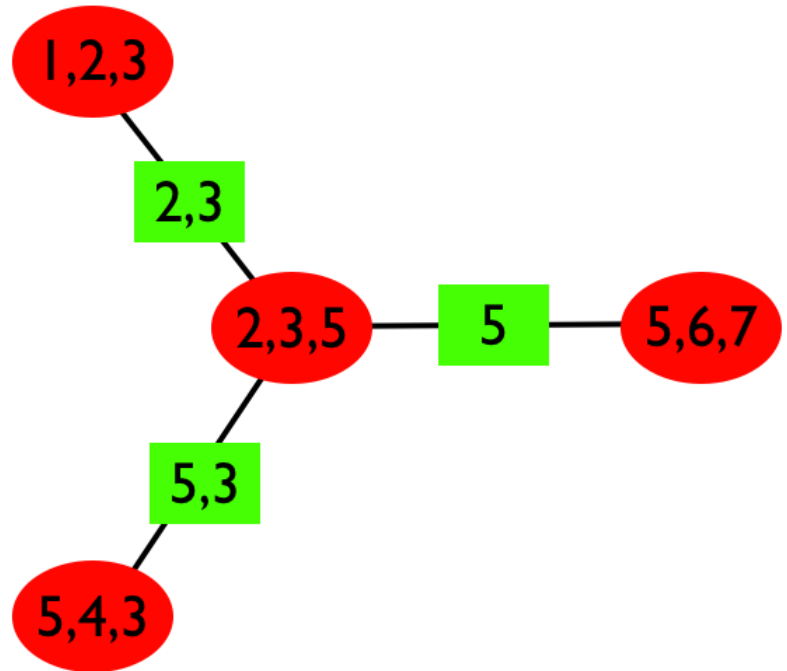
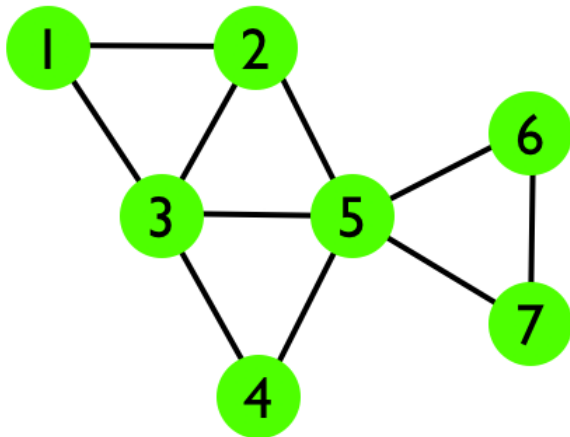
Theorem

The message passing algorithm converges after n iterations (n is diameter of graph).

Hack

Use this for graphs with **loops** and hope ...

Junction Trees



Stock standard algorithms available to transform graph into junction tree. Now we can use message passing . . .

Junction Tree Algorithm

Idea

Messages involve variables in the separator sets.

Algorithm

- Given clique potentials $M_c(y_c)$ and separator sets s .
- Initialize messages $\mu_{c,s}(y_s) = 1$
- Update outgoing messages by

$$\mu_{c,s}(y_s) = \sum_{y_c \setminus y_s} \prod_{s' \neq s} \mu_{c',s'}(y_{s'}) M_c(y_c)$$

Here s' is a separator set connecting c with c' .

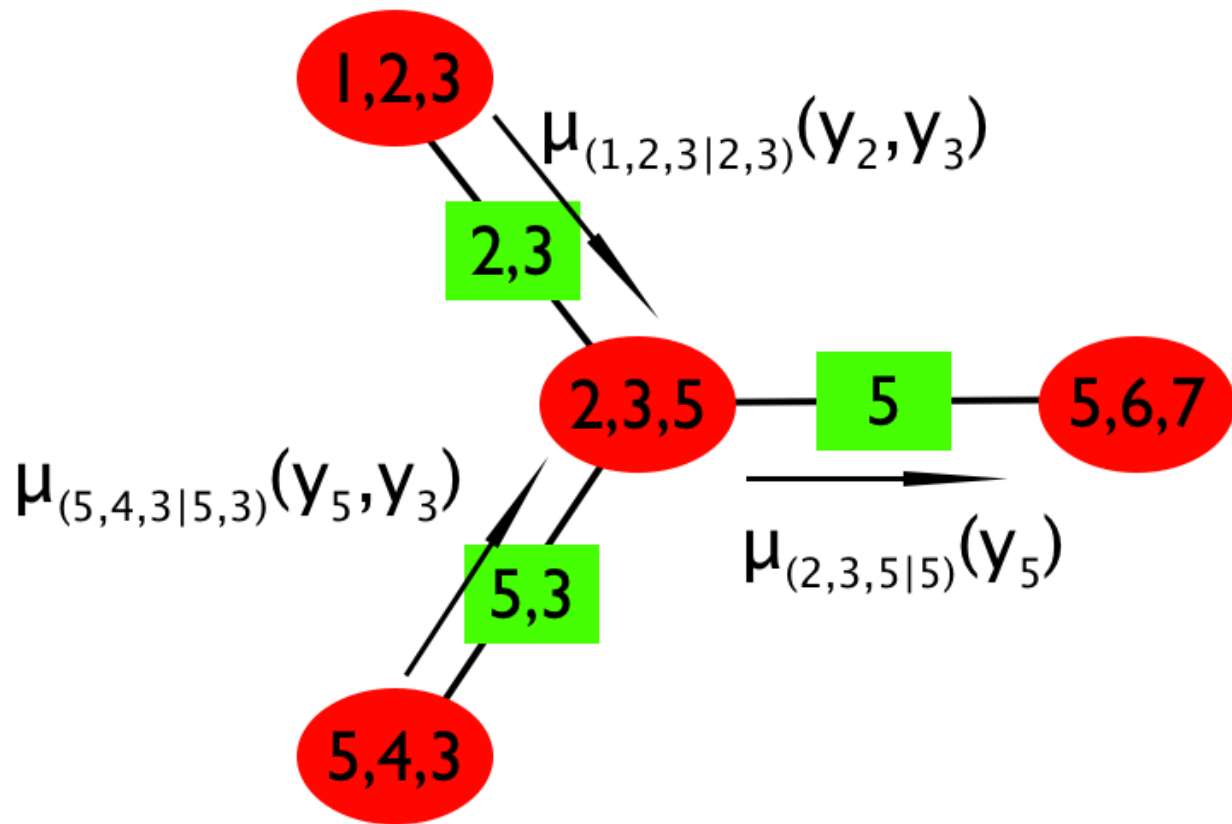
Theorem

The message passing algorithm converges after n iterations (n is diameter of the **hypergraph**).

Hack

Use this for graphs with **loops** and hope ...

Example



Problems

Scaling

The algorithm scales exponentially in the treewidth. Messages are of size $d^{|\mathcal{Y}_s|}$.

Convergence with loops

Use of message passing may or may not converge. No real proof available.

Workaround

Use a subset of the graph and solve the inference problem with this. Average over spanning trees.

Workaround

Use sampling methods for inference.

A Better Way

Fenchel Duality

Compute dual of log-partition function via

$$g^*(\mu) = \sup_{\theta \in \Theta} \langle \mu, \theta \rangle - g(\theta) \quad (\Theta \text{ is a convex domain})$$

Entropy and Expectation Parameters

The maximum of the optimization problem is obtained for $\mu = \partial_{\theta} g(\theta)$. This leads to

$$H = \int -\log p(x|\theta) p(x|\theta) d\theta = -\langle \mu(\theta), \theta \rangle + g(\theta) = -g^*(\mu)$$

Strong Duality

Dualizing again leads to

$$g(\theta) = \sup_{\mu \in M} \langle \theta, \mu \rangle + H(\mu)$$

Semidefinite Relaxation

Optimization Problem

$$g(\theta) = \sup_{\mu \in M} \langle \theta, \mu \rangle + H(\mu)$$

Here M is the set of all possible marginals.

Relaxations on M

The polytope M is convex (by duality), however it is hard to compute (as hard as $g(\theta)$). So we relax it to \tilde{M} by impose constraints on higher order moments, such as

- Interval and linear inequality constraints.
- SDP constraints on the covariance matrix.

Upper bound on $H(\mu)$

Gaussian bound on the covariance via $G(\mu)$. So we get

$$g(\theta) \leq \sup_{\mu \in \tilde{M}} \langle \theta, \mu \rangle + G(\mu)$$

Application to CRFs

Optimization Problem

$$\begin{aligned} & -\log p(\theta|X, Y) \\ &= \sum_{i=1}^m -\langle \phi(x_i, y_i), \theta \rangle + g(\theta|x_i) + \frac{1}{2\sigma^2} \|\theta\|^2 + c \\ &\leq \sum_{i=1}^m \sup_{\mu_i \in \tilde{M}_i} \langle \theta, \mu_i - \phi(x_i, y_i) \rangle + H(\mu_i) + \frac{1}{2\sigma^2} \|\theta\|^2 \end{aligned}$$

Technical Details

- Minimization over θ and μ_i can be swapped (saddle-point property of a convex-concave problem) to obtain dual problem in θ .
- Map from μ to moments in $y|x$ via invertible sufficient statistics map.
- Constrained max-det problem.

Summary

Conditional Random Fields

- Structured random variables
- Subspace representer theorem and decomposition
- Derivatives and conditional expectations

Inference and Message Passing

- Dynamic programming
- Message passing and junction trees
- Intractable cases

Semidefinite Relaxations

- Marginal polytopes
- Fenchel duality and entropy
- Relaxations for conditional random fields

Shameless Plugs

We are hiring. For details contact

- Alex.Smola@nicta.com.au (<http://www.nicta.com.au>)

Positions

- PhD scholarships
- Postdoctoral positions, Senior researchers
- Long-term visitors (sabbaticals etc.)

More details on kernels

- <http://www.kernel-machines.org>
- <http://www.learning-with-kernels.org>
Schölkopf and Smola: Learning with Kernels

Machine Learning Summer School

- <http://www.mlss.cc>
- MLSS'05 Canberra, Australia, 23/1-5/2/2005