Best Choice Edge Grafting For Efficient Learning of Markov Random Fields

Walid Chaabene, Bert Huang

Virginia Tech

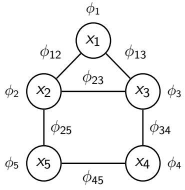
walidch, bhuang @vt.edu

December 11, 2018

Overview

Outline

Introduction



Pairwise Markov Random Fields

A graphical model that represents joint probability distributions.

$$G(V,E) : \begin{cases} V : \text{set of } n \text{ nodes (variables);} \\ E : \text{set of edges (parametric interactions).} \end{cases}$$

$$p_{\mathbf{w}}(X) = \frac{1}{Z(\mathbf{w})} \prod_{i \in V} \phi_i(x;\mathbf{w}) \prod_{(i,j) \in E} \phi_{ij}(x;\mathbf{w}), \qquad (1)$$

where:

$$\phi_c(x;\mathbf{w}) = \exp\left(\sum_{k \in c} w_k f_k(x)\right) = \exp\left(\mathbf{w}^\top f(x)\right). \qquad (2)$$

f_k : state indicator functions (assigned one parameter each). For example:

$$f_{k_{\{x_1=1\}}} = \begin{cases} 1 & \text{if } x_1 = 1 \\ 0 & \text{otherwise.} \end{cases} \qquad f_{k_{\{x_1=0,x_2=1\}}} = \begin{cases} 1 & \text{if } x_1 = 0 \text{ and } x_2 = 1 \\ 0 & \text{otherwise.} \end{cases}$$

Introduction

Structure learning problem:
Given N observations of n variables (V), find all relevant edges (E) and estimate their corresponding parameters.

Challenges

- n variables $\Rightarrow$ $\mathrm{O}(\mathrm{n}^2)$ possible edges.
- Learning requires large datasets.

This work

- Investigate major computational bottlenecks of ℓ_1-based learning techniques of Markov Random Fields.
- Propose scalable structure learning approach with controllable trade-off between learning speed and quality.

Outline

ℓ_1-Based Learning

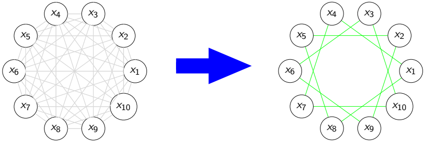
Minimizing ℓ_1-Regularized Negative Log-Likelihood

$$L(\mathbf{w}) = -\frac{1}{N}\sum_{m=1}^{N} \log p_{\mathbf{w}}(x^{(m)}) = -\frac{1}{N}\sum_{m=1}^{N} (\mathbf{w}^\top f(x^{(m)})) + \log Z(\mathbf{w}) \quad (3)$$

$$\mathbb{L}(\mathbf{w}) = L(\mathbf{w}) + \lambda\ ||\mathbf{w}||_1 \quad (4)$$

$$\min_{\mathbf{w}} \mathbb{L}(\mathbf{w}) \quad (5)$$

$$\delta_k L = -\frac{1}{N}\sum_{m=1}^{N} f_k(x^{(m)}) + E_{\mathbf{w}}[f_k(x)] = E_{\mathbf{w}}[f_k(x)] - E_D[f_k(x)] \quad (6)$$



Limitation:

- $E_w[f_k(x)]$: performs inference at each gradient step (Message passing methods are expensive on fully graphs).
- $E_D[f_k(x)]$: requires pre-computing data expectations of each possible state (sufficient statistics).

Feature Grafting[1]

Idea

Assume that all variables are independent and iteratively activate parameters (introduce dependency).

Approach

Active-set method: a working set S and a search set F.

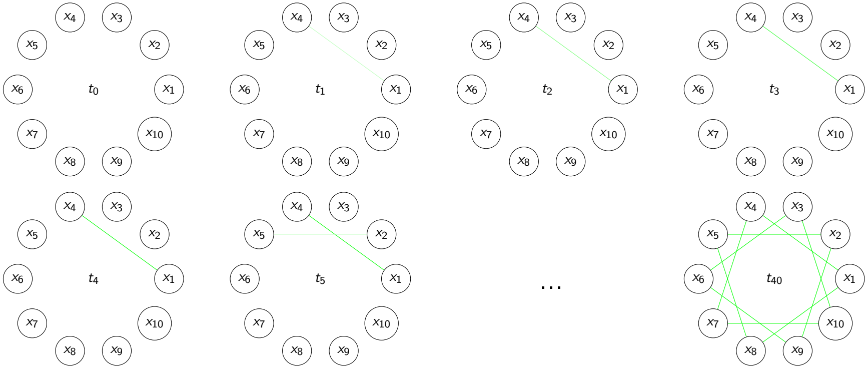
- $S = \{\text{unary parameters}\}$; $F = \{\text{pairwise parameters}\}$.
- Alternate between two steps until convergence:
  - *Step 1*: Optimizing over the active set S using a sub-gradient method.
  - *Step 2*: Select top violating parameter from F and add to S.
- Feature Activation Condition:

$$\text{KKT optimality condition: } \begin{cases} \delta_k L = 0 \text{ if } w_k \neq 0 \\ |\delta_k L| \leq \lambda \text{ if } w_k = 0 \end{cases} \tag{7}$$

$$\Rightarrow \quad C_1 : j = \arg\max_k |\delta_k L| \ s.t. \ |\delta_k L| > \lambda \tag{8}$$

[1] Lee et al, 2007

Feature Grafting



$t_0 : S = \emptyset$
$t_1 : S = \{w_{x_1=0,x_4=1}\}$
$t_2 : S = \{w_{x_1=0,x_4=1}, w_{x_1=1,x_4=1}\}$
$t_3 : S = \{w_{x_1=0,x_4=1}, w_{x_1=1,x_4=1}, w_{x_1=1,x_4=0}\}$
$t_4 : S = \{w_{x_1=0,x_4=1}, w_{x_1=1,x_4=1}, w_{x_1=1,x_4=0}, w_{x_1=0,x_4=0}\}$
$t_5 : S = \{w_{x_1=0,x_4=1}, w_{x_1=1,x_4=1}, w_{x_1=1,x_4=0}, w_{x_1=0,x_4=0}, w_{x_2=1,x_5=0}\}$
$t_{40} : S = S^*$

Feature Grafting

Algorithm 1 Grafting

Initialize $\mathcal{F} = \{\text{set of all pairwise parameters}\}$
Compute sufficient statistics of f $\forall f \in \mathcal{F}$ # cost: $O(n^2 N s_{\max}^2)$
repeat
    Select the top violating feature f^* # cost: $O(n^2 s_{\max}^2)$
    Activate f^*
    Optimize the ℓ_1-regularized L over the active set
until convergence

Limitations:

- Parameters are treated as one homogeneous group. No structure information is used.
- Requires computing $O(n^2 N s_{\max}^2)$ sufficient statistics and performing $O(n^2 s_{\max}^2)$ parameter activation tests.

Outline

Edge Grafting

Problem reformulation: Grafting Edges

- Redefine the search space: $F = \{\text{Edge-wise parameter groups}\}$
- Introduce groups sparsity regularization in the loss function.

$$\mathbb{L}(\mathbf{w}) = L(\mathbf{w}) + \sum_{g \in G} \lambda d_g ||\mathbf{w_g}||_2 + \lambda_2 ||\mathbf{w}||_2^2, \tag{9}$$

where g refers to either a node or an edge and d_g compensates for different groups' cardinalities.

$$\min_{\mathbf{w}} \mathbb{L}(\mathbf{w}) \tag{10}$$

$$\text{KKT optimality condition: } \begin{cases} \frac{||\delta_g L||_2}{d_g} + \lambda_2 ||\mathbf{w_g}||_2^2 = 0 \text{ if } ||\mathbf{w_g}||_2 \neq 0 \\ \frac{||\delta_g L||_2}{d_g} \leq \lambda \text{ if } ||\mathbf{w_g}||_2 = 0 \end{cases} \tag{11}$$

Edge Grafting

Grafting Edges

- Edge score:

$$s_e = \frac{||\delta_e L||_2}{d_e} \tag{12}$$

- Group-wise gradient (pairwise probability error between model and data observations):

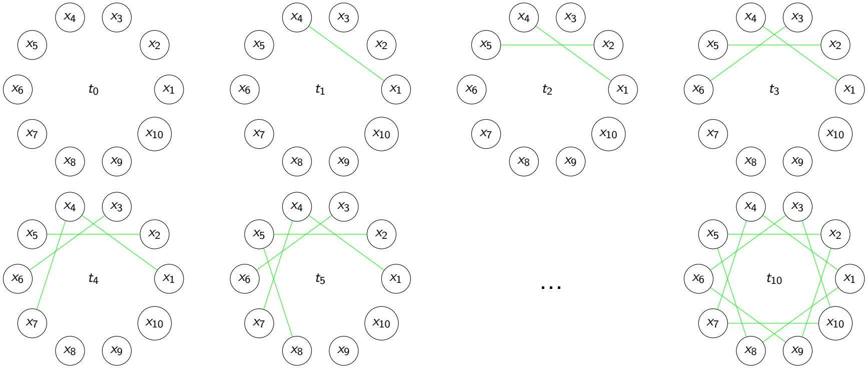
$$\delta_e L = \hat{p}_{\mathbf{w}}(e) - p_D(e) \tag{13}$$

- Necessary edge activation condition:

$$C_2 : \arg\max_{e} |s_e| \; s.t. \; s_e > \lambda \tag{14}$$

Limitations: Requires computing $O(n^2 N s_{\max}^2)$ sufficient statistics and performing $O(n^2)$ edge activation tests.

Edge Grafting



$t_0 : S = \emptyset$
$t_1 : S = \{w_{x_1=0,x_4=1}, w_{x_1=1,x_4=1}, w_{x_1=1,x_4=0}, w_{x_1=0,x_4=0}\}$
$t_2 : S = \{w_{x_1=0,x_4=1}, w_{x_1=1,x_4=1}, w_{x_1=1,x_4=0}, w_{x_1=0,x_4=0}, w_{x_2=0,x_5=1}, w_{x_2=1,x_5=1}, w_{x_2=1,x_5=0}, w_{x_2=0,x_5=0}\}$
$t_{10} : S = S^*$

Best Choice Edge Grafting

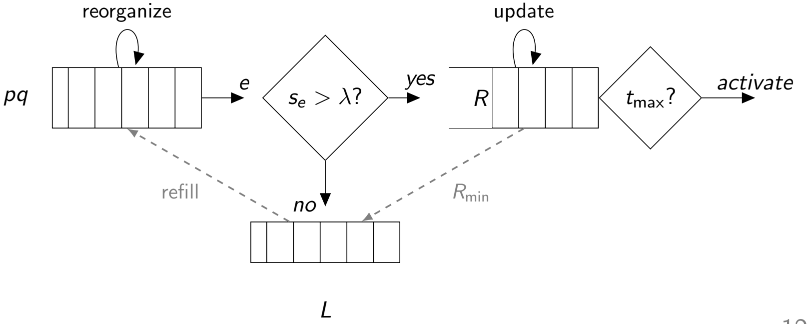
Best Choice Problem

Given a set of streaming candidates, make a decision without testing all possible ones. Similar to a hiring process.

Best Choice Edge Grafting Mechanism

- On-demand edge sufficient statistics computation.
- Reduced number of activation tests

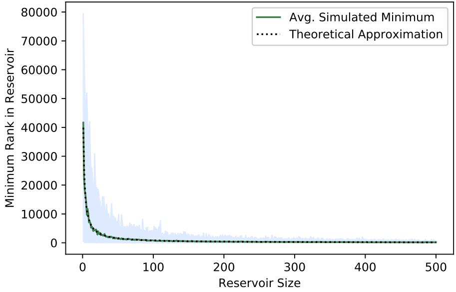
Figure: High-level operational scheme of the edge activation mechanism.



Reservoir Sampling

Benefits of reservoir sampling We simulate the behavior in finite settings, sampling $|R|$ ranks from the list of all possible numbers from 1 to $\binom{n}{2}$ and taking the minimum.

Figure: Simulated edge ranks using the reservoir. (50 nodes).



Two extremes

- **First Hit** ($|R| = 1$) $\rightarrow$ Bad quality edges.
- **Edge Grafting** (using an unlimited reservoir) $\rightarrow$ Negligible gains over a small reservoir.

Reservoir Sampling

Reservoir management

- Before $t_{\max}$ is reached:
  - If reservoir full: replace minimum scoring edge $R_{\min}$ with incoming edge e if $s_{R_{\min}} < s_e$.
- When $t_{\max}$ is reached:
  - Compute mean reservoir scores:

$$\mu = \frac{1}{|R|} \sum_{e \in R} s_e \tag{15}$$

  - Activation threshold as:

$$\tau_\alpha = (1 - \alpha)\mu + \alpha \max_{e \in R} s_e, \tag{16}$$

where $\alpha \in [0, 1]$ controls a trade-off between quality of added edges and speed of edge activation.

Search Space Reorganization

Reorganizing search space

- Search History:
  - Edge violation offset v_e:

$$v_e = 1 - \frac{s_e}{\lambda} \ . \tag{17}$$

  - Store failing edges in L and refill pq when it is empty:

$$pq[e] = v_e \tag{18}$$

- Partial structure information:
  - Idea: Promote a scale-free structure.
  - Detect hubs using degree centrality:

$$c_i = \frac{|\mathcal{N}_i|}{|V| - 1} \tag{19}$$

  - Construct Hub set:

$$H = \{i \in V \text{ such that } c_i > \hat{c}\} \tag{20}$$

  - Prioritizing edges incident to hubs such that $\forall h \in H$ and $\forall n \in V$:

$$pq[(h, n)] = pq[(h, n)] - 1 \tag{21}$$

Summary of Complexities

Algorithm	Suff. stats. at j^{th} edge	Activation step
Feature grafting	$O\left(n^2 N s_{\max}^2\right)$	$O\left(n^2 s_{\max}^2\right)$
Edge grafting	$O\left(n^2 N s_{\max}^2\right)$	$O\left(n^2 s_{\max}^2\right)$
Best choice edge grafting	$O\left((n + j t_{\max}) N s_{\max}^2\right)$	$O\left(t_{\max} s_{\max}^2\right)$

Outline

Synthetic Experiments

Synthetic Data

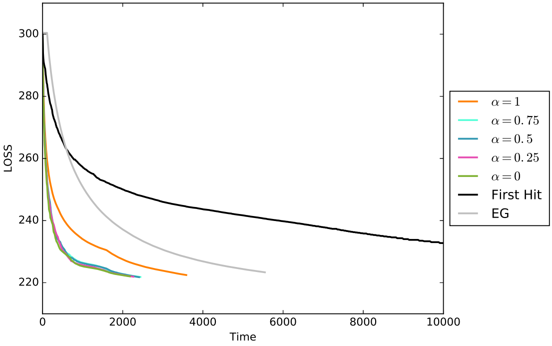
Number of nodes	200	400	600
Number of states per variable	5	5	5
Number of parameters	498, 500	1, 997, 000	4, 495, 500

- Scale-free-structures: Few dominant hubs.
- Data generated using Gibbs sampler: 20, 000 data points from each network, randomly split into train and held-out testing sets.

Synthetic Experiments

Synthetic results

Figure: Full convergence of different methods (200 nodes).

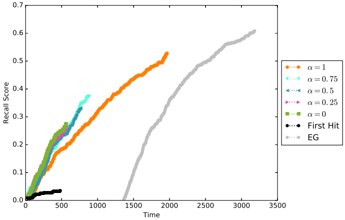


$$\tau_\alpha = (1-\alpha)\mu + \alpha \max_{e \in R} s_e$$

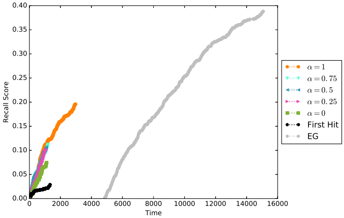
Synthetic Experiments

Synthetic results

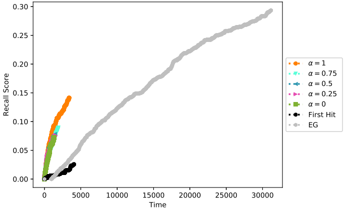
Figure: Learning objectives vs time for varying MRFs sizes.



(a) 200 nodes and 600 edges



(b) 400 nodes and 1,200 edges

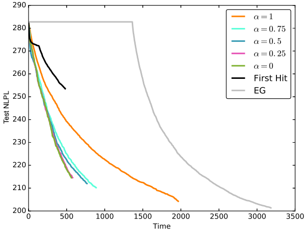


(c) 600 nodes and 1,800 edges

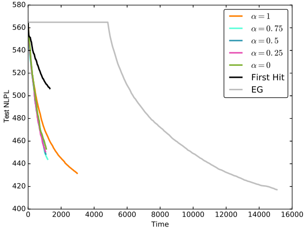
Synthetic Experiments

Synthetic results

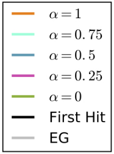
Figure: Negative Log Pseudo-Likelihood vs time for varying MRFs sizes.

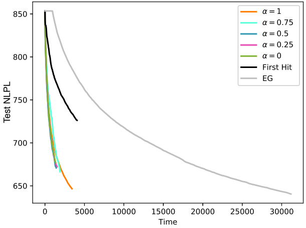


(a) 200 nodes and 600 edges



(b) 400 nodes and 1,200 edges



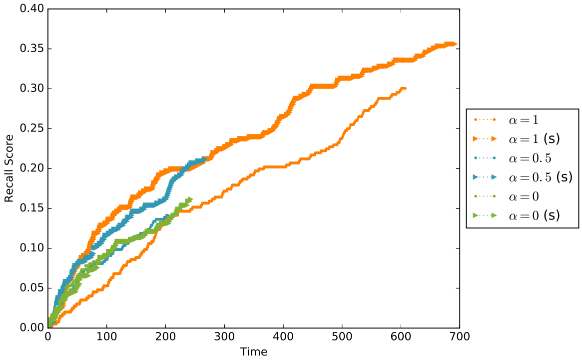


(c) 600 nodes and 1,800 edges

Synthetic Experiments

Synthetic results

Figure: Role of structure heuristics in improving the quality of the learned MRF.(200 nodes)



Real Data Experiments

Real data

Dataset	Jester	Yummly recipes
Number of variables	100	153
Number of States per variable	5	2
Number of parameters	124,250	36,450
Dataset size	73,421	10,000

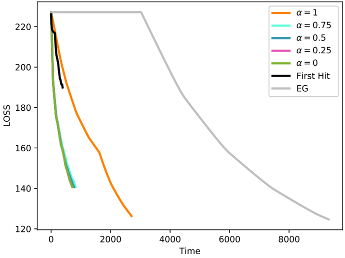
- Jester[2]: user ratings of jokes.
- Yummly recipes[3]: recipes with different ingredients.

[2] http://goldberg.berkeley.edu/jester-data/
[3] https://www.kaggle.com/c/whats-cooking

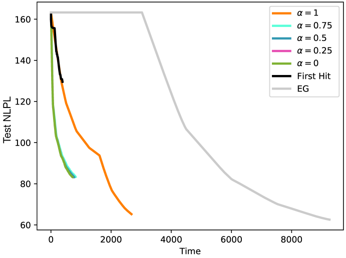
Real Data Experiments

Real data results

Figure: Negative Log Pseudo-Likelihood vs time for varying MRFs sizes.



(a) Yummly Objective



(b) Yummly NLPL

Outline

Conclusion

Proposed work

- Reformulate learning problem by introducing structure information.
- Avoid costly batch ℓ_1-learning on the entire problem space. Informed edge search through reservoir sampling and search space reorganization.

Result

- Faster edge activation and convergence.
- Controllable trade-off between learning speed and quality.
- Achieved better scalability.

Limitations and future work

- Assumption of scale free structure: Investigate better structure heuristics for a more efficient search space reorganization.
- Applied on pairwise MRFs: Generalize approach for higher order MRFs.

Contact us: walidch@vt.edu; bhuang@vt.edu