CS 446/ECE 449: Machine Learning

A. G. Schwing

University of Illinois at Urbana-Champaign, 2020

L13: Structured Prediction (ILP, LP relaxation, message passing, graph cut)

Goals of this lecture

- Getting to know structured inference algorithms

Reading material:

- D. Koller and N. Friedman; Probabilistic Graphical Models: Principles and Techniques;

Recap: Inference Program

$$\boldsymbol{y}^* = \arg\max_{\hat{\boldsymbol{y}}} \sum_r f_r(\hat{\boldsymbol{y}}_r)$$

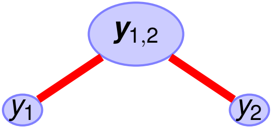
Algorithms:

- Exhaustive search
- Dynamic programming
- Integer linear program
- Linear programming relaxation
- Message passing
- Graph-cut

Integer Linear Program

Example:

$$\mathbf{y}^* = \arg\max_{\hat{\mathbf{y}}} \sum_r f_r(\hat{\mathbf{y}}_r)$$



Integer Linear Program (LP) equivalence: variables $b_r(\mathbf{y}_r)$

$$\max_{b_1,b_2,b_{12}} \begin{bmatrix} b_1(1) \\ b_1(2) \\ b_2(1) \\ b_2(2) \\ b_{12}(1,1) \\ b_{12}(2,1) \\ b_{12}(1,2) \\ b_{12}(2,2) \end{bmatrix}^\top \begin{bmatrix} f_1(1) \\ f_1(2) \\ f_2(1) \\ f_2(2) \\ f_{12}(1,1) \\ f_{12}(2,1) \\ f_{12}(1,2) \\ f_{12}(2,2) \end{bmatrix} \quad \text{s.t.} \quad \begin{array}{lr} b_r(\mathbf{y}_r) \in \{0,1\} & \forall r, \mathbf{y}_r \\ \sum_{\mathbf{y}_r} b_r(\mathbf{y}_r) = 1 & \forall r \\ \sum_{\mathbf{y}_p \setminus \mathbf{y}_r} b_p(\mathbf{y}_p) = b_r(\mathbf{y}_r) & \forall r, \mathbf{y}_r, p \in P(r) \end{array}$$

Example:

$$\max_{b_1,b_2,b_{12}} \begin{bmatrix} b_1(1) \\ b_1(2) \\ b_2(1) \\ b_2(2) \\ b_{12}(1,1) \\ b_{12}(2,1) \\ b_{12}(1,2) \\ b_{12}(2,2) \end{bmatrix}^\top \begin{bmatrix} f_1(1) \\ f_1(2) \\ f_2(1) \\ f_2(2) \\ f_{12}(1,1) \\ f_{12}(2,1) \\ f_{12}(1,2) \\ f_{12}(2,2) \end{bmatrix} \quad \text{s.t.} \quad \begin{array}{lr} b_r(\boldsymbol{y}_r) \in \{0,1\} & \forall r, \boldsymbol{y}_r \\ \sum_{\boldsymbol{y}_r} b_r(\boldsymbol{y}_r) = 1 & \forall r \\ \sum_{\boldsymbol{y}_p \setminus \boldsymbol{y}_r} b_p(\boldsymbol{y}_p) = b_r(\boldsymbol{y}_r) & \forall r, \boldsymbol{y}_r, p \in P(r) \end{array}$$

Last constraint explicitly:

$$\begin{aligned} b_{12}(1,1) + b_{12}(1,2) &= b_1(1) \\ b_{12}(2,1) + b_{12}(2,2) &= b_1(2) \\ b_{12}(1,1) + b_{12}(2,1) &= b_2(1) \\ b_{12}(1,2) + b_{12}(2,2) &= b_2(2) \end{aligned}$$

Linear Programming Relaxation

$$\boldsymbol{y}^* = \arg\max_{\hat{\boldsymbol{y}}} \sum_r f_r(\hat{\boldsymbol{y}}_r)$$

LP relaxation:

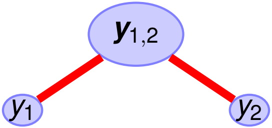
$$\underbrace{\cancel{b_r(\boldsymbol{y}_r) \in \{0,1\} \quad \forall r, \boldsymbol{y}_r}}$$

$$\max_{b_r} \quad \sum_{r,\boldsymbol{y}_r} b_r(\boldsymbol{y}_r) f_r(\boldsymbol{y}_r) \quad \underbrace{\text{s.t.} \quad \begin{array}{l} \text{Local probability } b_r \\ \text{Marginalization} \end{array}}_{\text{s.t.} \quad b \in \mathcal{C}}$$

- **Advantage:** global optimum for LP, very good solvers available
- **Disadvantage:** no global optimum for ILP, slow for larger problems

Message Passing ([Loopy] Belief Propagation)

Exploit: Graph structure defined via marginalization constraints



How: Compute the dual function

Message passing solvers:

- **Advantage:** Efficient due to analytically computable sub-problems
- **Disadvantage:** Special care required to find LP relaxation optimum

Computing the dual of

$$\max_b \sum_{r,\boldsymbol{y}_r} b_r(\boldsymbol{y}_r) f_r(\boldsymbol{y}_r) \quad \text{s.t.} \quad \begin{cases} b_r(\boldsymbol{y}_r) \geq 0 & \forall r, \boldsymbol{y}_r \\ \sum_{\boldsymbol{y}_r} b_r(\boldsymbol{y}_r) = 1 & \forall r \\ \sum_{\boldsymbol{y}_p \setminus \boldsymbol{y}_r} b_p(\boldsymbol{y}_p) = b_r(\boldsymbol{y}_r) & \forall r, p \in P(r), \boldsymbol{y}_r \end{cases}$$

Lagrangian:

$$\begin{aligned} L() &= \sum_{r,\boldsymbol{y}_r} b_r(\boldsymbol{y}_r) f_r(\boldsymbol{y}_r) + \sum_{r,p \in P(r),\boldsymbol{y}_r} \lambda_{r \to p}(\boldsymbol{y}_r) \left(\sum_{\boldsymbol{y}_p \setminus \boldsymbol{y}_r} b_p(\boldsymbol{y}_p) - b_r(\boldsymbol{y}_r) \right) \\ &= \sum_{r,\boldsymbol{y}_r} b_r(\boldsymbol{y}_r) \left(f_r(\boldsymbol{y}_r) - \sum_{p \in P(r)} \lambda_{r \to p}(\boldsymbol{y}_r) + \sum_{c \in C(r)} \lambda_{c \to r}(\boldsymbol{y}_c) \right) \end{aligned}$$

Intermediate result:

$$
\begin{aligned}
\sum_{r,p\in P(r),\boldsymbol{y}_r} \lambda_{r\to p}(\boldsymbol{y}_r) \sum_{\boldsymbol{y}_p\setminus\boldsymbol{y}_r} b_p(\boldsymbol{y}_p) &= \sum_{r,p\in P(r),\boldsymbol{y}_p} \lambda_{r\to p}(\boldsymbol{y}_r) b_p(\boldsymbol{y}_p) \\
&= \sum_{p,r\in C(p),\boldsymbol{y}_p} \lambda_{r\to p}(\boldsymbol{y}_r) b_p(\boldsymbol{y}_p) \\
&= \sum_{r,c\in C(r),\boldsymbol{y}_r} \lambda_{c\to r}(\boldsymbol{y}_c) b_r(\boldsymbol{y}_r) \\
&= \sum_{r,\boldsymbol{y}_r,c\in C(r)} \lambda_{c\to r}(\boldsymbol{y}_c) b_r(\boldsymbol{y}_r)
\end{aligned}
$$

Lagrangian:

$$L() = \sum_{r,\boldsymbol{y}_r} b_r(\boldsymbol{y}_r)\left(f_r(\boldsymbol{y}_r) - \sum_{p\in P(r)} \lambda_{r\to p}(\boldsymbol{y}_r) + \sum_{c\in C(r)} \lambda_{c\to r}(\boldsymbol{y}_c)\right)$$

Maximize Lagrangian w.r.t. primal variables subject to remaining constraints:

$$\max_b L() \quad \text{s.t.} \quad \begin{cases} b_r(\boldsymbol{y}_r) \geq 0 & \forall r, \boldsymbol{y}_r \\ \sum_{\boldsymbol{y}_r} b_r(\boldsymbol{y}_r) = 1 & \forall r \end{cases}$$

Dual function:

$$g(\lambda) = \sum_r \max_{\boldsymbol{y}_r}\left(f_r(\boldsymbol{y}_r) - \sum_{p\in P(r)} \lambda_{r\to p}(\boldsymbol{y}_r) + \sum_{c\in C(r)} \lambda_{c\to r}(\boldsymbol{y}_c)\right)$$

Dual function:

$$g(\lambda) = \sum_r \max_{\boldsymbol{y}_r} \left(f_r(\boldsymbol{y}_r) - \sum_{p \in P(r)} \lambda_{r \to p}(\boldsymbol{y}_r) + \sum_{c \in C(r)} \lambda_{c \to r}(\boldsymbol{y}_c) \right)$$

Dual program:

$$\min_\lambda g(\lambda)$$

Original primal:

$$\max_b \sum_{r, \boldsymbol{y}_r} b_r(\boldsymbol{y}_r) f_r(\boldsymbol{y}_r) \quad \text{s.t.} \quad \left\{ \begin{array}{ll} b_r(\boldsymbol{y}_r) \geq 0 & \forall r, \boldsymbol{y}_r \\ \sum_{\boldsymbol{y}_r} b_r(\boldsymbol{y}_r) = 1 & \forall r \\ \sum_{\boldsymbol{y}_p \setminus \boldsymbol{y}_r} b_p(\boldsymbol{y}_p) = b_r(\boldsymbol{y}_r) & \forall r, p \in P(r), \boldsymbol{y}_r \end{array} \right.$$

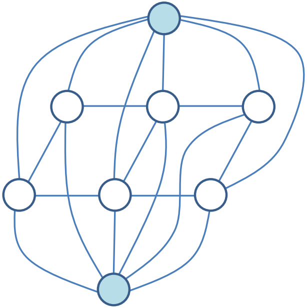
Properties of dual program:

- Convex program
- Not strongly convex
- Piecewise linear
- **Unconstrained**
- Lagrange multipliers are messages

Lagrange multipliers are messages defined on edges of the graph. They shift 'energy' such that local maximization (dual) is identical to global maximization (primal).

Graph-cut Solvers

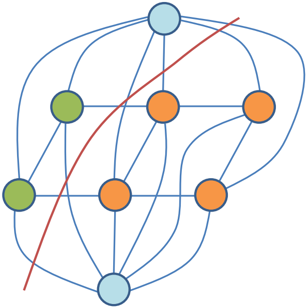
- Efficient algorithms to compute the minimum cost cut in a weighted graph
- Efficient algorithms to compute the maximum flow through a weighted graph



Graph-cut Solvers

For binary problems $y_d \in \{1, 2\}$:

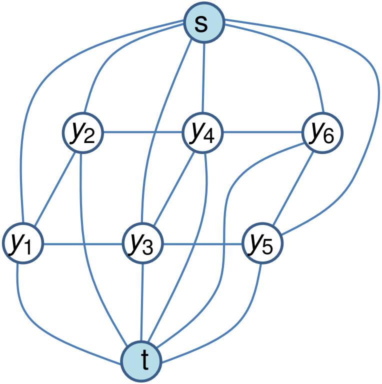
- Convert scoring function F into auxiliary graph (not the same graph as before!)
- Compute a weighted cut cost corresponding to the labeling score



What are the nodes and what are the weights on the edges in this auxiliary graph?

Graph-cut Solvers

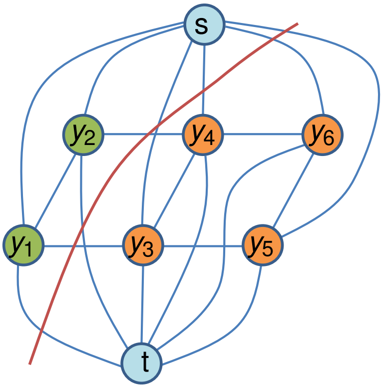
What are the nodes in the auxiliary graph?



- Two special nodes called 'source' and 'terminal'
- Variables y_d as nodes

Graph-cut Solvers

What are the nodes in the auxiliary graph?



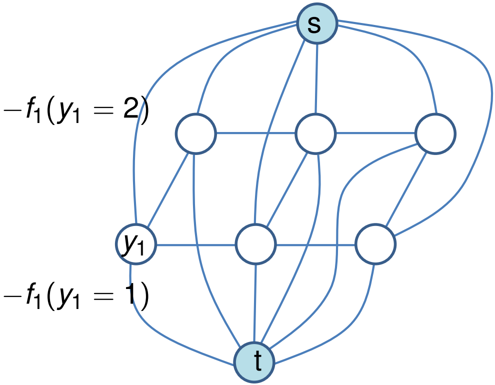
- Two special nodes called ‘source’ and ‘terminal’
- Variables y_d as nodes

Graph-cut Solvers

What weights do we assign to edges?
Recall that local scoring functions are arrays:

$$\left[\begin{array}{cc} f_1(y_1 = 1) & f_1(y_1 = 2) \end{array}\right]$$

Graph-cut solvers compute a min-cut:

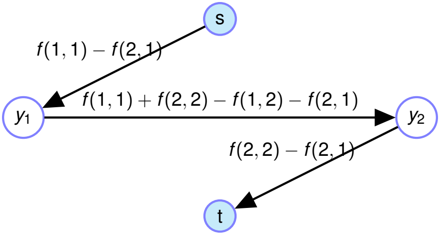


Graph-cut Solvers

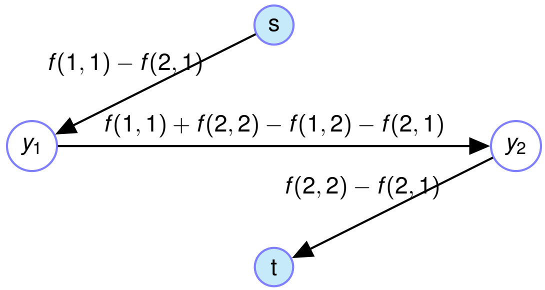
What weights do we assign to edges?
Recall that local scoring functions are arrays:

$$\begin{aligned}
\begin{bmatrix} f_{12}(1,1) & f_{12}(1,2) \\ f_{12}(2,1) & f_{12}(2,2) \end{bmatrix} &= f(1,1) - f(2,1) + f(2,2) \\
&+ \begin{bmatrix} 0 & 0 \\ f(2,1) - f(1,1) & f(2,1) - f(1,1) \end{bmatrix} \\
&+ \begin{bmatrix} f(2,1) - f(2,2) & 0 \\ f(2,1) - f(2,2) & 0 \end{bmatrix} \\
&+ \begin{bmatrix} 0 & f(1,2) + f(2,1) - f(1,1) - f(2,2) \\ 0 & 0 \end{bmatrix}
\end{aligned}$$

Graph-cut solvers compute a min-cut:



Graph-cut Solvers



Requirement for optimality: Pairwise edge weights are positive

$$f(1,1) + f(2,2) - f(1,2) - f(2,1) \geq 0 \quad \text{sub-modularity}$$

For higher order functions? More complicated graph constructions
For more than two labels? Move making algorithms

Structured Prediction

Inference:

$$\boldsymbol{y}^* = \arg\max_{\hat{\boldsymbol{y}}} \sum_r f_r(\boldsymbol{w}, x, \hat{\boldsymbol{y}}_r)$$

Efficiency and accuracy of inference algorithms is problem dependent:

- Exhaustive search
- Dynamic programming
- Integer linear program
- Linear programming relaxation
- Message passing
- Graph-cut

Quiz:

- What is the advantage of ILP and LP relaxation compared to dynamic programming and exhaustive search?
- When is a graph-cut algorithm optimal?

Important topics of this lecture

- More inference algorithms for structured spaces

Up next:

- Learning models for structured output spaces