Max-Margin Markov Networks

by Ben Taskar, Carlos Guestrin and Daphne Koller

Moontae Lee and Ozan Sener

Cornell University

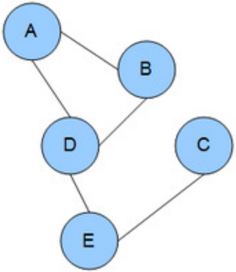
February 11, 2014

Overview

- Quiz
- Introduction to Markov Network
- Pairwise Log-linear Model
- Margin-based Formulation
- Exploiting Network Structure
- Polytope Constraints
- Coordinate-wise Optimization
- Training Methods
- Summary and Further Readings

Markov Random Field

- Temporal/Spatial relations need to be modelled by most of the ML systems
- Markov Random Field (MRF) is a way to model such structures.



Markov Random Field

Given a graph $G(V, E)$, a set of variables $(X_v)_{v \in V}$ is a MRF if a variable is conditionally independent of all other variables given its neighbors. $ex. P(X_E | X_A, X_B, X_C, X_D) = P(X_E | X_C, X_D)$

How to do Inference - $\arg\max P(\{X_v\}_{v \in V})$

- If it is a Markov Chain, we can use Viterbi algorithm.
- What if it is not ?

Hammersley & Clifford theorem

If MRF has positive measure, its probability density can be decomposed over set of cliques.

- $P(X_A, X_B, X_C, X_D, X_E) = e^{-E(X_A, X_B, X_C, X_D, X_E)}$ where, $E(X_{A:E}) = E(X_A, X_B, X_D) + E(X_D, X_E) + E(X_C, X_E)$

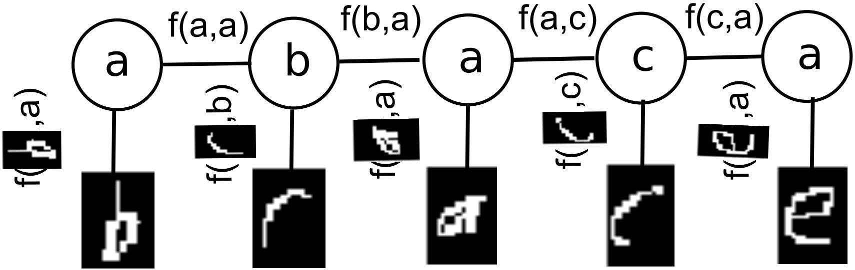
Pairwise Log-linear Model

- Assume pairwise MRF (any two non-adjacent variables are conditionally independent given all other variables)
- Energy function is defined over edges $E(X) = \sum_{(u,v)\in\mathcal{E}} E(X_u, X_v)$
- If we use indicator functions, resultant energy is linear. Consider two nodes (x_1, x_2) Markov network;

$$
\begin{aligned}
f_1(x) &= 1 \quad \textit{if} \quad x_1 = 0, x_2 = 0 & w_1 &= E(x_1 = 0, x_2 = 0) \\
f_2(x) &= 1 \quad \textit{if} \quad x_1 = 0, x_2 = 1 & w_2 &= E(x_1 = 0, x_2 = 1) \\
f_3(x) &= 1 \quad \textit{if} \quad x_1 = 1, x_2 = 0 & w_3 &= E(x_1 = 1, x_2 = 0) \\
f_4(x) &= 1 \quad \textit{if} \quad x_1 = 1, x_2 = 1 & w_4 &= E(x_1 = 1, x_2 = 1)
\end{aligned}
$$

$$E(x_1, x_2) = \sum_{i=1}^{4} f_i w_i = f(x_1, x_2)^T w$$

Problem to be Solved



- Energy function is log-likelihood ($E = w^T f$) where f is the concatenation of all edge features.

f = (f(a,b) f(b,a) f(a,c) f(c,a) , f( ,a), f( ,b), f( ,a), f( ,c), f( ,a))

- And we solve the energy minimization problem which corresponds to ML problem.

$$y = \arg\max w^T f(\;\cdot\;, y)$$

Margin-based Formulation

- We want to learn a weight vector w such that

$\arg\max w^T f($  $, y) = \text{"}brace\text{"}$

$w^T f($  $, \text{“}brace\text{”}) > w^T f($  $, \text{“}aaaaa\text{”})$

$$\vdots$$

$w^T f($  $, \text{“}brace\text{”}) > w^T f($  $\text{“}zzzzz\text{”})$

- Our goal is to maximize the margin constraining $\|w\| \leq 1$

$\max \lambda \quad s.t \quad w^T f($  $, \text{“}brace\text{”}) - w^T f($  $, y) \geq \lambda \quad \forall_y$

Max-Margin Markov Network (MMMN)

- Primal Formulation:

$$\min \ \frac{1}{2}\|w\|^2 + C\sum_x \xi_x \quad s.t \quad w^T\Delta f_x(y) \geq \Delta t_x(y) - \xi_x \ \ \forall_{x,y}$$

$$where \ \ \Delta f_x(y) = f(x, t(x)) - f(x, y), \ \ \Delta t_x(y) = loss\ against\ the\ true\ label\ t(x)$$

- Dual Formulation:

$$\max \ \sum_{x,y} \alpha_x(y)\Delta t_x(y) - \frac{1}{2}\left\|\sum_{x,y} \alpha_x(y)\Delta f_x(y)\right\|^2$$

$$s.t \quad \sum_y \alpha_x(y) = C \ \ \forall_x \qquad \alpha_x(y) \geq 0 \ \ \forall_{x,y}$$

Max-Margin Markov Network (MMMN)

- Primal Formulation:

$$\min \ \frac{1}{2}\|w\|^2 + C\sum_x \xi_x \quad s.t \quad w^T \Delta f_x(y) \geq \Delta t_x(y) - \xi_x \ \ \forall_{x,y}$$

$$where \ \ \Delta f_x(y) = f(x, t(x)) - f(x, y), \ \ \Delta t_x(y) = loss\ against\ the\ true\ label\ t(x)$$

- Dual Formulation:

$$\max \ \sum_{x,y} \alpha_x(y)\Delta t_x(y) - \frac{1}{2}\left\| \sum_{x,y} \alpha_x(y)\Delta f_x(y) \right\|^2$$

$$s.t \quad \sum_y \alpha_x(y) = C \ \ \forall_x \qquad \alpha_x(y) \geq 0 \ \ \forall_{x,y}$$

Q1. # of dual variables? (m examples, l binary outputs)

Max-Margin Markov Network (MMMN)

- Primal Formulation:

$$\min \quad \frac{1}{2}\|w\|^2 + C\sum_x \xi_x \quad s.t \quad w^T \Delta f_x(y) \geq \Delta t_x(y) - \xi_x \ \ \forall_{x,y}$$

$$where \ \ \Delta f_x(y) = f(x, t(x)) - f(x, y), \ \ \Delta t_x(y) = loss\ against\ the\ true\ label\ t(x)$$

- Dual Formulation:

$$\max \quad \sum_{x,y} \alpha_x(y)\Delta t_x(y) - \frac{1}{2}\left\|\sum_{x,y} \alpha_x(y)\Delta f_x(y)\right\|^2$$

$$s.t \quad \sum_y \alpha_x(y) = C \ \ \forall_x \qquad \alpha_x(y) \geq 0 \ \ \forall_{x,y}$$

Q1. # of dual variables? (*m* examples, *l* binary outputs)

Q2. # of addends?

Max-Margin Markov Network (MMMN)

- Primal Formulation:

$$\min \quad \frac{1}{2}\|w\|^2 + C\sum_x \xi_x \quad s.t \quad w^T \Delta f_x(y) \geq \Delta t_x(y) - \xi_x \;\; \forall_{x,y}$$

$$where \;\; \Delta f_x(y) = f(x, t(x)) - f(x, y), \;\; \Delta t_x(y) = loss\ against\ the\ true\ label\ t(x)$$

- Dual Formulation:

$$\max \quad \sum_{x,y} \alpha_x(y)\Delta t_x(y) - \frac{1}{2}\left\|\sum_{x,y} \alpha_x(y)\Delta f_x(y)\right\|^2$$

$$s.t \quad \sum_y \alpha_x(y) = C \;\; \forall_x \qquad \alpha_x(y) \geq 0 \;\; \forall_{x,y}$$

Q1. # of dual variables? (m examples, l binary outputs)

Q2. # of addends $= m \cdot 2^l + m \cdot 2^l$

Max-Margin Markov Network (MMMN)

- Primal Formulation:

$$\min \quad \frac{1}{2}\|w\|^2 + C\sum_x \xi_x \quad s.t \quad w^T \Delta f_x(y) \geq \Delta t_x(y) - \xi_x \;\; \forall_{x,y}$$

$$where \;\; \Delta f_x(y) = f(x, t(x)) - f(x, y), \;\; \Delta t_x(y) = loss\ against\ the\ true\ label\ t(x)$$

- Dual Formulation:

$$\max \; \sum_{x,y} \alpha_x(y)\Delta t_x(y) - \frac{1}{2}\left\|\sum_{x,y} \alpha_x(y)\Delta f_x(y)\right\|^2$$

$$s.t \quad \sum_y \alpha_x(y) = C \;\; \forall_x \qquad \alpha_x(y) \geq 0 \;\; \forall_{x,y}$$

Q1. # of dual variables? (m examples, l binary outputs)

Q2. # of addends $= m \cdot 2^l + m \cdot 2^l$

Q3. Is it equivalent to Structural SVM (SSVM)?

Exploiting Structure in MMMN (1)

(Observation 1) Dual variables $\{\alpha_x(y)\}_{x,y}$ satisfy

$$\sum_y \alpha_x(y) = C \ \ and \ \ \alpha_x(y) \geq 0 \ \ \forall_y$$

So, $\alpha_x(y)$ can be an unnormalized density function over y given x

Exploiting Structure in MMMN (1)

(Observation 1) Dual variables $\{\alpha_x(y)\}_{x,y}$ satisfy

$$\sum_y \alpha_x(y) = C \ \ and \ \ \alpha_x(y) \geq 0 \ \ \forall_y$$

So, $\alpha_x(y)$ can be an unnormalized density function over y given x

(Observation 2) Both are decomposed into

$$\Delta t_x(y) = loss\ against\ t(x) = \#\ of\ disagreements = \sum_{i \in V} I[\, y_i \neq (t(x))_i \,] = \sum_{i \in V} \Delta t_x(y_i)$$

$$\Delta f_x(y) = f(x, t(x)) - f(x, y) = \sum_{(i,j) \in E} \left(f(x, t(x)_i, t(x)_j) - f(x, y_i, y_j) \right) = \sum_{(i,j) \in E} \Delta f_x(y_i, y_j)$$

The decompositions are sums over edges and nodes coherent to our network structure $G = (V, E)$!

Exploiting Structure in MMMN (2)

- Define new dual variables via marginalizations $\{\alpha_x(y)\}_{x,y}$

$$\mu_x(y_i) = \sum_{y\sim[y_i]} \alpha_x(y) \quad \forall i \in V, \;\; \forall y, \;\; \forall x$$
$$\mu_x(y_i, y_j) = \sum_{y\sim[y_i, y_j]} \alpha_x(y) \quad \forall (i,j) \in E, \;\; \forall y_i, y_j, \;\; \forall x$$

- Then the 1st term has a new representation such that

$$\sum_y \alpha_x(y)\Delta t_x(y) = \sum_y \alpha_x(y)\left(\sum_{i\in V}\Delta t_x(y_i)\right) = \sum_y \sum_{i\in V} \alpha_x(y)\Delta t_x(y_i)$$
$$= \sum_{i\in V}\left(\sum_{y_i}\Delta t_x(y_i)\sum_{y\sim[y_i]}\alpha_x(y)\right) = \sum_{i\in V}\sum_{y_i}\mu_x(y_i)\Delta t_x(y_i)$$

Exploiting Structure in MMMN (3)

- (Example) Given a sample x, see the following transformation:

	y_1	y_2	y_3	$\Delta t_x(y_1)$	$\Delta t_x(y_2)$	$\Delta t_x(y_3)$	$\Delta t_x(y)$	$\alpha_x(y)$
$t(x)$	1	0	1	true label				
all possible labels y	0	0	0	1	0	1	2	0.1
	0	0	1	1	0	0	1	0.2
	0	1	0	1	1	1	3	0.1
	0	1	1	1	1	0	2	0.1
	1	0	0	0	0	1	1	0.1
	1	0	1	0	0	0	0	0.1
	1	1	0	0	1	1	2	0.2
	1	1	1	0	1	0	1	0.1
$\mu_x(y_i=0)$	0.5	0.5	0.5	0.5*1	0.5*0	0.5*1	$\Sigma = 1.5$	
$\mu_x(y_i=1)$	0.5	0.5	0.5	0.5*0	0.5*1	0.5*0		

$$\sum_y \alpha_x(y)\Delta t_x(y) = \text{sum of 8 terms} = 1.5 \quad (\because y \in \{0,1\}^3)$$

$$\sum_i \sum_{y_i} \mu_x(y_i)\Delta t_x(y_i) = \text{sum of 6 terms} = 1.5 \quad (\because i \in \{1,2,3\} \;\; y_i \in \{0,1\})$$

Exploiting Structure in MMMN (4)

- Similarly the 2nd term has a new representation such that

$$\| \sum_{x,y} \alpha_x(y) \Delta f_x(y) \|^2 = \sum_{x,x'} \sum_{(i,j)\in E} \sum_{(i',j')\in E} \sum_{y_i,y_j} \sum_{y_{i'},y_{j'}} \mu_x(y_i,y_j)\mu_{x'}(y_{i'},y_{j'})\Delta f_x(y_i,y_j)^T \Delta f_{x'}(y_{i'},y_{j'})$$

- Therefore the new equivalent formulation is to maximize

$$\sum_x \sum_{i\in V} \sum_{y_i} \mu_x(y_i)\Delta t_x(y_i) - \frac{1}{2} \sum_{x,x'} \sum_{(i,j)\in E} \sum_{(i',j')\in E} \sum_{y_i,y_j} \sum_{y_{i'},y_{j'}} \mu_x(y_i,y_j)\mu_{x'}(y_{i'},y_{j'})\Delta f_x(y_i,y_j)^T \Delta f_{x'}(y_{i'},y_{j'})$$

Exploiting Structure in MMMN (4)

- Similarly the 2nd term has a new representation such that

$$\|\sum_{x,y}\alpha_x(y)\Delta f_x(y)\|^2 = \sum_{x,x'}\sum_{(i,j)\in E}\sum_{(i',j')\in E}\sum_{y_i,y_j}\sum_{y_{i'},y_{j'}}\mu_x(y_i,y_j)\mu_{x'}(y_{i'},y_{j'})\Delta f_x(y_i,y_j)^T\Delta f_{x'}(y_{i'},y_{j'})$$

- Therefore the new equivalent formulation is to maximize

$$\sum_{x}\sum_{i\in V}\sum_{y_i}\mu_x(y_i)\Delta t_x(y_i) - \frac{1}{2}\sum_{x,x'}\sum_{(i,j)\in E}\sum_{(i',j')\in E}\sum_{y_i,y_j}\sum_{y_{i'},y_{j'}}\mu_x(y_i,y_j)\mu_{x'}(y_{i'},y_{j'})\Delta f_x(y_i,y_j)^T\Delta f_{x'}(y_{i'},y_{j'})$$

Q1. # of dual variables? (m examples, l binary outputs)

Exploiting Structure in MMMN (4)

- Similarly the 2nd term has a new representation such that

$$\|\sum_{x,y}\alpha_x(y)\Delta f_x(y)\|^2 = \sum_{x,x'}\sum_{(i,j)\in E}\sum_{(i',j')\in E}\sum_{y_i,y_j}\sum_{y_{i'},y_{j'}}\mu_x(y_i,y_j)\mu_{x'}(y_{i'},y_{j'})\Delta f_x(y_i,y_j)^T\Delta f_{x'}(y_{i'},y_{j'})$$

- Therefore the new equivalent formulation is to maximize

$$\sum_x\sum_{i\in V}\sum_{y_i}\mu_x(y_i)\Delta t_x(y_i) - \frac{1}{2}\sum_{x,x'}\sum_{(i,j)\in E}\sum_{(i',j')\in E}\sum_{y_i,y_j}\sum_{y_{i'},y_{j'}}\mu_x(y_i,y_j)\mu_{x'}(y_{i'},y_{j'})\Delta f_x(y_i,y_j)^T\Delta f_{x'}(y_{i'},y_{j'})$$

Q1. # of dual variables? (m examples, l binary outputs)

$$|\{\mu_x(y_i)\}_{x,y_i}| = ml \quad |\{\mu_x(y_i,y_j)\}_{x,y_i,y_j}| = ml^2 \Rightarrow ml(1+l)$$

Exploiting Structure in MMMN (4)

- Similarly the 2nd term has a new representation such that

$$\|\sum_{x,y}\alpha_x(y)\Delta f_x(y)\|^2 = \sum_{x,x'}\sum_{(i,j)\in E}\sum_{(i',j')\in E}\sum_{y_i,y_j}\sum_{y_{i'},y_{j'}}\mu_x(y_i,y_j)\mu_{x'}(y_{i'},y_{j'})\Delta f_x(y_i,y_j)^T\Delta f_{x'}(y_{i'},y_{j'})$$

- Therefore the new equivalent formulation is to maximize

$$\sum_x\sum_{i\in V}\sum_{y_i}\mu_x(y_i)\Delta t_x(y_i) - \frac{1}{2}\sum_{x,x'}\sum_{(i,j)\in E}\sum_{(i',j')\in E}\sum_{y_i,y_j}\sum_{y_{i'},y_{j'}}\mu_x(y_i,y_j)\mu_{x'}(y_{i'},y_{j'})\Delta f_x(y_i,y_j)^T\Delta f_{x'}(y_{i'},y_{j'})$$

Q1. # of dual variables? (m examples, l binary outputs)

$$|\{\mu_x(y_i)\}_{x,y_i}| = ml \quad |\{\mu_x(y_i,y_j)\}_{x,y_i,y_j}| = ml^2 \Rightarrow ml(1+l)$$

Q2. # of addends?

Exploiting Structure in MMMN (4)

- Similarly the 2nd term has a new representation such that

$$\|\sum_{x,y}\alpha_x(y)\Delta f_x(y)\|^2 = \sum_{x,x'}\sum_{(i,j)\in E}\sum_{(i',j')\in E}\sum_{y_i,y_j}\sum_{y_{i'},y_{j'}}\mu_x(y_i,y_j)\mu_{x'}(y_{i'},y_{j'})\Delta f_x(y_i,y_j)^T\Delta f_{x'}(y_{i'},y_{j'})$$

- Therefore the new equivalent formulation is to maximize

$$\sum_x\sum_{i\in V}\sum_{y_i}\mu_x(y_i)\Delta t_x(y_i) - \frac{1}{2}\sum_{x,x'}\sum_{(i,j)\in E}\sum_{(i',j')\in E}\sum_{y_i,y_j}\sum_{y_{i'},y_{j'}}\mu_x(y_i,y_j)\mu_{x'}(y_{i'},y_{j'})\Delta f_x(y_i,y_j)^T\Delta f_{x'}(y_{i'},y_{j'})$$

Q1. # of dual variables? (m examples, l binary outputs)

$$|\{\mu_x(y_i)\}_{x,y_i}| = ml \quad |\{\mu_x(y_i,y_j)\}_{x,y_i,y_j}| = ml^2 \Rightarrow ml(1+l)$$

Q2. # of addends $= ml \cdot 2 + m^2 \cdot {}_lC_2{}^2 \cdot 2^4$

Exploiting Structure in MMMN (4)

- Similarly the 2nd term has a new representation such that

$$\|\sum_{x,y}\alpha_x(y)\Delta f_x(y)\|^2 = \sum_{x,x'}\sum_{(i,j)\in E}\sum_{(i',j')\in E}\sum_{y_i,y_j}\sum_{y_{i'},y_{j'}}\mu_x(y_i,y_j)\mu_{x'}(y_{i'},y_{j'})\Delta f_x(y_i,y_j)^T\Delta f_{x'}(y_{i'},y_{j'})$$

- Therefore the new equivalent formulation is to maximize

$$\sum_x\sum_{i\in V}\sum_{y_i}\mu_x(y_i)\Delta t_x(y_i) - \frac{1}{2}\sum_{x,x'}\sum_{(i,j)\in E}\sum_{(i',j')\in E}\sum_{y_i,y_j}\sum_{y_{i'},y_{j'}}\mu_x(y_i,y_j)\mu_{x'}(y_{i'},y_{j'})\Delta f_x(y_i,y_j)^T\Delta f_{x'}(y_{i'},y_{j'})$$

Q1. # of dual variables? (m examples, l binary outputs)

$$|\{\mu_x(y_i)\}_{x,y_i}| = ml \quad |\{\mu_x(y_i,y_j)\}_{x,y_i,y_j}| = ml^2 \Rightarrow ml(1+l)$$

Q2. # of addends $= ml\cdot 2 + m^2\cdot {}_lC_2{}^2\cdot 2^4$

Q3. What is a computational trade-off?

Polytope Constraints (1)

- New formulation is subject to *marginal polytope* constraint

$$\sum_{y_i} \mu_x(y_i) = C \ \ \forall_x, \forall_{i \in V}; \quad \sum_{y_i} \mu_x(y_i, y_j) = \mu_x(y_j) \quad \mu_x(y_i, y_j) \geq 0 \ \ \forall_x, \forall_{(i,j) \in E}$$

Polytope Constraints (1)

- New formulation is subject to *marginal polytope* constraint

$$\sum_{y_i} \mu_x(y_i) = C \ \ \forall_x, \forall_{i \in V}; \quad \sum_{y_i} \mu_x(y_i, y_j) = \mu_x(y_j) \quad \mu_x(y_i, y_j) \geq 0 \ \ \forall_x, \forall_{(i,j) \in E}$$

(Define 1) For given graph $G = (V, E)$, $Marg[G] :=$

$\{\{\mu_i(C_i)\}_{i \in V} \cup \{\mu_{ij}(S_{ij})\}_{(i,j) \in E} \mid {}^{\exists}$legal distribution Q_G such that $\{\mu_i\}$ & $\{\mu_{ij}\}$ are correct marginals of $Q_G\}$

Polytope Constraints (1)

- New formulation is subject to *marginal polytope* constraint

$$\sum_{y_i} \mu_x(y_i) = C \ \ \forall_x, \forall_{i \in V}; \quad \sum_{y_i} \mu_x(y_i, y_j) = \mu_x(y_j) \quad \mu_x(y_i, y_j) \geq 0 \ \ \forall_x, \forall_{(i,j) \in E}$$

(Define 1) For given graph $G = (V, E)$, $Marg[G] :=$

$$\{\{\mu_i(C_i)\}_{i \in V} \cup \{\mu_{ij}(S_{ij})\}_{(i,j) \in E} \mid {}^{\exists}\text{legal distribution } Q_G \text{ such that} \{\mu_i\} \ \& \ \{\mu_{ij}\} \text{ are correct marginals of } Q_G\}$$

(Define 2) For given graph $G = (V, E)$, $Local[G] :=$

$$\{\{\mu_i(C_i)\}_{i \in V} \cup \{\mu_{ij}(S_{ij})\}_{(i,j) \in E} \mid \text{marginals are locally consistent satisfying the calibration constraints}\}$$

Polytope Constraints (2)

Q1. Between $Marg[G]$ and $Local[G]$, which is the superset?

Polytope Constraints (2)

Q1. Between $Marg[G]$ and $Local[G]$, which is the superset?

(Fact) For general graph G,

Local[G] is the superset. That means Local[G] $\supseteq$ $Marg[G]$

Polytope Constraints (2)

Q1. Between $Marg[G]$ and $Local[G]$, which is the superset?

(Fact) For general graph G,

Local[G] is the superset. That means Local[G] $\supseteq Marg[G]$

Q2. Can you come up with an example in $Local[G] - Marg[G]$?

Polytope Constraints (2)

Q1. Between $Marg[G]$ and $Local[G]$, which is the superset?

(Fact) For general graph G,

Local[G] is the superset. That means Local[G] $\supseteq$ $Marg[G]$

Q2. Can you come up with an example in $Local[G] - Marg[G]$?

Think about the example given in the black board

Polytope Constraints (2)

Q1. Between $Marg[G]$ and $Local[G]$, which is the superset?

(Fact) For general graph G,

Local[G] is the superset. That means Local[G] $\supseteq Marg[G]$

Q2. Can you come up with an example in $Local[G] - Marg[G]$?

Think about the example given in the black board

- Q2-a. Is $\{\{\mu_1, \mu_2, \mu_3\}, \{\mu_{12}, \mu_{23}, \mu_{13}\}\} \in Local[G]$?

Polytope Constraints (2)

Q1. Between $Marg[G]$ and $Local[G]$, which is the superset?

(Fact) For general graph G,

Local[G] is the superset. That means Local[G] $\supseteq Marg[G]$

Q2. Can you come up with an example in $Local[G] - Marg[G]$?

Think about the example given in the black board

- Q2-a. Is $\{\{\mu_1, \mu_2, \mu_3\}, \{\mu_{12}, \mu_{23}, \mu_{13}\}\} \in Local[G]$?
- Q2-b. Is $\{\{\mu_1, \mu_2, \mu_3\}, \{\mu_{12}, \mu_{23}, \mu_{13}\}\} \in Marg[G]$?

Polytope Constraints (3)

- Our formulation requires marginal polytope constraint on tree-structured graph G

Polytope Constraints (3)

- Our formulation requires marginal polytope constraint on tree-structured graph G

(Theorem) If G:tree-structured

$Local[G] = Marg[G]$ (i.e., two polytopes are consistent)

Polytope Constraints (3)

- Our formulation requires marginal polytope constraint on tree-structured graph G

(Theorem) If G:tree-structured

$Local[G] = Marg[G]$ (i.e., two polytopes are consistent)

- Thus constraints coincide with the local consistency polytope

Polytope Constraints (3)

- Our formulation requires marginal polytope constraint on tree-structured graph G

(Theorem) If G:tree-structured

$Local[G] = Marg[G]$ (i.e., two polytopes are consistent)

- Thus constraints coincide with the local consistency polytope

Q. If the given graph G is not tree-structured?

Polytope Constraints (3)

- Our formulation requires marginal polytope constraint on tree-structured graph G

(Theorem) If G:tree-structured

$Local[G] = Marg[G]$ (i.e., two polytopes are consistent)

- Thus constraints coincide with the local consistency polytope

Q. If the given graph G is not tree-structured?

$\Rightarrow$ Solve the relaxed optimization on Local[G] via approximate algorithms such as loopy belief propagation.

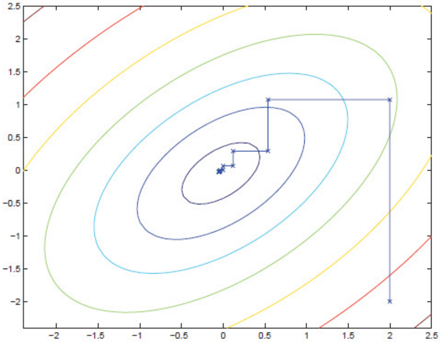
Coordinate Ascent/Descent

- Consider the problem of $\max_{\alpha_0,\ldots,\alpha_n} f(\alpha_0,\ldots,\alpha_n)$
- If we only want to reach local maximum (it is global if KKT is satisfied), we can replace the gradient with gradient in a predefined direction.

Coordinate Ascent

while until convergence:
for i $= 0$ **to** n :
$\alpha_i := \arg\max_{\alpha_i} f(\alpha_0,\ldots,\alpha_i,\ldots\alpha_n)$
Check KKT Conditions

Convergence: Same as gradient descent



Sequential Minimal Optimization (SMO)

- Recall the initial dual formulation.

$$\max \quad f = \sum_{x,y} \alpha_x(y)\Delta t_x(y) - \frac{1}{2}\left\|\sum_{x,y} \alpha_x(y)\Delta f_x(y)\right\|^2$$

$$s.t \quad \sum_{y} \alpha_x(y) = C \ \ \forall_x \qquad \alpha_x(y) \geq 0 \ \ \forall_{x,y}$$

- If we choose a specific coordinate $\alpha_x(y^1)$;

$$\alpha_x(y^1) = C - \sum_{y \in Y/y^1} \alpha_x(y)$$

- We can choose two coordinates y^1, y^2; then,

$$\alpha_x(y^1) + \alpha_x(y^2) = C - \textstyle\sum_{y \in Y/\{y^1,y^2\}} \alpha_x(y) = \gamma \implies \alpha_x(y^2) = \gamma - \alpha_x(y^1)$$
$$\max_{\alpha_x(y^1),\alpha_x(y^2)} f = \max_{\alpha_x(y^1)} a\alpha_x(y^1)^2 + b\alpha_x(y^1) + c$$

- Corresponding update in primal

$$\lambda = \alpha_x(y^1) - \alpha_x(y^1)'$$

$$\mu_x(y_i, y_j)' = \mu_x(y_i, y_j) + \lambda I[y_i = y_i^1, y_j = y_j^1] - \lambda I[y_i = y_i^2, y_j = y_j^2]$$

How to Train MMMN/SSVM in General?

- Polynomial-Size Reformulation
  - Exploit sparse dependency structure in underlying distribution
  - Implicit representation requires an inference in graphical model
- Cutting-plane Method
  - Efficiently manage only polynomially many working constraints
  - The next quadratic programming has only a different constraint
  - # of constraints needed can be large for good approximation
- Subgradient Method
  - Formulate the optimization objective as an unconstrained non-differentiable function having a maximum operation
  - # of iterations needed is improved $((O(1/\epsilon^2)$ vs $(O(1/\epsilon))$
  - The problem is that we haven't seen it yet!

Summary and Further Reading

- MMMN/SSVM allow us to encode various dependencies on completely general graph structures whereas HMM/CRF is mostly about linear/skip chain dependencies
- When a graph satisfies sub-modularity, computing maximum in min-max formulation can be efficiently solved by linear program via finding min-cut
- The exact inference to train the CRF is intractable in this case
- *Associative Max-Margin Markov Netowrks* by [Taskar 2004]
- *Dual Extragradient and Bregman Projections* by [Taskar 2006]
- *Learning Structural SVM with Latent Variables* by [Yu/Joachim 2009]

The End

Do you have any question?

Question

...Which tool do you use?...

Answer

...ShareLaTeX...