Probabilistic Graphical Models

Lecture 9: Variational Inference Relaxations

Volkan Cevher, Matthias Seeger
Ecole Polytechnique Fédérale de Lausanne

24/10/2011

Announcement

- No assignment this week
- Deadline programming assignment: June 18 (next lecture)
  `bayesml09lecture@googlemail.com`

Outline

1. Structured Mean Field (Variational Bayes)

2. Moment Parameters. Variational Relaxations

Variational Mean Field

$$\log Z \geq \sup_{Q \in \mathcal{Q}} \left\{ \mathrm{E}_Q[\Psi(\boldsymbol{x})] + \mathrm{H}[Q(\boldsymbol{x})] \right\}$$

- $\mathcal{Q}$: Tractable subset of all distributions (factorization constraints)

$$\mathcal{Q} = \left\{ Q(\boldsymbol{x}) = \prod_k Q_k(\boldsymbol{x}_{S_k}) \right\}, \quad S_k \text{ disjoint}$$

Tractable? For any k,
$\mathcal{N}_k$: Factor nodes j connected to any $i \in S_k$ ($S_k \cap C_j \neq \emptyset$)

$$Q_k'(\boldsymbol{x}_{S_k}) \propto \exp\left(\sum_{j \in \mathcal{N}_k} \mathrm{E}_{Q(\boldsymbol{x}_{C_j \setminus S_k})}[\Psi_j(\boldsymbol{x}_{C_j})] \right)$$

tractable to handle

F1

Variational Mean Field

$$\log Z \geq \sup_{Q \in \mathcal{Q}} \{ \mathrm{E}_Q[\Psi(\boldsymbol{x})] + \mathrm{H}[Q(\boldsymbol{x})] \}$$

- $\mathcal{Q}$: Tractable subset of all distributions (factorization constraints)

$$\mathcal{Q} = \left\{ Q(\boldsymbol{x}) = \prod_k Q_k(\boldsymbol{x}_{S_k}) \right\}, \quad S_k \text{ disjoint}$$

  Tractable? For any k,
  $\mathcal{N}_k$: Factor nodes j connected to any $i \in S_k$ ($S_k \cap C_j \neq \emptyset$)

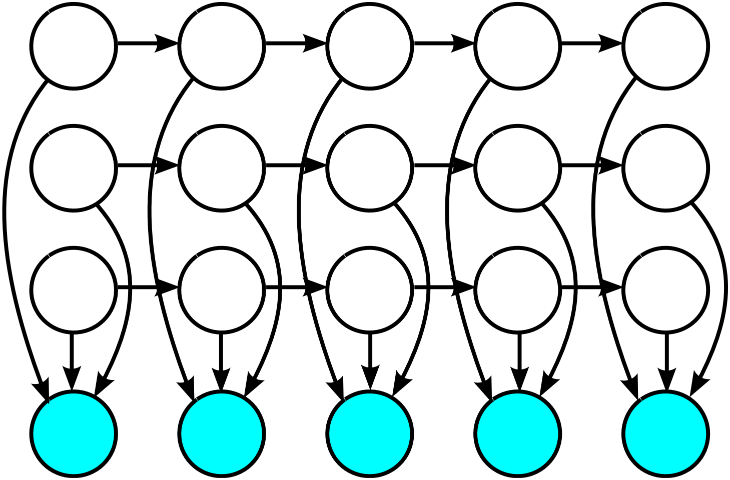
$$Q'_k(\boldsymbol{x}_{S_k}) \propto \exp\left(\sum_{j \in \mathcal{N}_k} \mathrm{E}_{Q(\boldsymbol{x}_{C_j \setminus S_k})}[\Psi_j(\boldsymbol{x}_{C_j})] \right)$$

  tractable to handle

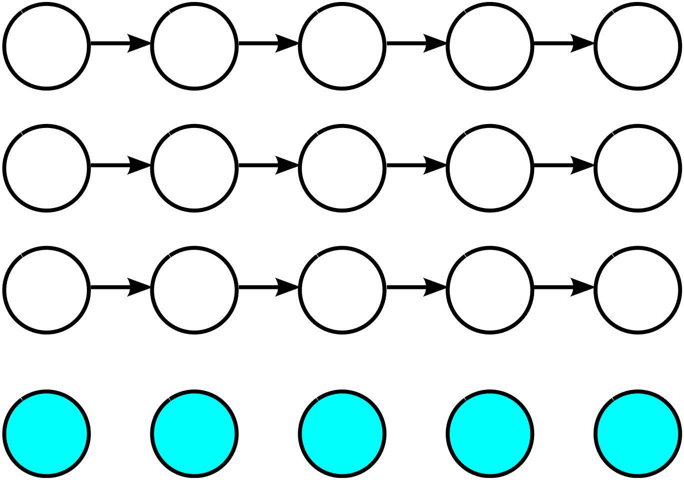
- $Q(\boldsymbol{x})$ completely factorized? Naive mean field
  Anything more elaborate? Structured mean field

Factorial Hidden Markov Model

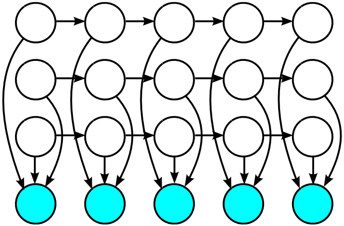
F2



Factorial Hidden Markov Model



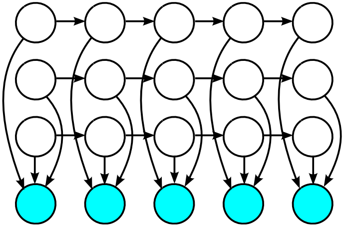
Factorial Hidden Markov Model

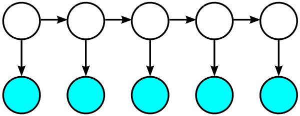


- S_1 = uppermost chain. Update?

F3

Factorial Hidden Markov Model





- S_1 = uppermost chain. Update?
- $Q(\boldsymbol{x}_{S_1})$: Markov chain (variable single node potentials)
  - Double node (transition) potentials of $Q(\boldsymbol{x}_{S_k})$? Fixed up front!
  - Forward-backward for single node marginals to update $Q(\boldsymbol{x}_{S_1})$. Implementation reduces to single HMM code, called with changing evidence potentials
- Not magic, but as expected:
  If this does not happen, you made a mistake

Variational Bayes

- Another instance of re-naming game:
  Nothing else than structured mean field
- Often applied to $P(\boldsymbol{x}, \theta | \boldsymbol{y})$
  ($\boldsymbol{y}$ observed, $\boldsymbol{x}$ latent nuisance, θ latent parameters)

Variational Bayes

- Another instance of re-naming game:
  Nothing else than structured mean field
- Often applied to $P(\boldsymbol{x}, \theta | \boldsymbol{y})$
  ($\boldsymbol{y}$ observed, $\boldsymbol{x}$ latent nuisance, θ latent parameters)

Expectation maximization	Variational Bayes
$\max_{\theta} \log \int P(\boldsymbol{y}, \boldsymbol{x} \| \theta)\, d\boldsymbol{x}$	$\log \int P(\boldsymbol{y}, \boldsymbol{x} \| \theta)\, d\boldsymbol{x}\, d\theta$
$\geq \max_{\theta, Q(\boldsymbol{x})} \left\{ \mathrm{E}_Q[\log P(\boldsymbol{y}, \boldsymbol{x} \| \theta)] + \mathrm{H}[Q(\boldsymbol{x})] \right\}$	$\geq \max_{Q(\theta), Q(\boldsymbol{x})} \left\{ \mathrm{E}_Q[\log P(\boldsymbol{y}, \boldsymbol{x} \| \theta)] + \mathrm{H}[Q(\boldsymbol{x})] + \mathrm{H}[Q(\theta)] \right\}$

Factorization assumption: $Q(\boldsymbol{x}, \theta) = Q(\boldsymbol{x}) Q(\theta)$

Variational Bayes

- Another instance of re-naming game:
  Nothing else than structured mean field
- Often applied to $P(\boldsymbol{x}, \boldsymbol{\theta} | \boldsymbol{y})$
  ($\boldsymbol{y}$ observed, $\boldsymbol{x}$ latent nuisance, $\boldsymbol{\theta}$ latent parameters)

Expectation maximization	Variational Bayes
$\max_{\boldsymbol{\theta}} \log \int P(\boldsymbol{y}, \boldsymbol{x} \vert \boldsymbol{\theta})\, d\boldsymbol{x}$	$\log \int P(\boldsymbol{y}, \boldsymbol{x} \vert \boldsymbol{\theta})\, d\boldsymbol{x}\, d\boldsymbol{\theta}$
$\geq \max_{\boldsymbol{\theta}, Q(\boldsymbol{x})} \Big\{ \mathrm{E}_Q[\log P(\boldsymbol{y}, \boldsymbol{x} \vert \boldsymbol{\theta})] + \mathrm{H}[Q(\boldsymbol{x})] \Big\}$	$\geq \max_{Q(\boldsymbol{\theta}), Q(\boldsymbol{x})} \Big\{ \mathrm{E}_Q[\log P(\boldsymbol{y}, \boldsymbol{x} \vert \boldsymbol{\theta})] + \mathrm{H}[Q(\boldsymbol{x})] + \mathrm{H}[Q(\boldsymbol{\theta})] \Big\}$

Factorization assumption: $Q(\boldsymbol{x}, \boldsymbol{\theta}) = Q(\boldsymbol{x}) Q(\boldsymbol{\theta})$

- Easy to write generic code (bit like MCMC Gibbs sampling)
- Good approximation?
  Can do better today for almost any well-studied model

Moment Parameters

$$\log Z = \sup_Q \{ \mathrm{E}_Q[\Psi(\boldsymbol{x})] + \mathrm{H}[Q(\boldsymbol{x})] \}$$

Moment Parameters

$$\log Z \geq \sup_{Q \in \mathcal{Q}} \{ \mathrm{E}_Q[\Psi(\boldsymbol{x})] + \mathrm{H}[Q(\boldsymbol{x})] \}$$

- $\mathcal{Q}$: Tractable subset of all distributions (factorization constraints)
  $\Rightarrow$ Seems whole story. What else could there be?

Moment Parameters

$$\log Z \geq \sup_{Q \in \mathcal{Q}} \{ \mathrm{E}_Q[\Psi(\boldsymbol{x})] + \mathrm{H}[Q(\boldsymbol{x})] \}$$

- $\mathcal{Q}$: Tractable subset of all distributions (factorization constraints)
  $\Rightarrow$ Seems whole story. What else could there be?
- Consider log-linear models: $\Psi_j(\boldsymbol{x}_{C_j}) = \boldsymbol{\theta}_j^T \boldsymbol{f}_j(\boldsymbol{x}_{C_j})$, $\boldsymbol{\theta} = (\boldsymbol{\theta}_j)$ F5

$$\mathrm{E}_Q[\Psi(\boldsymbol{x})] = \sum_j \boldsymbol{\theta}_j^T \boldsymbol{\mu}_j, \quad \boldsymbol{\mu}_j := \mathrm{E}_Q[\boldsymbol{f}_j(\boldsymbol{x}_{C_j})], \; \boldsymbol{\mu} = (\boldsymbol{\mu}_j)$$

Moment Parameters

$$\log Z \geq \sup_{Q \in \mathcal{Q}} \{\mathrm{E}_Q[\Psi(\boldsymbol{x})] + \mathrm{H}[Q(\boldsymbol{x})]\}$$

- $\mathcal{Q}$: Tractable subset of all distributions (factorization constraints) $\Rightarrow$ Seems whole story. What else could there be?
- Consider log-linear models: $\Psi_j(\boldsymbol{x}_{C_j}) = \boldsymbol{\theta}_j^T \boldsymbol{f}_j(\boldsymbol{x}_{C_j})$, $\boldsymbol{\theta} = (\boldsymbol{\theta}_j)$

$$\mathrm{E}_Q[\Psi(\boldsymbol{x})] = \sum_j \boldsymbol{\theta}_j^T \boldsymbol{\mu}_j, \quad \boldsymbol{\mu}_j := \mathrm{E}_Q[\boldsymbol{f}_j(\boldsymbol{x}_{C_j})], \ \boldsymbol{\mu} = (\boldsymbol{\mu}_j)$$

- Moment parameters: Under mild assumptions on $\boldsymbol{f}_j(\boldsymbol{x}_{C_j})$: Just another way (instead of $\boldsymbol{\theta}$) of parameterizing $P(\boldsymbol{x})$

Moment Parameters: Examples

$$P(\boldsymbol{x};\theta) = Z^{-1} e^{\theta^T \boldsymbol{f}(\boldsymbol{x})}$$

θ	Natural parameters
$\boldsymbol{f}(\boldsymbol{x})$	Statistics, representation
$\mu = \mathrm{E}_\theta[\boldsymbol{f}(\boldsymbol{x})]$	Moment parameters

Representation minimal: For every $\boldsymbol{z} \neq \mathbf{0}$, there is $\boldsymbol{x}$: $\boldsymbol{z}^T(\boldsymbol{f}(\boldsymbol{x})^T\,1)^T \neq 0$
Otherwise: Representation overcomplete F6

Moment Parameters: Examples

$$P(\boldsymbol{x};\theta) = Z^{-1} e^{\theta^T \boldsymbol{f}(\boldsymbol{x})}$$

$\boldsymbol{\theta}$	Natural parameters
$\boldsymbol{f}(\boldsymbol{x})$	Statistics, representation
$\boldsymbol{\mu} = \mathrm{E}_{\boldsymbol{\theta}}[\boldsymbol{f}(\boldsymbol{x})]$	Moment parameters

Representation minimal: For every $\boldsymbol{z} \neq \mathbf{0}$, there is $\boldsymbol{x}$:
$\boldsymbol{z}^T(\boldsymbol{f}(\boldsymbol{x})^T\, 1)^T \neq 0$
Otherwise: Representation overcomplete

1. Multinomial on graph with cliques C_j F6b
   Convenient overcomplete representation: Components of $\boldsymbol{f}(\boldsymbol{x})$: Indicators on cliques C_j, indicators on intersections of cliques, indicators on intersections of cliques, intersections, ...
   Equality constraints for $\boldsymbol{\mu}$:
   - Consistency on nonempty intersections
   - Sum to one on smallest intersections

Moment Parameters: Examples

$$P(\boldsymbol{x};\boldsymbol{\theta}) = Z^{-1} e^{\boldsymbol{\theta}^T \boldsymbol{f}(\boldsymbol{x})}$$

$\boldsymbol{\theta}$	Natural parameters
$\boldsymbol{f}(\boldsymbol{x})$	Statistics, representation
$\boldsymbol{\mu} = \mathrm{E}_{\boldsymbol{\theta}}[\boldsymbol{f}(\boldsymbol{x})]$	Moment parameters

Representation minimal: For every $\boldsymbol{z} \neq \boldsymbol{0}$, there is $\boldsymbol{x}$: $\boldsymbol{z}^T(\boldsymbol{f}(\boldsymbol{x})^T\ 1)^T \neq 0$

Otherwise: Representation overcomplete

2. Gaussian MRF

   Overcomplete representation:

F6c

$$\boldsymbol{f}(\boldsymbol{x}) = \begin{pmatrix} \boldsymbol{x} \\ \mathrm{vec}(-\boldsymbol{x}\boldsymbol{x}^T/2) \end{pmatrix}, \quad \boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{r} \\ \mathrm{vec}(\boldsymbol{A}) \end{pmatrix}$$

   Not minimal: $\boldsymbol{A}$ symmetric. $\{ij\} \notin E \rightarrow a_{ij} = a_{ji} = 0$.

Variational Formulation of Bayesian Inference

$$\log Z = \sup_Q \left\{ \theta^T \mathrm{E}_Q[\boldsymbol{f}(\boldsymbol{x})] + \mathrm{H}[Q(\boldsymbol{x})] \right\}, \quad \boldsymbol{f}(\boldsymbol{x}) = [\boldsymbol{f}_j(\boldsymbol{x}_{C_j})]$$

- Transform to moment parameters

Variational Formulation of Bayesian Inference

$$\log Z = \sup_{\boldsymbol{\mu} \in \mathcal{M}} \left\{ \boldsymbol{\theta}^T \boldsymbol{\mu} + \mathrm{H}[Q(\boldsymbol{x})] \right\}, \quad \boldsymbol{\mu}_j = \mathrm{E}_Q[\boldsymbol{f}_j(\boldsymbol{x}_{C_j})]$$

- Transform to moment parameters

$$\mathcal{M} = \left\{ (\boldsymbol{\mu}_j) \,\middle|\, \boldsymbol{\mu}_j = \mathrm{E}_Q[\boldsymbol{f}_j(\boldsymbol{x}_{C_j})] \text{ for some } Q(\boldsymbol{x}) \right\}$$

$\Rightarrow$ Marginal polytope F7

- What about the entropy?

Variational Formulation of Bayesian Inference

$$\log Z = \sup_{\boldsymbol{\mu} \in \mathcal{M}} \left\{ \boldsymbol{\theta}^T \boldsymbol{\mu} + \mathrm{H}[\boldsymbol{\mu}] \right\}$$

- Transform to moment parameters

$$\mathcal{M} = \left\{ (\boldsymbol{\mu}_j) \,\middle|\, \boldsymbol{\mu}_j = \mathrm{E}_Q[\boldsymbol{f}_j(\boldsymbol{x}_{C_j})] \text{ for some } Q(\boldsymbol{x}) \right\}$$

  $\Rightarrow$ Marginal polytope

- What about the entropy?
  $\boldsymbol{\mu} \in \mathcal{M} \leftrightarrow Q(\boldsymbol{x})$ unique: Entropy is function of $\boldsymbol{\mu}$
- Point of this exercise: $\mathcal{M}$ convex set of vectors, more useful relaxation target than set of distributions

Variational Formulation of Bayesian Inference

$$\log Z = \sup_{\boldsymbol{\mu} \in \mathcal{M}} \left\{ \boldsymbol{\theta}^T \boldsymbol{\mu} + \mathrm{H}[\boldsymbol{\mu}] \right\}$$

- Transform to moment parameters

$$\mathcal{M} = \left\{ (\boldsymbol{\mu}_j) \,\middle|\, \boldsymbol{\mu}_j = \mathrm{E}_Q[\boldsymbol{f}_j(\boldsymbol{x}_{C_j})] \text{ for some } Q(\boldsymbol{x}) \right\}$$

$\Rightarrow$ Marginal polytope

- What about the entropy?
  $\boldsymbol{\mu} \in \mathcal{M} \leftrightarrow Q(\boldsymbol{x})$ unique: Entropy is function of $\boldsymbol{\mu}$
- Point of this exercise: $\mathcal{M}$ convex set of vectors, more useful relaxation target than set of distributions
- Close now: Exponential families, Fenchel duality, maximum entropy. Full story:
  Wainwright, Jordan: Graphical Models, Exponential Families, and Variational Inference
  Foundations and Trends in Machine Learning, 1(1–2), pp. 1–305

Bayesian Inference is Convex Optimization

$$\log Z = \sup_{\mu \in \mathcal{M}} \left\{ \theta^T \mu + \mathrm{H}[\mu] \right\}$$

- Marginal polytope $\mathcal{M}$: Convex set F8

Bayesian Inference is Convex Optimization

$$\log Z = \underbrace{\sup_{\mu \in \mathcal{M}}}_{\mathcal{M}\ \text{convex}} \left\{ \theta^T \mu + \mathrm{H}[\mu] \right\}$$

- Marginal polytope $\mathcal{M}$: Convex set
- Entropy $\mu \mapsto \mathrm{H}[\mu]$: Concave function on $\mathcal{M}$ F8b

Bayesian Inference is Convex Optimization

$$\log Z = \underbrace{\sup_{\mu \in \mathcal{M}}}_{\mathcal{M}\text{ convex}} \underbrace{\left\{ \theta^T \mu + \mathrm{H}[\mu] \right\}}_{\text{concave}}$$

- Marginal polytope $\mathcal{M}$: Convex set
- Entropy $\mu \mapsto \mathrm{H}[\mu]$: Concave function on $\mathcal{M}$
- Posterior: Unique solution to convex optimization problem

Bayesian Inference is Convex Optimization

$$\log Z = \underbrace{\sup_{\mu \in \mathcal{M}}}_{\mathcal{M}\text{ convex}} \underbrace{\left\{ \theta^T \mu + \mathrm{H}[\mu] \right\}}_{\text{concave}}$$

- Marginal polytope $\mathcal{M}$: Convex set
- Entropy $\mu \mapsto \mathrm{H}[\mu]$: Concave function on $\mathcal{M}$
- Posterior: Unique solution to convex optimization problem
- Convex optimization can be intractable F8c

$\mathcal{M}$	can be hard to fence in
$\theta \leftrightarrow \mu$	can be hard to compute
$\mathrm{H}[\mu]$	can be hard to compute

Bayesian Inference is Convex Optimization

$$\log Z = \underbrace{\sup_{\mu \in \mathcal{M}}}_{\mathcal{M}\text{ convex}} \underbrace{\left\{ \theta^T \mu + \mathrm{H}[\mu] \right\}}_{\text{concave}}$$

- Marginal polytope $\mathcal{M}$: Convex set
- Entropy $\mu \mapsto \mathrm{H}[\mu]$: Concave function on $\mathcal{M}$
- Posterior: Unique solution to convex optimization problem
- Convex optimization can be intractable

$\mathcal{M}$	can be hard to fence in
$\theta \leftrightarrow \mu$	can be hard to compute
$\mathrm{H}[\mu]$	can be hard to compute

- Took some steps. But worth it:
  Rich literature on relaxations of hard convex problems

Variational Mean Field Revisited

$$\log Z = \sup_{\boldsymbol{\mu} \in \mathcal{M}} \left\{ \boldsymbol{\theta}^T \boldsymbol{\mu} + \mathrm{H}[\boldsymbol{\mu}] \right\}$$

- Have to approximate $\mathcal{M}$, $\mathrm{H}[\boldsymbol{\mu}]$. One way you already know ...

Variational Mean Field Revisited

$$\log Z = \sup_{\boldsymbol{\mu} \in \mathcal{M}} \left\{ \boldsymbol{\theta}^T \boldsymbol{\mu} + \mathrm{H}[\boldsymbol{\mu}] \right\}$$

- Have to approximate $\mathcal{M}$, $\mathrm{H}[\boldsymbol{\mu}]$. One way you already know . . .

$$\mathcal{M} \supset \mathcal{M}_{\text{NMF}} := \left\{ \boldsymbol{\mu} \,\middle|\, \mu_{C_j}(\boldsymbol{x}_{C_j}) = \sum\nolimits_{\boldsymbol{x}_{C_j}} \left(\prod\nolimits_{i \in C_j} Q(x_i) \right) \boldsymbol{f}_j(\boldsymbol{x}_{C_j}) \right\}$$

Inner approximation, induced by factorized distributions

- Entropy decomposes just as distribution: $\mathrm{H}[\boldsymbol{\mu}] = \sum_i \mathrm{H}[\boldsymbol{\mu}_i]$

$$\log Z \geq \sup_{\boldsymbol{\mu} \in \mathcal{M}_{\text{NMF}}} \left\{ \boldsymbol{\theta}^T \boldsymbol{\mu} + \sum\nolimits_i \mathrm{H}[\boldsymbol{\mu}_i] \right\}$$

Variational Mean Field Revisited

$$\log Z = \sup_{\boldsymbol{\mu} \in \mathcal{M}} \left\{ \boldsymbol{\theta}^T \boldsymbol{\mu} + \mathrm{H}[\boldsymbol{\mu}] \right\}$$

- Have to approximate $\mathcal{M}$, $\mathrm{H}[\boldsymbol{\mu}]$. One way you already know ...

$$\mathcal{M} \supset \mathcal{M}_{\text{NMF}} := \left\{ \boldsymbol{\mu} \;\middle|\; \mu_{C_j}(\boldsymbol{x}_{C_j}) = \sum_{\boldsymbol{x}_{C_j}} \left(\prod_{i \in C_j} Q(x_i) \right) \boldsymbol{f}_j(\boldsymbol{x}_{C_j}) \right\}$$

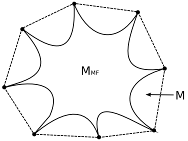
Inner approximation, induced by factorized distributions

- Entropy decomposes just as distribution: $\mathrm{H}[\boldsymbol{\mu}] = \sum_i \mathrm{H}[\boldsymbol{\mu}_i]$

$$\log Z \geq \sup_{\boldsymbol{\mu} \in \mathcal{M}_{\text{NMF}}} \left\{ \boldsymbol{\theta}^T \boldsymbol{\mu} + \sum_i \mathrm{H}[\boldsymbol{\mu}_i] \right\}$$

- Non-convex relaxation: $\mathcal{M}_{\text{NMF}}$ not convex

F9



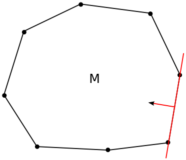
The Marginal Polytope

$$\mathcal{M} = \left\{ (\boldsymbol{\mu}_j) \,\middle|\, \boldsymbol{\mu}_j = \mathrm{E}_Q[\boldsymbol{f}_j(\boldsymbol{x}_{C_j})] \text{ for some } Q(\boldsymbol{x}) \right\}$$

Multinomial on graph. Minimal representation.

- $\mathcal{M}$ convex polytope: Described by finite number inequalities.
  Complexity of $\mathcal{M}$: Number of inequalities

F10

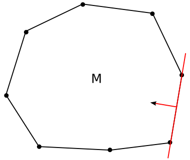


The Marginal Polytope

$$\mathcal{M} = \left\{ (\boldsymbol{\mu}_j) \,\middle|\, \boldsymbol{\mu}_j = \mathrm{E}_Q[\boldsymbol{f}_j(\boldsymbol{x}_{C_j})] \text{ for some } Q(\boldsymbol{x}) \right\}$$

Multinomial on graph. Minimal representation.

- $\mathcal{M}$ convex polytope: Described by finite number inequalities.
  Complexity of $\mathcal{M}$: Number of inequalities



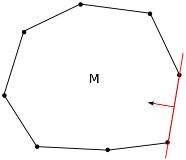
- Complexity of $\mathcal{M}$ $\rightarrow$ complexity of exact inference [we'll see why]
- $\mathcal{G}$ tree: $\mathcal{M}$ described by $O(n)$ inequalities [next lecture]

The Marginal Polytope

$$\mathcal{M} = \left\{ (\boldsymbol{\mu}_j) \,\middle|\, \boldsymbol{\mu}_j = \mathrm{E}_Q[\boldsymbol{f}_j(\boldsymbol{x}_{C_j})] \text{ for some } Q(\boldsymbol{x}) \right\}$$

Multinomial on graph. Minimal representation.

- $\mathcal{M}$ convex polytope: Described by finite number inequalities.
  Complexity of $\mathcal{M}$: Number of inequalities



- Complexity of $\mathcal{M} \to$ complexity of exact inference [we'll see why]
- $\mathcal{G}$ tree: $\mathcal{M}$ described by $O(n)$ inequalities [next lecture]
- Many graphs $\mathcal{G}$ with cycles: $\mathcal{M}$ polytope description provably hard ($\mathrm{poly}(n)$ inequalities would imply P=NP)

The Marginal Polytope

$$\mathcal{M} = \left\{ (\boldsymbol{\mu}_j) \,\middle|\, \boldsymbol{\mu}_j = \mathrm{E}_Q[\boldsymbol{f}_j(\boldsymbol{x}_{C_j})] \text{ for some } Q(\boldsymbol{x}) \right\}$$

Gaussian MRF. Minimal representation (upper triangle of $\boldsymbol{A}$).

- $\mathcal{M}$ exactly characterized by $\boldsymbol{\Sigma} = \boldsymbol{A}^{-1} \succ \boldsymbol{0}$.
  Convex cone (not polytope): Tractable to describe F10b
- $\mathcal{G}$ tree: $\mathcal{M}$ described by $O(n)$ inequalities
- General sparse $\mathcal{G}$: Approximate inference still of interest, if exact cost $O(n^3)$ too high

Wrap-Up

- Structured Mean Field: $Q(\boldsymbol{x})$ product of tractable, disjoint factors
- Variational Bayes: Another name for structured mean field
- Bayesian (marginal) inference is a convex optimization problem
- Variational approximations: Inner / outer bounds to marginal polytope