Probabilistic Graphical Models

Lecture 8: Information Theory. First Variational Approximation

Volkan Cevher, Matthias Seeger
Ecole Polytechnique Fédérale de Lausanne

24/10/2011

Outline

Why Approximate Inference?

- Why inference (computing marginal posterior distributions)? Essential backbone for (almost) anything todo with probabilistic model
  - Answering queries (honest answer: with uncertainties)
  - Learning model parameters
  - Making good decisions
  - Direct further data acquisition
  - Planning strategies (beyond single decisions)

Why Approximate Inference?

- Why inference (computing marginal posterior distributions)?
  Essential backbone for (almost) anything todo with probabilistic model
  - Answering queries (honest answer: with uncertainties)
  - Learning model parameters
  - Making good decisions
  - Direct further data acquisition
  - Planning strategies (beyond single decisions)
- Why approximate inference?
  Exact inference intractable for almost all real-world models
  - Loops in graphical model: Blow-up of intermediate representations, with no efficient (dynamic programming) way around
  - Potentials not closed under conditioning / marginalization: Blow-up of messages even for tree graphical models
- Bottomline: Bayesian inference powerful, consistent idea.
  Without approximate inference: Entirely academic exercise

Conjugate Priors

Transition Part I $\rightarrow$ Part II

I'll mention some things quickly, without us looking at them in more detail

Sometimes, inference is simple

$\boldsymbol{y}$	Observation
$\boldsymbol{\theta}$	Latent parameters (query)
$P(\boldsymbol{y}\vert\boldsymbol{\theta})$	Likelihood potential (positive function of $\boldsymbol{\theta}$)

Family of distributions $\mathcal{F} = \{P(\boldsymbol{\theta}|\boldsymbol{\alpha})\}$, $\boldsymbol{\alpha}$ fixed size: F2

- For every $\boldsymbol{y}$: $P(\boldsymbol{\theta}) \in \mathcal{F} \;\Rightarrow\; P(\boldsymbol{\theta}|\boldsymbol{y}) \in \mathcal{F}$
- If $P(\boldsymbol{\theta}) = P(\boldsymbol{\theta}|\boldsymbol{\alpha}_0)$, $P(\boldsymbol{\theta}|\boldsymbol{y}) = P(\boldsymbol{\theta}|\boldsymbol{\alpha}_1)$: For every $(\boldsymbol{\alpha}_0, \boldsymbol{y})$: $\boldsymbol{\alpha}_1$ easy to find

$\Rightarrow$ Inference a piece of cake! $\mathcal{F}$ conjugate to $P(\boldsymbol{y}|\boldsymbol{\theta})$ (or to $\{P(\boldsymbol{y}|\boldsymbol{\theta})\}$)

Markov Chain Monte Carlo

- General, maybe most flexible framework for approximate inference. Ideas from physics (thermodynamics, statistical mechanics)
- Not covered here (would need own course). I'll just give you cocktail party summary

Markov Chain Monte Carlo

1. Inference needs integrals $\int f(\boldsymbol{x})P(\boldsymbol{x})\,d\boldsymbol{x}$, $\boldsymbol{x}$ high-dimensional, $P(\boldsymbol{x})$ coupled, complicated (posterior)
2. Law of large numbers: $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N \sim P(\boldsymbol{x})$ independent: $N^{-1}\sum_i f(\boldsymbol{x}_i) \to \mathrm{E}_P[f(\boldsymbol{x})]$ almost surely.
   Central limit theorem: P, f nice $\Rightarrow$ Convergence as $1/\sqrt{N}$ independent of $\boldsymbol{x}$ dimensionality.
   Catch: Sampling from $P(\boldsymbol{x})$ hard as well

Markov Chain Monte Carlo

1. Inference needs integrals $\int f(\boldsymbol{x})P(\boldsymbol{x})\,d\boldsymbol{x}$, $\boldsymbol{x}$ high-dimensional, $P(\boldsymbol{x})$ coupled, complicated (posterior)
2. Law of large numbers: $\boldsymbol{x}_1,\dots,\boldsymbol{x}_N \sim P(\boldsymbol{x})$ independent:
   $N^{-1}\sum_i f(\boldsymbol{x}_i) \to \mathrm{E}_P[f(\boldsymbol{x})]$ almost surely.
   Central limit theorem: P, f nice $\Rightarrow$ Convergence as $1/\sqrt{N}$ independent of $\boldsymbol{x}$ dimensionality.
   Catch: Sampling from $P(\boldsymbol{x})$ hard as well
3. Let's just do something: Start with some $\boldsymbol{x}$, draw $\boldsymbol{x}' \sim K(\boldsymbol{x}'|\boldsymbol{x})$, keep doing that. At the very least:

$$P(\boldsymbol{x}') = \int K(\boldsymbol{x}'|\boldsymbol{x})P(\boldsymbol{x})\,d\boldsymbol{x}$$

   Such kernels K exist, need evaluation of $\propto P(\boldsymbol{x})$ only

Markov Chain Monte Carlo

$$P(\boldsymbol{x}') = \int K(\boldsymbol{x}'|\boldsymbol{x})P(\boldsymbol{x})\,d\boldsymbol{x}$$

4. MCMC magic: Under mild assumptions, that's all we need:
   $\boldsymbol{x}^{(j+1)} \sim K(\cdot|\boldsymbol{x}^{(j)}) \Rightarrow$ Marginal $\boldsymbol{x}^{(j)} \stackrel{D}{\to} P(\boldsymbol{x})$ as $j \to \infty$
   Rough idea why:
   - $K(\boldsymbol{x}'|\boldsymbol{x})$ contraction of probability mass. Information propagation with K brings marginal distributions closer together
   - There is only one fixed point (here: mild assumptions)

Markov Chain Monte Carlo

- MCMC used for many things things besides approximate inference
  - Theoretical CS: Counting of combinatorial sets. Volume estimation
  - Statistical physics: Evaluation of thermodynamical numbers (entropy, volume of macrostates). Studying phase transitions of coupled spin systems (magnets, spin glasses)
- Rich theory in the discrete case
- Related to, but different from stochastic optimization

Beware

BEWARE! MCMC sampling can be dangerous!

[OpenBUGS User Manual, page 1]

- MCMC: Simple to code. Hard to use properly
- You never exactly know when you're done
  - No definite convergence test in general
  - Hard to spot failures. Very hard to debug
  - Slow convergence can happen even with unimodal distributions, Gaussian tails

Beware

BEWARE! MCMC sampling can be dangerous!

[OpenBUGS User Manual, page 1]

- MCMC: Simple to code. Hard to use properly
- You never exactly know when you're done
  - No definite convergence test in general
  - Hard to spot failures. Very hard to debug
  - Slow convergence can happen even with unimodal distributions, Gaussian tails
- MCMC: Black box (in most cases), for good and for bad
  - Easy to code. For some problems, nothing else works. Safe if answers can be checked (search, exploration)
  - Can be very slow, or fail without you noticing. Always compare against something else if you can

Beware

BEWARE! MCMC sampling can be dangerous!

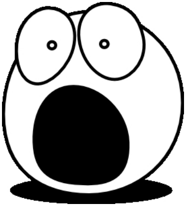
[OpenBUGS User Manual, page 1]

Dare to find out for yourself?

- Neal: Probabilistic Inference using Markov Chain Monte Carlo Methods (1993)

  [http://www.cs.toronto.edu/∼radford/papers-online.html]

- Gilks *et.al.*: Markov Chain Monte Carlo in Practice (1996)



Elements of Information Theory

Wake Up!

Transition time is over

Elements of Information Theory

Information Theory (Shannon, 1948)

- Narrow sense:
  - Limits of data compression (and how to achieve them)
  - Limits of error-free(!) communication over noisy channel
- Wide sense:
  - Basis of communication (language)
  - What is information? How to best encode it
  - Basis of anything adaptive, of learning
  - Source of great simplifications in number of mathematical domains
  - Information theory $\leftrightarrow$ applied probability / decision theory: Essentially equivalent in basic concepts, problems, methods

Good luck for students: Amazing textbook available:

- Cover, Thomas: Elements of Information Theory (1991)

Entropy of Distribution

$$\mathrm{H}[P(\boldsymbol{x})] = \mathrm{E}_P[-\log P(\boldsymbol{x})] = -\sum_{\boldsymbol{x}} P(\boldsymbol{x}) \log P(\boldsymbol{x})$$

- Game of questions: I draw $\boldsymbol{x} \sim P(\boldsymbol{x})$, give you P but not $\boldsymbol{x}$. How many questions $[\boldsymbol{x} \in \mathcal{E}]$ do you need to pin down $\boldsymbol{x}$? F6

Entropy of Distribution

$$\mathrm{H}[P(\boldsymbol{x})] = \mathrm{E}_P[-\log P(\boldsymbol{x})] = -\sum_{\boldsymbol{x}} P(\boldsymbol{x}) \log P(\boldsymbol{x})$$

- Game of questions: I draw $\boldsymbol{x} \sim P(\boldsymbol{x})$, give you P but not $\boldsymbol{x}$. How many questions $[\boldsymbol{x} \in \mathcal{E}]$ do you need to pin down $\boldsymbol{x}$?
- Shannon: On average: $\leq \mathrm{H}[P(\boldsymbol{x})] + 1$ questions if you're smart, no less than $\mathrm{H}[P(\boldsymbol{x})]$ even for a genius (log to base 2)
  $\Rightarrow$ Equivalent: Number bits needed to encode $\boldsymbol{x}$
  $\Rightarrow$ Amount of uncertainty in $P(\boldsymbol{x})$ F6b

Entropy of Distribution

$$\mathrm{H}[P(\boldsymbol{x})] = \mathrm{E}_P[-\log P(\boldsymbol{x})] = -\sum_{\boldsymbol{x}} P(\boldsymbol{x}) \log P(\boldsymbol{x})$$

- Game of questions: I draw $\boldsymbol{x} \sim P(\boldsymbol{x})$, give you P but not $\boldsymbol{x}$. How many questions $[\boldsymbol{x} \in \mathcal{E}]$ do you need to pin down $\boldsymbol{x}$?
- Shannon: On average: $\leq \mathrm{H}[P(\boldsymbol{x})] + 1$ questions if you're smart, no less than $\mathrm{H}[P(\boldsymbol{x})]$ even for a genius (log to base 2)
  $\Rightarrow$ Equivalent: Number bits needed to encode $\boldsymbol{x}$
  $\Rightarrow$ Amount of uncertainty in $P(\boldsymbol{x})$

Joint entropy	$\mathrm{H}[P(\boldsymbol{y}, \boldsymbol{x})] = \mathrm{E}_P[-\log P(\boldsymbol{y}, \boldsymbol{x})]$
Conditional entropy	$\mathrm{H}[P(\boldsymbol{y} \mid \boldsymbol{x})] = \mathrm{E}_P[-\log P(\boldsymbol{y} \mid \boldsymbol{x})]$

F6c

Entropy of Distribution

$$\mathrm{H}[P(\boldsymbol{x})] = \mathrm{E}_P[-\log P(\boldsymbol{x})] = -\sum_{\boldsymbol{x}} P(\boldsymbol{x}) \log P(\boldsymbol{x})$$

- Game of questions: I draw $\boldsymbol{x} \sim P(\boldsymbol{x})$, give you P but not $\boldsymbol{x}$. How many questions $[\boldsymbol{x} \in \mathcal{E}]$ do you need to pin down $\boldsymbol{x}$?
- Shannon: On average: $\leq \mathrm{H}[P(\boldsymbol{x})] + 1$ questions if you're smart, no less than $\mathrm{H}[P(\boldsymbol{x})]$ even for a genius (log to base 2)
  $\Rightarrow$ Equivalent: Number bits needed to encode $\boldsymbol{x}$
  $\Rightarrow$ Amount of uncertainty in $P(\boldsymbol{x})$

Joint entropy	$\mathrm{H}[P(\boldsymbol{y}, \boldsymbol{x})] = \mathrm{E}_P[-\log P(\boldsymbol{y}, \boldsymbol{x})]$
Conditional entropy	$\mathrm{H}[P(\boldsymbol{y} \vert \boldsymbol{x})] = \mathrm{E}_P[-\log P(\boldsymbol{y} \vert \boldsymbol{x})]$

Chain rule of entropy: F6d

$$\mathrm{H}[P(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)] = \sum_{i=1}^{n} \mathrm{H}[P(\boldsymbol{x}_i \vert \boldsymbol{x}_{<i})]$$

Relative Entropy

$$\mathrm{D}[P(\boldsymbol{x}) \,\|\, Q(\boldsymbol{x})] = \mathrm{E}_P\left[\log \frac{P(\boldsymbol{x})}{Q(\boldsymbol{x})}\right]$$

- Game of questions. This time, you get it wrong. You think $\boldsymbol{x} \sim Q(\boldsymbol{x})$, but in fact $\boldsymbol{x} \sim P(\boldsymbol{x})$. How many questions? F7

Relative Entropy

$$\mathrm{D}[P(\boldsymbol{x}) \,\|\, Q(\boldsymbol{x})] = \mathrm{E}_P\left[\log \frac{P(\boldsymbol{x})}{Q(\boldsymbol{x})}\right]$$

- Game of questions. This time, you get it wrong. You think $\boldsymbol{x} \sim Q(\boldsymbol{x})$, but in fact $\boldsymbol{x} \sim P(\boldsymbol{x})$. How many questions?
- On average: $\mathrm{E}_P[-\log Q(\boldsymbol{x})] = \mathrm{H}[P(\boldsymbol{x})] + \mathrm{D}[P(\boldsymbol{x}) \,\|\, Q(\boldsymbol{x})]$
  $\Rightarrow$ Number of additional bits for using Q instead of true P
  $\Rightarrow$ Natural divergence (distance) measure between distributions F7b

Relative Entropy

$$\mathrm{D}[P(\boldsymbol{x}) \,\|\, Q(\boldsymbol{x})] = \mathrm{E}_P\left[\log \frac{P(\boldsymbol{x})}{Q(\boldsymbol{x})}\right]$$

- Game of questions. This time, you get it wrong. You think $\boldsymbol{x} \sim Q(\boldsymbol{x})$, but in fact $\boldsymbol{x} \sim P(\boldsymbol{x})$. How many questions?
- On average: $\mathrm{E}_P[-\log Q(\boldsymbol{x})] = \mathrm{H}[P(\boldsymbol{x})] + \mathrm{D}[P(\boldsymbol{x}) \,\|\, Q(\boldsymbol{x})]$
  $\Rightarrow$ Number of additional bits for using Q instead of true P
  $\Rightarrow$ Natural divergence (distance) measure between distributions
- Other name: Kullback-Leibler divergence.
  No distance: $\mathrm{D}[P \,\|\, Q] \neq \mathrm{D}[Q \,\|\, P]$

Relative Entropy

$$\mathrm{D}[P(\boldsymbol{x}) \,\|\, Q(\boldsymbol{x})] = \mathrm{E}_P\left[\log \frac{P(\boldsymbol{x})}{Q(\boldsymbol{x})}\right]$$

- Game of questions. This time, you get it wrong. You think $\boldsymbol{x} \sim Q(\boldsymbol{x})$, but in fact $\boldsymbol{x} \sim P(\boldsymbol{x})$. How many questions?
- On average: $\mathrm{E}_P[-\log Q(\boldsymbol{x})] = \mathrm{H}[P(\boldsymbol{x})] + \mathrm{D}[P(\boldsymbol{x}) \,\|\, Q(\boldsymbol{x})]$
  $\Rightarrow$ Number of additional bits for using Q instead of true P
  $\Rightarrow$ Natural divergence (distance) measure between distributions
- Other name: Kullback-Leibler divergence.
  No distance: $\mathrm{D}[P \,\|\, Q] \neq \mathrm{D}[Q \,\|\, P]$

Conditional relative entropy:
$\mathrm{D}[P(\boldsymbol{y}|\boldsymbol{x}) \,\|\, Q(\boldsymbol{y}|\boldsymbol{x})] = \mathrm{E}_P[\log\{P(\boldsymbol{y}|\boldsymbol{x})/Q(\boldsymbol{y}|\boldsymbol{x})\}]$
Chain rule of relative entropy:
$\mathrm{D}[P(\boldsymbol{y},\boldsymbol{x}) \,\|\, Q(\boldsymbol{y},\boldsymbol{x})] = \mathrm{D}[P(\boldsymbol{y}|\boldsymbol{x}) \,\|\, Q(\boldsymbol{y}|\boldsymbol{x})] + \mathrm{D}[P(\boldsymbol{x}) \,\|\, Q(\boldsymbol{x})]$ F7c

Mutual Information

$$\mathrm{I}(\boldsymbol{x};\boldsymbol{y}) = \mathrm{D}[P(\boldsymbol{x},\boldsymbol{y}) \,\|\, P(\boldsymbol{x})P(\boldsymbol{y})] = \mathrm{E}_P\left[\log\frac{P(\boldsymbol{x},\boldsymbol{y})}{P(\boldsymbol{x})P(\boldsymbol{y})}\right]$$

- $\boldsymbol{x}$, $\boldsymbol{y}$ may be dependent. How many additional questions (bits) for ignoring that? F8

Mutual Information

$$\mathrm{I}(\boldsymbol{x};\boldsymbol{y}) = \mathrm{D}[P(\boldsymbol{x},\boldsymbol{y}) \,\|\, P(\boldsymbol{x})P(\boldsymbol{y})] = \mathrm{E}_P\left[\log \frac{P(\boldsymbol{x},\boldsymbol{y})}{P(\boldsymbol{x})P(\boldsymbol{y})}\right]$$

- $\boldsymbol{x}$, $\boldsymbol{y}$ may be dependent. How many additional questions (bits) for ignoring that?
- Mutual information: Reduction in uncertainty of one random variable due to knowledge of other

$$\mathrm{I}(\boldsymbol{x};\boldsymbol{y}) = \mathrm{H}[P(\boldsymbol{x})] - \mathrm{H}[P(\boldsymbol{x}|\boldsymbol{y})] = \mathrm{H}[P(\boldsymbol{y})] - \mathrm{H}[P(\boldsymbol{y}|\boldsymbol{x})]$$

$\Rightarrow$ Amount of information $\boldsymbol{x}$ about $\boldsymbol{y}$, or $\boldsymbol{y}$ about $\boldsymbol{x}$

Mutual Information

$$\mathrm{I}(\boldsymbol{x};\boldsymbol{y}) = \mathrm{D}[P(\boldsymbol{x},\boldsymbol{y}) \,\|\, P(\boldsymbol{x})P(\boldsymbol{y})] = \mathrm{E}_P\left[\log \frac{P(\boldsymbol{x},\boldsymbol{y})}{P(\boldsymbol{x})P(\boldsymbol{y})}\right]$$

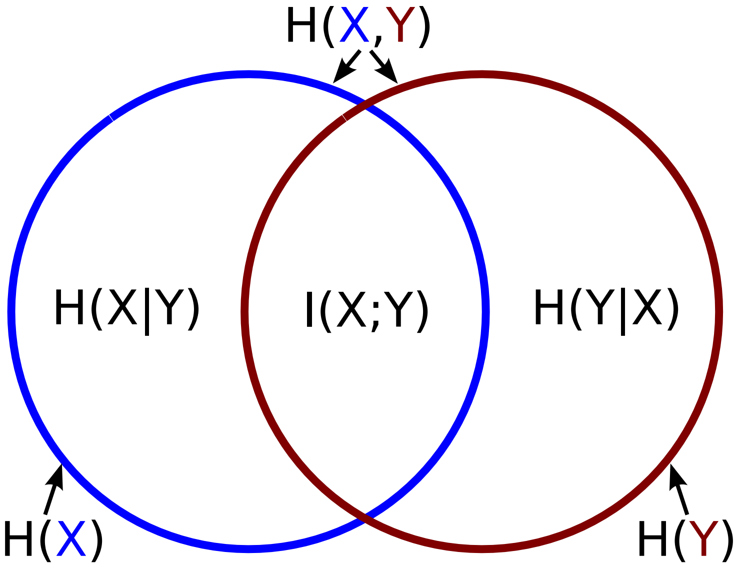
- $\boldsymbol{x}$, $\boldsymbol{y}$ may be dependent. How many additional questions (bits) for ignoring that?
- Mutual information: Reduction in uncertainty of one random variable due to knowledge of other

$$\mathrm{I}(\boldsymbol{x};\boldsymbol{y}) = \mathrm{H}[P(\boldsymbol{x})] - \mathrm{H}[P(\boldsymbol{x}|\boldsymbol{y})] = \mathrm{H}[P(\boldsymbol{y})] - \mathrm{H}[P(\boldsymbol{y}|\boldsymbol{x})]$$

$\Rightarrow$ Amount of information $\boldsymbol{x}$ about $\boldsymbol{y}$, or $\boldsymbol{y}$ about $\boldsymbol{x}$

- Note: $\boldsymbol{x} \perp \boldsymbol{y}$ (independent) $\Rightarrow P(\boldsymbol{x},\boldsymbol{y}) = P(\boldsymbol{x})P(\boldsymbol{y}) \Rightarrow \mathrm{I}(\boldsymbol{x};\boldsymbol{y}) = 0$. We'll see $\Leftarrow$. Mutual information: Measure of dependence

Venn Diagram for Information



Information Inequality

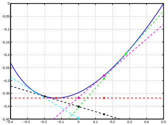
- Something missing here
  - More questions for getting it wrong: $\mathrm{H}[P(\boldsymbol{x})] \rightarrow \mathrm{H}[P(\boldsymbol{x})] + \mathrm{D}[P(\boldsymbol{x}) \,\|\, Q(\boldsymbol{x})]$
  - $\mathrm{I}(\boldsymbol{x};\boldsymbol{y})$ measures dependence. $\mathrm{I}(\boldsymbol{x};\boldsymbol{y}) = 0$ for $\boldsymbol{x} \perp \boldsymbol{y}$

Is $\mathrm{D}[P(\boldsymbol{x}) \,\|\, Q(\boldsymbol{x})] \geq 0$? Is $\mathrm{I}(\boldsymbol{x};\boldsymbol{y}) \geq 0$?

Information Inequality

- Something missing here
  - More questions for getting it wrong:
    $\mathrm{H}[P(\boldsymbol{x})] \to \mathrm{H}[P(\boldsymbol{x})] + \mathrm{D}[P(\boldsymbol{x}) \parallel Q(\boldsymbol{x})]$
  - $\mathrm{I}(\boldsymbol{x};\boldsymbol{y})$ measures dependence. $\mathrm{I}(\boldsymbol{x};\boldsymbol{y}) = 0$ for $\boldsymbol{x} \perp \boldsymbol{y}$

  Is $\mathrm{D}[P(\boldsymbol{x}) \parallel Q(\boldsymbol{x})] \geq 0$? Is $\mathrm{I}(\boldsymbol{x};\boldsymbol{y}) \geq 0$?
- Convexity comes to the rescue. Jensen's inequality F10

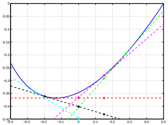


$$\mathrm{E}_P[f(\boldsymbol{x})] \geq f(\mathrm{E}_P[\boldsymbol{x}]), \quad f \text{ convex}$$

Information Inequality

- Something missing here
  - More questions for getting it wrong: $\mathrm{H}[P(\boldsymbol{x})] \to \mathrm{H}[P(\boldsymbol{x})] + \mathrm{D}[P(\boldsymbol{x}) \,\|\, Q(\boldsymbol{x})]$
  - $\mathrm{I}(\boldsymbol{x};\boldsymbol{y})$ measures dependence. $\mathrm{I}(\boldsymbol{x};\boldsymbol{y}) = 0$ for $\boldsymbol{x} \perp \boldsymbol{y}$

  Is $\mathrm{D}[P(\boldsymbol{x}) \,\|\, Q(\boldsymbol{x})] \geq 0$? Is $\mathrm{I}(\boldsymbol{x};\boldsymbol{y}) \geq 0$?
- Convexity comes to the rescue. Jensen's inequality



$$\mathrm{E}_P[f(\boldsymbol{x})] \geq f(\mathrm{E}_P[\boldsymbol{x}]), \quad f \text{ convex}$$

- Information inequality: $\mathrm{D}[P(\boldsymbol{x}) \,\|\, Q(\boldsymbol{x})] \geq 0$.
  Since $-\log(\cdot)$ strictly convex (nowhere linear):
  $\mathrm{D}[P(\boldsymbol{x}) \,\|\, Q(\boldsymbol{x})] = 0 \quad \Leftrightarrow \quad P(\boldsymbol{x}) = Q(\boldsymbol{x})$ P-almost everywhere.

$$\mathrm{I}(\boldsymbol{x};\boldsymbol{y}) \geq 0; \quad \mathrm{I}(\boldsymbol{x};\boldsymbol{y}) = 0 \;\Leftrightarrow\; \boldsymbol{x} \perp \boldsymbol{y}$$

Corollaries

Raking in the fruits

- Conditioning reduces entropy (learning always helps) F11

$$\mathrm{H}[P(\boldsymbol{x}|\boldsymbol{y})] \leq \mathrm{H}[P(\boldsymbol{x})]$$

Corollaries

Raking in the fruits

- Conditioning reduces entropy (learning always helps)

$$\mathrm{H}[P(\boldsymbol{x}|\boldsymbol{y})] \leq \mathrm{H}[P(\boldsymbol{x})]$$

- Conditional mutual information:
  Measure for conditional independence F11b

$$\mathrm{I}(\boldsymbol{x};\boldsymbol{y}|\boldsymbol{z}) = 0 \quad \Leftrightarrow \quad \boldsymbol{x} \perp \boldsymbol{y} \mid \boldsymbol{z}$$

Corollaries

Raking in the fruits

- Conditioning reduces entropy (learning always helps)

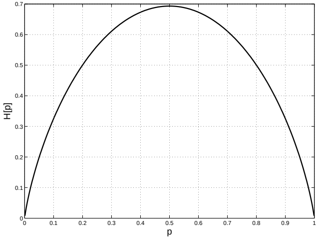
$$\mathrm{H}[P(\boldsymbol{x}|\boldsymbol{y})] \leq \mathrm{H}[P(\boldsymbol{x})]$$

- Conditional mutual information:
  Measure for conditional independence

$$\mathrm{I}(\boldsymbol{x};\boldsymbol{y}|\boldsymbol{z}) = 0 \quad \Leftrightarrow \quad \boldsymbol{x} \perp \boldsymbol{y} \mid \boldsymbol{z}$$

- Entropy: Concave function

$$\mathrm{H}[\lambda P(\boldsymbol{x}) + (1-\lambda) Q(\boldsymbol{x})]$$
$$\geq \lambda \mathrm{H}[P(\boldsymbol{x})] + (1-\lambda)\mathrm{H}[Q(\boldsymbol{x})]$$



Remember EM?

One approach to variational approximate inference:
Computations with $P(\boldsymbol{x}) = Z^{-1} e^{\Psi(\boldsymbol{x})}$ hard (even $\log Z$)?
$\Rightarrow$ Approximate it by $Q(\boldsymbol{x})$, for which computations simple

Remember derivation of EM? F12

Remember EM?

One approach to variational approximate inference:
Computations with $P(\boldsymbol{x}) = Z^{-1} e^{\Psi(\boldsymbol{x})}$ hard (even $\log Z$)?
$\Rightarrow$ Approximate it by $Q(\boldsymbol{x})$, for which computations simple

Remember derivation of EM?

$$\log Z = \log \int e^{\Psi(\boldsymbol{x})}\, d\boldsymbol{x} = \sup_Q \mathrm{E}_Q[\log\{e^{\Psi(\boldsymbol{x})}/Q(\boldsymbol{x})\}]$$
$$= \sup_Q \{\mathrm{E}_Q[\Psi(\boldsymbol{x})] + \mathrm{H}[Q(\boldsymbol{x})]\}$$

Was called variational mean field inequality. Let's see why

Remember EM?

One approach to variational approximate inference:
Computations with $P(\boldsymbol{x}) = Z^{-1} e^{\Psi(\boldsymbol{x})}$ hard (even $\log Z$)?
$\Rightarrow$ Approximate it by $Q(\boldsymbol{x})$, for which computations simple

Remember derivation of EM?

$$\log Z = \log \int e^{\Psi(\boldsymbol{x})}\, d\boldsymbol{x} = \sup_Q \mathrm{E}_Q[\log\{e^{\Psi(\boldsymbol{x})}/Q(\boldsymbol{x})\}]$$
$$= \sup_Q \{\mathrm{E}_Q[\Psi(\boldsymbol{x})] + \mathrm{H}[Q(\boldsymbol{x})]\}$$

Was called variational mean field inequality. Let's see why

- Maximizer: $Q(\boldsymbol{x}) = P(\boldsymbol{x})$ itself. Attains $\log Z$ F12b
- Any other $Q(\boldsymbol{x})$: Lower bound. $Q(\boldsymbol{x})$ closer to $P(\boldsymbol{x})$?
  $\Rightarrow$ Maximize the lower bound!

Remember EM?

One approach to variational approximate inference:
Computations with $P(\boldsymbol{x}) = Z^{-1} e^{\Psi(\boldsymbol{x})}$ hard (even $\log Z$)?
$\Rightarrow$ Approximate it by $Q(\boldsymbol{x})$, for which computations simple

Remember derivation of EM?

$$\log Z = \log \int e^{\Psi(\boldsymbol{x})}\, d\boldsymbol{x} = \sup_Q \mathrm{E}_Q[\log\{e^{\Psi(\boldsymbol{x})}/Q(\boldsymbol{x})\}]$$
$$= \sup_Q \{\mathrm{E}_Q[\Psi(\boldsymbol{x})] + \mathrm{H}[Q(\boldsymbol{x})]\}$$

Was called variational mean field inequality. Let's see why

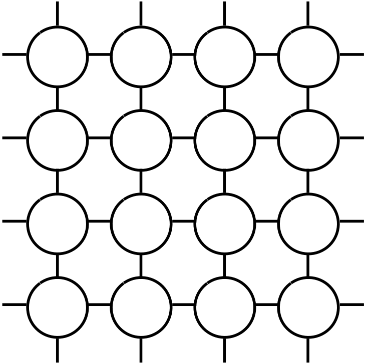
- Maximizer: $Q(\boldsymbol{x}) = P(\boldsymbol{x})$ itself. Attains $\log Z$
- Any other $Q(\boldsymbol{x})$: Lower bound. $Q(\boldsymbol{x})$ closer to $P(\boldsymbol{x})$?
  $\Rightarrow$ Maximize the lower bound!
- Relax this problem: Work with $\mathcal{Q} = \{Q(\boldsymbol{x})\}$: F12c
  - Lower bound can be evaluated for each $Q \in \mathcal{Q}$
  - Bayesian computations can be done with any $Q \in \mathcal{Q}$ (not with P)

Naive Mean Field for Markov Random Fields

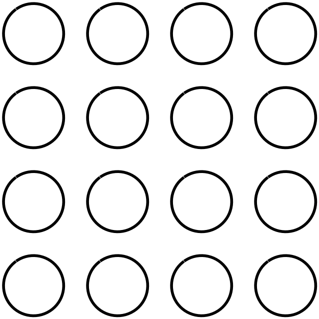
Distributions complicated, because they are coupled
$\Rightarrow$ Mean field: Approximate them by factorizing distributions

Naive Mean Field: Drop all edges

True MRF posterior $P(\boldsymbol{x})$



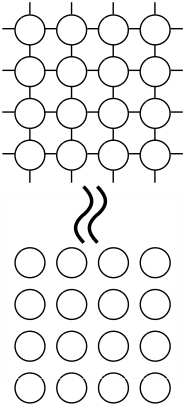
Approximations $Q(\boldsymbol{x}) \in \mathcal{Q}$



Naive Mean Field for Markov Random Fields

Variational problem:

$$\operatorname*{argmax}_{\{Q(x_k)\}} \left\{ \sum_j \mathrm{E}_Q[\Psi_j(\boldsymbol{x}_{C_j})] + \sum_k \mathrm{H}[Q(x_k)] \right\}$$



Our first variational algorithm:

Default-initialize $Q(\boldsymbol{x}_k)$ (say: uniform)
repeat
  Pick some node k at random
  Update $Q(x_k)$, keeping all others fixed

$$Q(x_k) \leftarrow \operatorname{argmax} \left\{ \sum_{j \in \mathcal{N}_k} \mathrm{E}_Q[\Psi_j(\boldsymbol{x}_{C_j})] + \mathrm{H}[Q(x_k)] \right\}$$

until Convergence

Prize question: How does that update look like?

F14

Remarks

- Does this always converge? Yes. To a unique solution? No F15
- How to compare different fixed points? Or even different $\mathcal{Q}$? You get lower bound to $\log Z$

Remarks

- Does this always converge? Yes. To a unique solution? No
- How to compare different fixed points? Or even different $\mathcal{Q}$?
  You get lower bound to $\log Z$
- Why "mean field"? $P(\boldsymbol{x})$: Random field. $Q(\boldsymbol{x}) \approx P(\boldsymbol{x})$ has no couplings ($\mathrm{E}_Q[x_j x_k] = \mathrm{E}_Q[x_j]\mathrm{E}_Q[x_k]$).
  True means at convergence ($\mathrm{E}_Q[x_j] = \mathrm{E}_P[x_j]$)? No
  (remember: $\mathrm{E}_P[x_j]$ hard as well!)

Remarks

- Does this always converge? Yes. To a unique solution? No
- How to compare different fixed points? Or even different $\mathcal{Q}$?
  You get lower bound to $\log Z$
- Why "mean field"? $P(\boldsymbol{x})$: Random field. $Q(\boldsymbol{x}) \approx P(\boldsymbol{x})$ has no couplings ($\mathrm{E}_Q[x_j x_k] = \mathrm{E}_Q[x_j]\mathrm{E}_Q[x_k]$).
  True means at convergence ($\mathrm{E}_Q[x_j] = \mathrm{E}_P[x_j]$)? No
  (remember: $\mathrm{E}_P[x_j]$ hard as well!)
- General idea here: Relax variational problem
  $$\sup_Q(\dots) \;\geq\; \sup_{Q\in\mathcal{Q}}(\dots)$$
  $\mathcal{Q}$: Subset of all distributions (factorization constraints). Each $Q(\boldsymbol{x})$ is distribution.
  $\Rightarrow$ Maximize lower bound over $\mathcal{Q}$
  Note: Might not find maximizer $Q \in \mathcal{Q}$, but local maximum

Variational Mean Field: Minimizing Relative Entropy

$$\log Z = \log \int e^{\Psi(\boldsymbol{x})}\, d\boldsymbol{x} \geq \mathrm{E}_Q[\Psi(\boldsymbol{x})] + \mathrm{H}[Q(\boldsymbol{x})], \quad P(\boldsymbol{x}) = Z^{-1} e^{\Psi(\boldsymbol{x})}$$

- What is the slack in this bound?
  Hint: ≥ 0, and $= 0$ iff $Q(\boldsymbol{x}) = P(\boldsymbol{x})$ F16

Variational Mean Field: Minimizing Relative Entropy

$$\log Z = \log \int e^{\Psi(\boldsymbol{x})}\, d\boldsymbol{x} \geq \mathrm{E}_Q[\Psi(\boldsymbol{x})] + \mathrm{H}[Q(\boldsymbol{x})], \quad P(\boldsymbol{x}) = Z^{-1} e^{\Psi(\boldsymbol{x})}$$

- What is the slack in this bound?
  Hint: ≥ 0, and $= 0$ iff $Q(\boldsymbol{x}) = P(\boldsymbol{x})$
- Variational mean field: Minimize slack (relative entropy)

$$\min_{Q \in \mathcal{Q}} \mathrm{D}[Q(\boldsymbol{x}) \,\|\, P(\boldsymbol{x})]$$

Does that fit relative entropy semantics?

Variational Mean Field: Minimizing Relative Entropy

$$\log Z = \log \int e^{\Psi(\boldsymbol{x})}\, d\boldsymbol{x} \geq \mathrm{E}_Q[\Psi(\boldsymbol{x})] + \mathrm{H}[Q(\boldsymbol{x})], \quad P(\boldsymbol{x}) = Z^{-1} e^{\Psi(\boldsymbol{x})}$$

- What is the slack in this bound?
  Hint: ≥ 0, and $= 0$ iff $Q(\boldsymbol{x}) = P(\boldsymbol{x})$
- Variational mean field: Minimize slack (relative entropy)

$$\min_{Q \in \mathcal{Q}} \mathrm{D}[Q(\boldsymbol{x}) \,\|\, P(\boldsymbol{x})]$$

  Does that fit relative entropy semantics?
- It's the wrong way around! We should minimize $\mathrm{D}[P \,\|\, Q]$.
  Alas, even that is hard. For naive mean field, unique solution is

$$Q(x_1, \dots, x_N) = P(x_1) \dots P(x_N)$$

  Variational mean field: a tractable compromise

Wrap-Up

- Information theory: Fundamental characteristics and limits to compression and faultless information transmission
- Statistical learning, information theory: Different sides of the same coin
- Variational mean field: Tractable approximate inference by factorization assumptions
- Naive mean field: Drop all edges, update node by node