Probabilistic Graphical Models

Lecture 8: Information Theory. First Variational Approximation

Volkan Cevher, Matthias Seeger
Ecole Polytechnique Fédérale de Lausanne

24/10/2011
Outline

1. Why Approximate Inference?
2. Some Information Theory
3. Variational Mean Field Approximations
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
Sometimes, inference is simple

\( y \) Observation
\( \theta \) Latent parameters (query)
\( P(y|\theta) \) Likelihood potential (positive function of \( \theta \))

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

- For every \( y \): \( P(\theta) \in \mathcal{F} \) \( \Rightarrow \) \( P(\theta|y) \in \mathcal{F} \)
- If \( P(\theta) = P(\theta|\alpha_0) \), \( P(\theta|y) = P(\theta|\alpha_1) \): For every \( (\alpha_0, y) \):
  \( \alpha_1 \) easy to find

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

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
1. Inference needs integrals $\int f(x)P(x)\,dx$, $x$ high-dimensional, $P(x)$ coupled, complicated (posterior).

2. Law of large numbers: $x_1, \ldots, x_N \sim P(x)$ independent:
   $N^{-1} \sum_i f(x_i) \to E_P[f(x)]$ almost surely.
   Central limit theorem: $P$, $f$ nice $\Rightarrow$ Convergence as $1/\sqrt{N}$
   independent of $x$ dimensionality.
   Catch: Sampling from $P(x)$ hard as well.
Inference needs integrals $\int f(x) P(x) \, dx$, $x$ high-dimensional, $P(x)$ coupled, complicated (posterior)

Law of large numbers: $x_1, \ldots, x_N \sim P(x)$ independent: $N^{-1} \sum_i f(x_i) \to E_P[f(x)]$ almost surely.
Central limit theorem: $P, f$ nice $\Rightarrow$ Convergence as $1/\sqrt{N}$ independent of $x$ dimensionality.
Catch: Sampling from $P(x)$ hard as well

Let's just do something: Start with some $x$, draw $x' \sim K(x'|x)$, keep doing that. At the very least:

$$P(x') = \int K(x'|x) P(x) \, dx$$

Such kernels $K$ exist, need evaluation of $\propto P(x)$ only
Why Approximate Inference?

Markov Chain Monte Carlo

\[ P(x') = \int K(x'|x)P(x) \, dx \]

**MCMC magic:** Under mild assumptions, that’s all we need:

\[ x^{(j+1)} \sim K(\cdot | x^{(j)}) \Rightarrow \text{Marginal } x^{(j)} \xrightarrow{D} P(x) \text{ as } j \to \infty \]

Rough idea why:

- \( K(x'|x) \) contraction of probability mass. Information propagation with \( K \) brings marginal distributions closer together
- There is only one fixed point (here: mild assumptions)
Why Approximate Inference?

Markov Chain Monte Carlo

- MCMC used for many 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!

Dare to find out for yourself?

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

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

\[ H[P(x)] = E_P[- \log P(x)] = - \sum_x P(x) \log P(x) \]

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

Shannon: On average: \( \leq H[P(x)] + 1 \) questions if you’re smart, no less than \( H[P(x)] \) even for a genius (log to base 2) ⇒ Equivalent: Number bits needed to encode \( x \) ⇒ Amount of uncertainty in \( P(x) \).
**Entropy of Distribution**

\[
H[P(x)] = EP[-\log P(x)] = - \sum_x P(x) \log P(x)
\]

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

- Shannon: On average: \( \leq H[P(x)] + 1 \) questions if you’re smart, no less than \( H[P(x)] \) even for a genius (log to base 2)

\( \Rightarrow \) Equivalent: Number bits needed to encode \( x \)

\( \Rightarrow \) Amount of uncertainty in \( P(x) \)
Entropy of Distribution

\[ H[P(x)] = E_P[- \log P(x)] = - \sum_x P(x) \log P(x) \]

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

- Shannon: On average: \( \leq H[P(x)] + 1 \) questions if you’re smart, no less than \( H[P(x)] \) even for a genius (log to base 2)
  \( \Rightarrow \) Equivalent: Number bits needed to encode \( x \)
  \( \Rightarrow \) **Amount of uncertainty** in \( P(x) \)

Joint entropy  \( H[P(y, x)] = E_P[- \log P(y, x)] \)

Conditional entropy  \( H[P(y|x)] = E_P[- \log P(y|x)] \)
Entropy of Distribution

\[ H[P(x)] = E_P[- \log P(x)] = - \sum_x P(x) \log P(x) \]

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

Joint entropy
\[ H[P(y, x)] = E_P[- \log P(y, x)] \]

Conditional entropy
\[ H[P(y|x)] = E_P[- \log P(y|x)] \]

Chain rule of entropy:
\[ H[P(x_1, \ldots, x_n)] = \sum_{i=1}^n H[P(x_i|x_{<i})] \]
Relative Entropy

\[ D[P(x) \parallel Q(x)] = E_P \left[ \log \frac{P(x)}{Q(x)} \right] \]

- Game of questions. This time, you get it wrong. You think \( x \sim Q(x) \), but in fact \( x \sim P(x) \). How many questions?
Relative Entropy

$$D[P(x) \parallel Q(x)] = E_P \left[ \log \frac{P(x)}{Q(x)} \right]$$

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

- On average: $E_P[-\log Q(x)] = H[P(x)] + D[P(x) \parallel Q(x)]$
  - Number of additional bits for using $Q$ instead of true $P$
  - Natural divergence (distance) measure between distributions

Other name: Kullback-Leibler divergence.

No distance: $D[P \parallel Q] \neq D[Q \parallel P]$
Relative Entropy

\[ D[P(x) \parallel Q(x)] = E_P \left[ \log \frac{P(x)}{Q(x)} \right] \]

- Game of questions. This time, you get it wrong. You think \( x \sim Q(x) \), but in fact \( x \sim P(x) \). How many questions?
- On average: \( E_P[− \log Q(x)] = H[P(x)] + D[P(x) \parallel Q(x)] \)  
  ⇒ Number of additional bits for using \( Q \) instead of true \( P \)  
  ⇒ Natural divergence (distance) measure between distributions

- Other name: Kullback-Leibler divergence.  
  No distance: \( D[P \parallel Q] \neq D[Q \parallel P] \)
Relative Entropy

\[ D[P(x) \parallel Q(x)] = \mathbb{E}_P \left[ \log \frac{P(x)}{Q(x)} \right] \]

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

- On average: \( \mathbb{E}_P[-\log Q(x)] = H[P(x)] + D[P(x) \parallel Q(x)] \)
  \( \Rightarrow \) 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: \( D[P \parallel Q] \neq D[Q \parallel P] \)

Conditional relative entropy:
\[ D[P(y|x) \parallel Q(y|x)] = \mathbb{E}_P[\log \{P(y|x)/Q(y|x)\}] \]

Chain rule of relative entropy:
\[ D[P(y, x) \parallel Q(y, x)] = D[P(y|x) \parallel Q(y|x)] + D[P(x) \parallel Q(x)] \]
Mutual Information

\[ I(x; y) = D[P(x, y) \parallel P(x)P(y)] = \mathbb{E}_P \left[ \log \frac{P(x, y)}{P(x)P(y)} \right] \]

- \( x, y \) may be dependent. How many additional questions (bits) for ignoring that?
Mutual Information

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

- \( x, y \) 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

\[ I(x; y) = H[P(x)] - H[P(x \mid y)] = H[P(y)] - H[P(y \mid x)] \]

⇒ Amount of information \( x \) about \( y \), or \( y \) about \( x \)
Some Information Theory

Mutual Information

\[ I(x; y) = D[P(x, y) \parallel P(x)P(y)] = E_P \left[ \log \frac{P(x, y)}{P(x)P(y)} \right] \]

- \( x, y \) 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

\[ I(x; y) = H[P(x)] - H[P(x | y)] = H[P(y)] - H[P(y | x)] \]

\( \Rightarrow \) Amount of information \( x \) about \( y \), or \( y \) about \( x \)
- Note: \( x \perp y \) (independent) \( \Rightarrow P(x, y) = P(x)P(y) \Rightarrow I(x; y) = 0. \) We’ll see \( \Leftrightarrow \). Mutual information: Measure of dependence
Venn Diagram for Information

\[ H(X,Y) \]

\[ H(X|Y) \quad I(X;Y) \quad H(Y|X) \]

\[ H(X) \]

\[ H(Y) \]
Some Information Theory

Information Inequality

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

Is \( D[P(x) \parallel Q(x)] \geq 0? \) Is \( I(x; y) \geq 0? \)
Some Information Theory

Information Inequality

- Something missing here
  - More questions for getting it wrong:
    \[ H[P(x)] \rightarrow H[P(x)] + D[P(x) \parallel Q(x)] \]
  - \( I(x; y) \) measures dependence. \( I(x; y) = 0 \) for \( x \perp y \)
  - Is \( D[P(x) \parallel Q(x)] \geq 0? \) Is \( I(x; y) \geq 0? \)

- Convexity comes to the rescue. Jensen’s inequality

\[ E_P[f(x)] \geq f(E_P[x]), \quad f \text{ convex} \]
Information Inequality

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

Is \( D[P(x) \parallel Q(x)] \geq 0? \) Is \( I(x; y) \geq 0? \)

- Convexity comes to the rescue. Jensen’s inequality

\[ E_P[f(x)] \geq f(E_P[x]), \quad f \text{ convex} \]

- Information inequality: \( D[P(x) \parallel Q(x)] \geq 0. \)
  Since \( -\log(\cdot) \) strictly convex (nowhere linear):
  \( D[P(x) \parallel Q(x)] = 0 \iff P(x) = Q(x) \text{ } P\text{-almost everywhere}. \)

\[ I(x; y) \geq 0; \quad I(x; y) = 0 \iff x \perp y \]
Corollaries

Raking in the fruits

- Conditioning reduces entropy (learning always helps)

\[ H[P(x|y)] \leq H[P(x)] \]
Corollaries

Raking in the fruits

- Conditioning reduces entropy (learning always helps)
  \[ H[P(x|y)] \leq H[P(x)] \]

- Conditional mutual information: Measure for conditional independence
  \[ I(x; y|z) = 0 \iff x \perp y | z \]
Corollaries

Raking in the fruits

- Conditioning reduces entropy (learning always helps)
  \[ H[P(x|y)] \leq H[P(x)] \]

- Conditional mutual information:
  Measure for conditional independence
  \[ I(x; y|z) = 0 \iff x \perp y | z \]

- Entropy: Concave function
  \[ H[\lambda P(x) + (1 - \lambda)Q(x)] \geq \lambda H[P(x)] + (1 - \lambda)H[Q(x)] \]
Remember EM?

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

Remember derivation of EM?
Remember EM?

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

Remember derivation of EM?

$$\log Z = \log \int e^{\psi(x)} \, dx = \sup_Q E_Q[\log\{e^{\psi(x)}/Q(x)\}]$$

$$= \sup_Q \{E_Q[\psi(x)] + H[Q(x)]\}$$

Was called variational mean field inequality. Let’s see why
Remember EM?

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

Remember derivation of EM?

\[
\log Z = \log \int e^{\psi(x)} \, dx = \sup_Q E_Q[\log\{e^{\psi(x)}/Q(x)\}] \\
= \sup_Q \{E_Q[\psi(x)] + H[Q(x)]\}
\]

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

- Maximizer: $Q(x) = P(x)$ itself. Attains log $Z$
- Any other $Q(x)$: Lower bound. $Q(x)$ closer to $P(x)$?
  ⇒ Maximize the lower bound!
Remember EM?

One approach to variational approximate inference:

Computations with $P(x) = Z^{-1} e^{\psi(x)}$ hard (even log $Z$)?

⇒ Approximate it by $Q(x)$, for which computations simple

Remember derivation of EM?

\[
\log Z = \log \int e^{\psi(x)} \, dx = \sup_Q E_Q[\log \{e^{\psi(x)}/Q(x)\}]
\]

\[
= \sup_Q \{E_Q[\psi(x)] + H[Q(x)]\}
\]

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

- Maximizer: $Q(x) = P(x)$ itself. Attains log $Z$
- Any other $Q(x)$: Lower bound. $Q(x)$ closer to $P(x)$?
  ⇒ Maximize the lower bound!
- Relax this problem: Work with $Q = \{Q(x)\}$:
  - Lower bound can be evaluated for each $Q \in Q$
  - Bayesian computations can be done with any $Q \in Q$ (not with $P$)
Distributions complicated, because they are coupled

⇒ Mean field: Approximate them by factorizing distributions

Naive Mean Field: Drop all edges

True MRF posterior $P(x)$

Approximations $Q(x) \in Q$
Variational problem:

$$\arg\max_{\{Q(x_k)\}} \left\{ \sum_j E_{Q}[\psi_j(x_{C_j})] + \sum_k H[Q(x_k)] \right\}$$

Our first variational algorithm:

- Default-initialize $Q(x_k)$ (say: uniform)
- repeat
  - Pick some node $k$ at random
  - Update $Q(x_k)$, keeping all others fixed
    $$Q(x_k) \leftarrow \arg\max \left\{ \sum_{j \in N_k} E_{Q}[\psi_j(x_{C_j})] + H[Q(x_k)] \right\}$$
- until Convergence

Prize question: How does that update look like?
Remarks

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

Why "mean field"? 
\[ P(x) \] Random field. 
\( Q(x) \approx P(x) \) has no couplings (\( E_Q[x_j x_k] = E_Q[x_j] E_Q[x_k] \)). 

True means at convergence (\( E_Q[x_j] = E_P[x_j] \))? No (remember: \( E_P[x_j] \) hard as well!)

General idea here: Relax variational problem 
\[ \sup_Q \ldots \geq \sup_{Q \in Q} \ldots \]

\( Q \): Subset of all distributions (factorization constraints). Each \( Q(x) \) is distribution.

\( \Rightarrow \) Maximize lower bound over \( Q \)

Note: Might not find maximizer \( Q \in Q \), but local maximum
Remarks

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

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

Variational Mean Field: Minimizing Relative Entropy

\[ \log Z = \log \int e^{\psi(x)} \, dx \geq E_Q[\psi(x)] + H[Q(x)], \quad P(x) = Z^{-1} e^{\psi(x)} \]

What is the slack in this bound?

Hint: \( \geq 0 \), and \( = 0 \) iff \( Q(x) = P(x) \)
Variational Mean Field: Minimizing Relative Entropy

$$\log Z = \log \int e^{\psi(x)} \, dx \geq E_Q[\psi(x)] + H[Q(x)], \quad P(x) = Z^{-1} e^{\psi(x)}$$

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

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

Does that fit relative entropy semantics?
Variational Mean Field: Minimizing Relative Entropy

\[ \log Z = \log \int e^{\psi(x)} \, dx \geq E_Q[\psi(x)] + H[Q(x)], \quad P(x) = Z^{-1} e^{\psi(x)} \]

- What is the slack in this bound?
  - Hint: \( \geq 0 \), and \( = 0 \) iff \( Q(x) = P(x) \)

- Variational mean field: Minimize slack (relative entropy)

\[ \min_{Q \in \mathcal{Q}} D[Q(x) \parallel P(x)] \]

Does that fit relative entropy semantics?

- It’s the wrong way around! We should minimize \( D[P \parallel Q] \).
  - Alas, even that is hard. For naive mean field, unique solution is

\[ Q(x_1, \ldots, x_N) = P(x_1) \ldots P(x_N) \]

Variational mean field: a **tractable compromise**
Variational Mean Field Approximations

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