Variational Methods
Information theory, pattern recognition, and neural networks

- Source coding (Data compression)
- Noisy-channel coding
- Inference + probabilistic methods
  - 9-10 Inference
  - 11 Clustering
  - 12 Monte Carlo methods
  - 13 Advanced Monte Carlo methods
  - 14 Variational methods
- Neural networks
- State-of-the-art error-correcting codes

www.inference.phy.cam.ac.uk/itprnn/
www.inference.phy.cam.ac.uk/itila/
Overview

- Data compression
- Noisy-channel coding
  - Chs 1-6, 8-10, 14
- Inference, data modelling
  - Clustering, pattern recognition
    - Chs 20, 22
- Probability toolbox
  - Monte Carlo methods
    - Ch 29
  - Variational methods
    - Ch 33
- Neural networks
  - Chs 38, 39, (& perhaps 41, 44), 42
- State-of-the-art error-correcting codes

Additional reading

- Laplace's method (Ch 27)
- Ising models (Ch 31)
The course
www.inference.phy.cam.ac.uk/itprnn/

The book
www.inference.phy.cam.ac.uk/itila/
Variational methods

Interested in $P(x) = \frac{1}{Z} P^*(x) = \frac{1}{Z} e^{-E(x)}$.

$E(x)$ is simple, but not simple enough.

**Idea**  Approximate $P(x)$ by a simpler distribution $Q(x; \theta)$.

Adjust $\theta$ to get the ‘best’ approximation.

Then approximate $\sum_x \phi(x) P(x)$ by $\sum_x \phi(x) Q(x; \theta^*)$

How to measure 'best'?  Possible ideas:

$$D_{KL}(Q||P) = \sum_x Q(x) \log \frac{Q(x)}{P(x)}$$

$$D_{KL}(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$
Distances between probability distributions

Q:

If \[ P(x) = \{1/3, 1/3, 1/3, \epsilon\} \]
and \[ Q(x) = \{1/4, 1/4, 1/4, 1/4\}, \]

which is bigger:

\[
D_{KL}(Q \| P) = \sum_x Q(x) \log \frac{Q(x)}{P(x)}
\]

or

\[
D_{KL}(P \| Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}
\]
Example

\[ \sum P \log \frac{P}{Q} = \]

\[ \sum Q \log \frac{Q}{P} = \]
\[ D(Q \parallel P) \geq \sqrt{D(P \parallel Q)} \leq D(P \parallel Q) \]
$\sum P \log \frac{P}{Q} = \log \frac{4}{3}$

$\sum Q \log \frac{Q}{P} = \frac{3}{4} \log \frac{3}{4} + \frac{1}{4} \log \frac{1}{4 \varepsilon}$

Bad news
\[ \frac{Q}{P} = \frac{3}{4} \log \frac{3}{x} + \frac{1}{4} \log \frac{1}{4} = \text{Bad news} \]
\[ \log \frac{3}{3} \]

\[ \frac{3}{4} \log \frac{3}{k} + \frac{1}{4} \log \frac{3}{4} \]

\[ \text{Big} \]

\[ \text{Bad news} \]
\[ \log \left( \frac{4}{3} \right) \]

\[ \sum p \log \frac{p}{\hat{q}} \rightarrow \text{A} \]

\[ \text{if } \sum q \log \frac{q}{p} \]

\[ \frac{3}{4} \log \left( \frac{3}{4} \right) + \frac{1}{4} \log \left( \frac{1}{4} \right) \]

\[ \text{BIG} \]

\[ \text{Bad news} \]
\[ x \sum P \frac{\log P}{Q} \text{ not on} \]
Variational methods

Interested in \( P(x) = \frac{1}{Z} P^*(x) = \frac{1}{Z} e^{-E(x)} \).

\( E(x) \) is simple, but not simple enough.

Idea  Approximate \( P(x) \) by a simpler distribution \( Q(x; \theta) \).

Adjust \( \theta \) to get the ‘best’ approximation.

Then approximate \( \sum_x \phi(x) P(x) \) by \( \sum_x \phi(x) Q(x; \theta^*) \)

Objective function: Variational free energy

\[
\tilde{F}(\theta) = \sum_x Q(x; \theta) E(x) - \sum_x Q(x; \theta) \ln \frac{1}{Q(x; \theta)}
\]

\( \tilde{F}(\theta) \) is lower-bounded by \(-\log Z\)

\[
\tilde{F}(\theta) = D_{KL}(Q || P) - \log Z
\]
\( \sum_x P \log \frac{P}{Q} \) not on

\( D_{KL}(Q \parallel P) = \sum_x Q \log \frac{Q}{P(x)} = \sum_x Q(x) E(x) \)

\( P(x) = \frac{e^{-E(x)}}{\mathcal{Z}} \)
\[ P(x) = \frac{e^{-E(x)}}{Z} \]

\[ = \sum_{x} q(x) E(x) + \log_e Z - H_{q(x)}(x) \]

\( \text{doesn't matter} \)
\[ \tilde{F}(\theta) = \sum \alpha \theta(x) E(x) - \sum x \phi(x) \ln \frac{1}{\tilde{q}(x)} \]
\[ \sum_Q \log \frac{Q(x)}{P(x)} \geq 0 \]

\[ P(x) = \frac{e^{-E(x)}}{Z} \]

\[ = \sum_Q \sum_x Q(x) E(x) \]
\[ \tilde{F}(\theta) = \sum_x q(x) E(x) - \sum_x q(x) \ln \frac{1}{Q(x)} \]

Variational free energy

\[ \tilde{F}(\theta) \geq -\log Z \]
Examples

Inferring mu and sigma

Two coupled spins
\[ N(\mu, \sigma^2) \rightarrow 3 \times \frac{\sum}{n} \]

\[ P = \text{Inference of } \mu, \sigma^2 \]
\[ Q(\mu, \sigma^2) = Q(\mu, \sigma^2; \mu, \sigma^2) \]
Approximate the spin system whose energy function is

\[ E(x; J) = -\frac{1}{2} \sum_{m,n} J_{mn} x_m x_n - \sum_n h_n x_n \]

with a separable distribution

\[ Q(x; a) = \frac{1}{Z_Q} \exp \left( \sum_n a_n x_n \right). \]

We optimize \( Q \) so as to minimize the variational free energy

\[ \beta \tilde{F}'(a) = \beta \sum_x Q(x; a) E(x; J) - \sum_x Q(x; a) \ln \frac{1}{Q(x; a)}, \]

\[ = \beta \left( -\frac{1}{2} \sum_{m,n} J_{mn} \bar{x}_m \bar{x}_n - \sum_n h_n \bar{x}_n \right) - \sum_n H_2^{(e)}(q_n). \]

- Minimum (with respect to \( a \)) can be found by the iterative equations

\[ a_m = \beta \left( \sum_n J_{mn} \bar{x}_n + h_m \right) \quad \text{and} \quad \bar{x}_n = \tanh(a_n) \]
\[ p(x) = e^{-\beta E(x; J)} \]
\[ \tilde{F}(\theta, \beta) = \sum_x \delta(x) V(x) - \sum_x \delta(x) \ln \frac{1}{Q_\theta(x)} \]

Variational free energy

\[ F(\theta) \geq -\log Z(\beta) \]
\[ E(x) = -\frac{1}{2} \sum_{mn} J_{mn} x_m x_n \]
\[ X \in \exists \, \xi^{-1}, \xi^{+1} \]

\[ \frac{1}{2} \sum_{mn} J_{mn} X_{mn} X_{nn} - \sum_n h_n x_n \]
\[ Q(x; a) = \frac{1}{\mathbb{Z}Q} \sum \nabla \phi \]
\[
= \prod_{n=1}^{N} Q_n(x_n; a_n)
\]
\[
= \frac{1}{\mathbb{Z}Q} \sum_{a_n, x_n} e^{\nabla \phi}
\]
\[ h_n x_n \]
\[
\frac{a_{n-1}}{a_{n+1}+1} = \frac{a_{n-1}}{2}
\]

\[
X_n = a_n + (-1)^n a_{n+1}
\]

\[
X_0 = \frac{a_0}{2} - (1-a)
\]

\[
= 2a - 1
\]

\[
\sum a(n+1) \prod \text{YEC}
\]

\[
F(\Theta, \beta) = \sum a(n+1) \prod \text{YEC}
\]
\[ \exists x_1, x_2 \in \{ \pm 1 \} \]

\[ E(x) = -x_1 x_2 \]
\[ x_1, x_2 \in \{ \pm 1 \} \]

\[ f(x) = -x_1 x_2 \]

\[ \beta \tilde{F}(a) = \beta \bar{x}_1 \bar{x}_2 - H_2^{(e)}(q_1) - H_2^{(e)}(q_2) \]

Diagram:

- Four points labeled as: \( x_1, -1 \), \( x_1, 1 \), \( -1, x_2 \), \( 1, x_2 \)

- Lines connecting the points

- A circle around the point \( (x_1, x_2) \) with label \( J = 1 \)
\[ \beta + (\alpha) = \beta \]

\[ Q(x_2) \]

\[ X_2 \]

\[ Q(x_1, x_2) \]
beta = 0.725
\( \beta = 0.4 \)

\( \beta = 2 \)
\{ x_1, x_2 \} \in \{ \pm 1 \}

E(x) = -x_1 x_2

J = 1

\beta \bar{F}(a) = \beta \bar{x}_1 \bar{x}_2 - H_{2e}
Bifurcation

\( q_1, q_2 = \frac{1}{2} \)

\( q^*, q^* \)

\( 0.4 \)

\( 1 \)

\( 2 \)

\( \beta \)
Have you seen this before?

\[ a_m = \beta \left( \sum_n J_{mn} \bar{x}_n + h_m \right) \quad \text{and} \quad \bar{x}_n = \tanh(a_n) \]

or equivalently...

\[ \bar{x}_m = \tanh \left( \beta h_m + \beta \sum_n J_{mn} \bar{x}_n \right) \]
\[ \bar{x}_m = \tanh \left( \beta h_m + \beta \sum_n J_{mn} \bar{x}_n \right) \]

'Mean field theory' is a variational method
Mean Field Theory
Curie-Weiss
Feynman trick
Bogoliubov
VFE view of MFT

1. Clear objective function

2. 
Overview of MFT

1. Clear objective function $\rightarrow$ derivation
   - Could choose more complex $O$ $\rightarrow$ generalize
   - Show that MFT gives a bound on $Z$