Variational Inference: a Short Introduction[1]

Dmitry Adamskiy, David Barber

University College London

[1] These slides accompany the book *Bayesian Reasoning and Machine Learning*. The book and demos can be downloaded from `www.cs.ucl.ac.uk/staff/D.Barber/brml`. Feedback and corrections are also available on the site. Feel free to adapt these slides for your own purposes, but please include a link the above website.

Overview

Problem Setting

KL-divergence

Mean Field Approximation

Coordinate Ascent

Example: Spin Systems

Toy Example: Super Simple Spin System

Variational Generative Models a.k.a. “Variational autoencoders”

Problem Setting

- Sometimes we have to deal with (reasonably) complex probability distributions: for example, we may know them up to normalising constant.
  - EM-like inference: posterior $p(h|v,\theta)$ could be nasty, but joint $p(h,v|\theta)$ is fine.
  - In undirected models, $p = \frac{1}{Z}e^{-E(x)}$, we know energy $E(x)$, but normalization constant is a problem.
  - ...
- Two approaches: deterministic appoximations (today) and sampling (next week).

VI: idea

Instead of dealing with the difficult distribution p we are going to find the best approximation to it in a class of simpler distributions q. This raises a number of questions...

- What does it mean to be a best approximation?
- Which family of distributions to choose?
- How to find it?

KL-divergence

A natural way to measure distance between distributions is to use KL-divergence:

$$\mathrm{KL}(q||p) = \sum_x q(x) \log \frac{q(x)}{p(x)}$$

Couple of toy examples to help us better understand it:

Example 1

Suppose $p = (1/3, 1/3, 1/3 - \varepsilon, \varepsilon)$ and $q = (1/4, 1/4, 1/4, 1/4)$. What is bigger, $\mathrm{KL}(p||q)$ or $\mathrm{KL}(q||p)$?

Example 2

Consider approximating bimodal distribution $p(x)$ with unimodal $q(x)$ using $\mathrm{KL}(p||q)$ and $\mathrm{KL}(q||p)$ as your measure of success. What would be the difference?

Practical considerations

Suppose $p(x) = \frac{e^{-E(x)}}{Z}$.

- Option 1: $\mathrm{KL}(p||q)$:

$$\mathrm{KL}(p||q) = \sum p(x) \log \frac{p(x)}{q(x)}$$

Computing expectations under $p(x)$ is hard, so that's not on. On the other hand...

- Option 2: $\mathrm{KL}(q||p)$

$$\mathrm{KL}(q||p) = \sum q(x) \log \frac{q(x)}{p(x)} = \sum q(x)E(x) + \log(Z) - H(q)$$

$\log(Z)$ is the hard bit, but that one does not depend on q, so it doesn't matter. In fact, $H(q) - \sum q(x)E(x)$ is a lower bound on it (since KL-divergence is non-negative).

Same in EM-language (this is from last lecture)

The key feature of the EM algorithm is to form an alternative objective function for which the parameter coupling effect is removed, meaning that individual parameter updates can be achieved, akin to the case of fully observed data. The way this works is to replace the marginal likelihood with a lower bound – it is this lower bound that has the decoupled form.
Consider the Kullback-Leibler divergence between a 'variational' distribution $q(h|v)$ and the parametric model $p(h|v,\theta)$:

$$\mathrm{KL}(q(h|v)||p(h|v,\theta)) \equiv \langle \log q(h|v) - \log p(h|v,\theta) \rangle_{q(h|v)} \geq 0$$

Using Bayes' rule, $p(h|v,\theta) = p(h,v|\theta)/p(v|\theta)$ and the fact that $p(v|\theta)$ does not depend on h,

$$\log p(v|\theta) \geq -\langle \log q(h|v) \rangle_{q(h|v)} + \langle \log p(h,v|\theta) \rangle_{q(h|v)}$$

Free Energy / ELBO

- So the main idea of Variational Inference is to find the best dsitribution q within certain family, that maximizes Free Energy (or as it sometimes called, Evidence Lower Bound (ELBO))
- Here it is in the language of Markov Net models ($p(x) = \frac{e^{-E(x)}}{Z}$):

$$F(q) = H(q) - \sum_x q(x)E(x) = \langle -E(x) \rangle_{q(x)} + H(q)$$

- And here it is for the case of learning with hidden variables (from last lecture):

$$F(q) = \langle \log p(v, h|\theta) \rangle_{q(h|\theta)} + H(q)$$

- Which family of to choose for q?

Mean Field Approximation

Mean Field Approximation

- A common approach to selecting a family of "easy" distributions Q is to select a factorized distribution:

$$Q(x) = \prod_{n=1}^{N} q(x_n)$$

  This is a simplification and typically this family does not contain true distribution of interest (for example, in the posterior it is often the dependencies that make it difficult).
- How to optimize the bound for this distribution? In principle, any optimization method you know!
- Here is one common approach...

Coordinate Ascent

Coordinate Ascent VI – 1

We can write the distribution p using chain rule for any ordering of variables $x_1, \ldots, x_n$:

$$p(x_{1:n}) = \prod_{j=1}^{n} p(x_j | x_{1:(j-1)}).$$

We could also decompose the entropy:

$$H(q) = \sum_{n=1}^{N} H(q_n)$$

This gives the following decomposition of the bounds:

$$F(q) = \sum_{j=1}^{n} \left\langle \log p(x_j | x_{1:(j-1)}) \right\rangle_q - \left\langle \log q(x_j) \right\rangle_{q_j}$$

Coordinate Ascent VI – 2

- Consider the bound as a function of $q(x_k)$:

$$F(q_k) = \langle E_{-k}[\log p(x_k|x_{-k})] \rangle_{q_k} + H(q_k) + \text{const.},$$

  where E_{-k} is expectation with respect to all $q_i(x_i)$, $i \neq k$.
- The solution to this is

$$q_k^*(x_k) \propto e^{E_{-k}[\log p(x_k|x_{-k})]}$$

  (to see that, observe that $F(q_k)$ is equal to $\mathrm{KL}(q_k||q_k^*)$ up to an additive constant).
- The Coordinate Ascent algorithm is then to iteratively update each q_k. This guaranteed convergence (of the lower bounds) to the local maximum.

Exercise: Bivariate Gaussian

Exercise

What would the mean field approximation to the following two-dimensional Gaussian distribution look like?

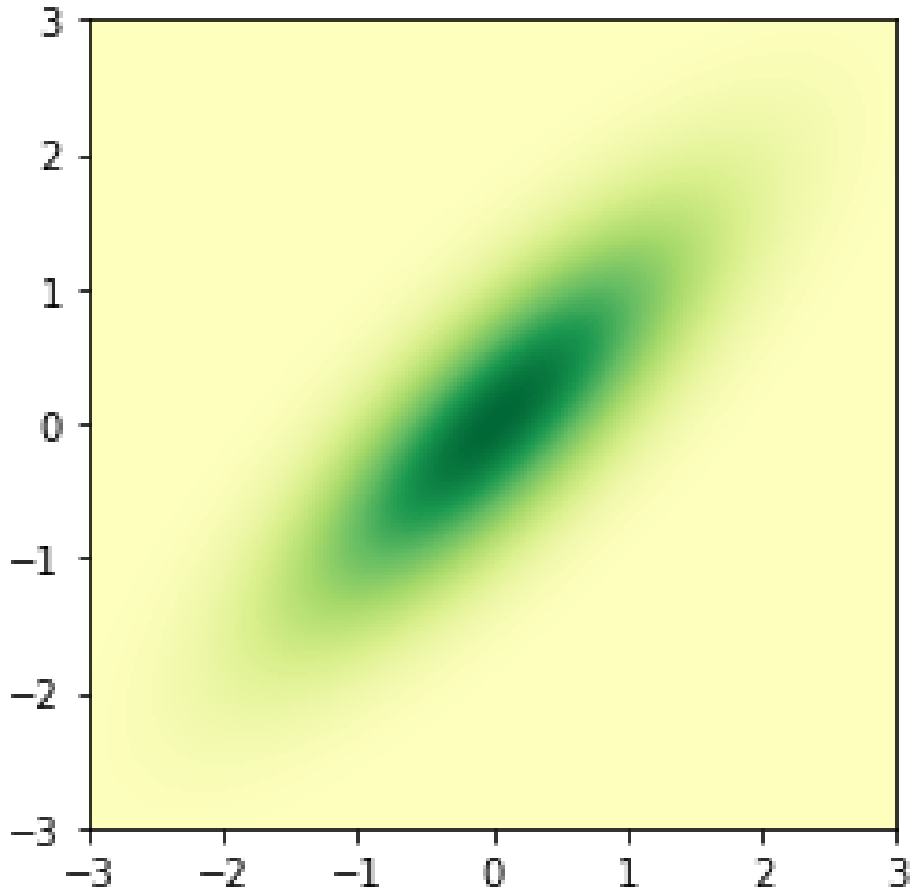


Figure: P(x)

Example: Spin Systems

Spin Systems

- Consider the Markov Net on binary variables with the distribution

$$P(x) = \frac{e^{-\beta E(x)}}{Z}$$

where $E(x|J,h) = -\frac{1}{2}\sum_{m,n} J_{mn}x_m x_n - \sum_n b_n x_n$.

- Mean Field Approximation: approximate P with a separable distribution

$$q(x|a) = \frac{1}{Z_q}\exp(\sum_n a_n x_n)$$

This means that the free energy becomes

$$-\beta \sum_x q(x|a)E(x|J,b) - \sum q(x|a)\log q(x|a)$$

which is further simplified to

$$\sum_n H(q_n) + \sum_{ij} J_{mn}\bar{x}_m\bar{x}_n + \sum_n b_n\bar{x}_n,$$

where $\bar{x}_n$ is an expectation of x_n under q_n.

CAVI updates for Mean Field Theory

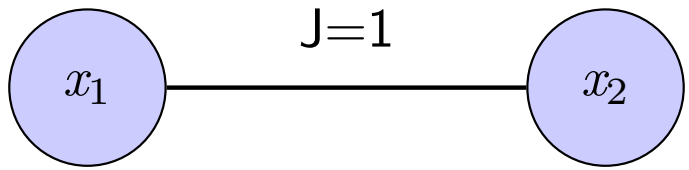
Minimum with respect to a is found by iterating these equations:

$$a_n = \beta(\sum_m J_{mn}\bar{x}_m + h_n)$$

and $\bar{x}_n = \tanh(a_n)$.
Let's derive them for the simplest possible spin system...

Toy Example: Super Simple Spin System

Simplest possible Markov Net / Spin System

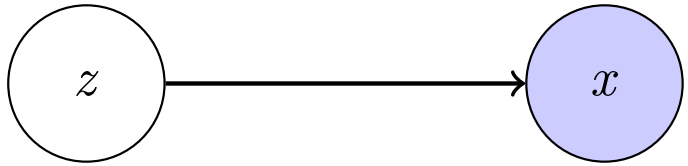


- Two binary variables $x_1, x_2 \in \{-1, +1\}$
- $P(x_1, x_2) = \frac{e^{\beta x_1 x_2}}{Z}$, β is a temperature (or rather coolness) parameter.
- What does this distribution look like?

Variational Generative Models a.k.a. “Variational autoencoders”

Variational Generative Models

- We want to build a generative model like this:



- Here z could be a hidden low-dimensional representation of a picture/video/sentence.
- The joint then is

$$p(x, z) = p(z)p_\theta(x|z),$$

where $p_\theta(x|z)$ is some rather complicated distribution (for example parametrised by neural network) where most parameters of the model live.

- We want to find ML-parameters, maximizing $\log p(x)$.
- Following the variational approach, we arrive at

$$\log p(x) \geq \int q_\phi(z|x) \log \frac{p(z)p_\theta(x|z)}{q_\phi(z|x)} dz$$

Here, family q_ϕ could be parametrised by another neural network.

Variational Generative Models

This is then approximated by taking samples from $q_\phi(z|x)$

$$\log p(x) \geq \int q_\phi(z|x) \log \frac{p(z)p_\theta(x|z)}{q_\phi(z|x)} dz \approx \frac{1}{S} \sum_{s=1}^{S} \log \frac{p(z^s)p_\theta(x|z^s)}{q_\phi(z^s|x)}$$

This then is iteratively optimized in θ and ϕ.