University of BRISTOL

Machine Learning

Deterministic Approximative Inference

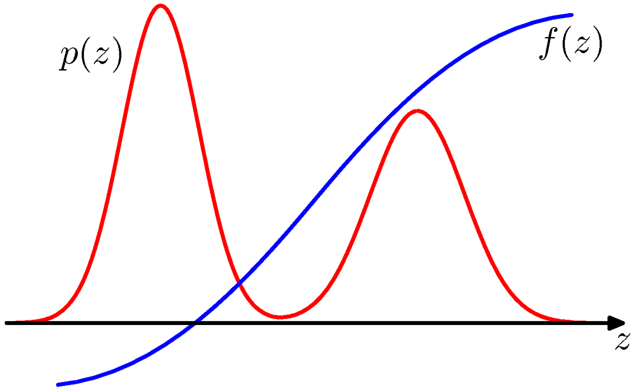
Carl Henrik Ek - carlhenrik.ek@bristol.ac.uk

November 7, 2017

http://www.carlhenrik.com

Introduction

Introduction Ch. 11.0 [1]



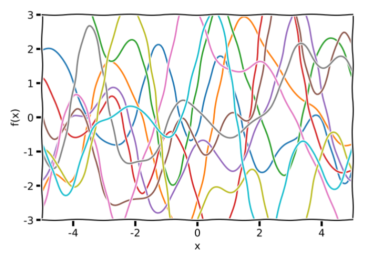
$$\mathbb{E}_{p(z)}[f] = \int f(z)p(z)\mathrm{d}z \approx \frac{1}{L}\sum_{l=1}^{L} f(z^{(l)})$$

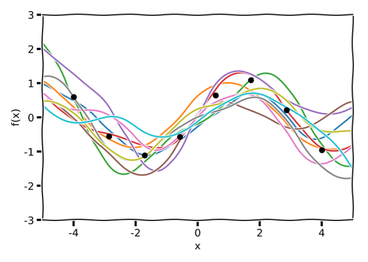
$$\mathbf{z}^{(l)} \sim p(\mathbf{z}), \qquad f(z) = p(x|z)$$

Big Number

$$\infty$$

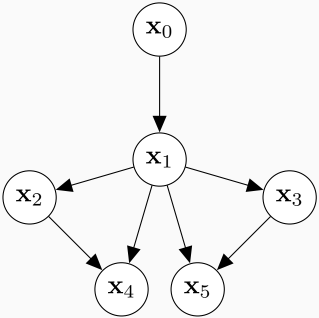
Big Number





$$p(\mathbf{y}|\mathbf{x}) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{x})\mathrm{d}\mathbf{f}$$

Ancestral Sampling

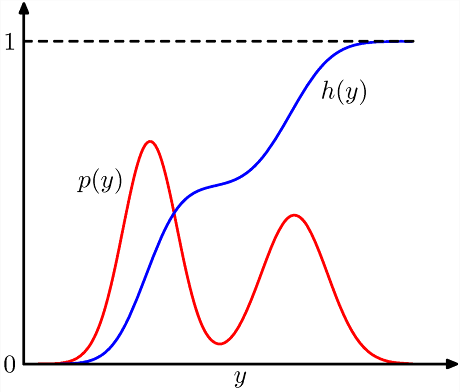


Sample from $p(\mathbf{x})$

1. pick top nodes and draw sample
2. fix the top nodes and sample from conditonals
3. arrive at sample from $\mathbf{x}$

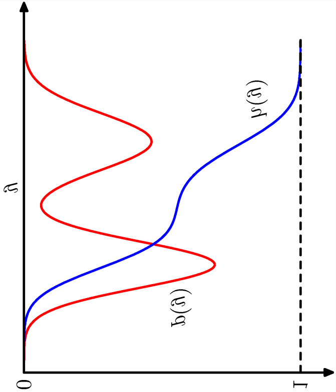
$$p(\mathbf{x}) = p(x_5|x_3, x_1)p(x_4|x_2, x_1)p(x_3|x_1)p(x_2|x_1)p(x_1|x_0)p(x_0)$$

Basic Probabilities



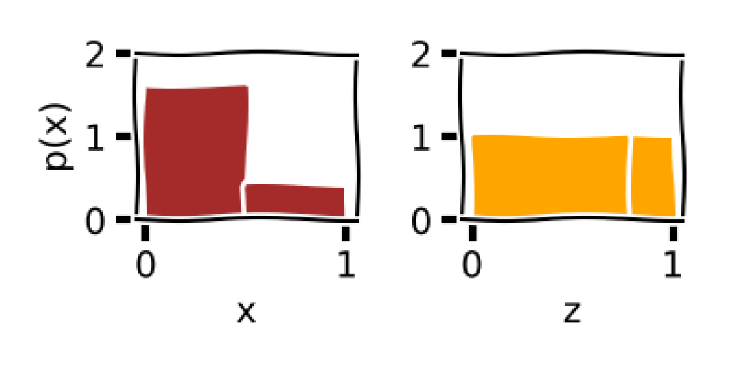
$$z = h(y) = \int_{-\infty}^{y} p(y)\mathrm{d}y$$

Basic Probabilities



$$y = h^{-1}(z)$$

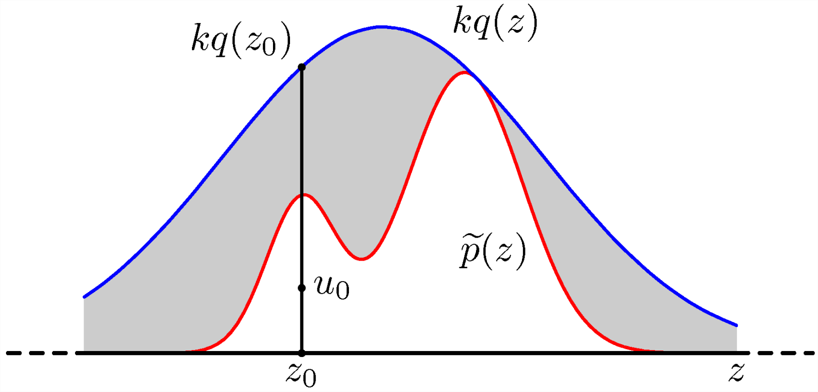
Change of Variables



Starting point

- We can sample random numbers from the uniform distribution
- We can using the indefinite integral transform the uniform
- Want to use these distributions as proxies

Rejection Sampling

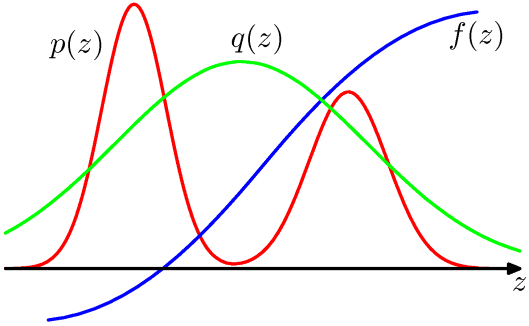


Importance Sampling Ch 11.1.4 [1]

$$\begin{aligned}
\mathbb{E}_{p(\mathbf{z})}[f] &= \int f(\mathbf{z})p(\mathbf{z})\mathrm{d}\mathbf{z} = \int f(\mathbf{z})\frac{p(\mathbf{z})}{q(\mathbf{z})}q(\mathbf{z})\mathrm{d}\mathbf{z} = \mathbb{E}_{q(\mathbf{z})}\left[f\frac{p(\mathbf{z})}{q(\mathbf{z})}\right] \\
&\approx \frac{1}{L}\sum_{l=1}^{L}\frac{p(z^{(l)})}{q(z^{(l)})}\cdot f(\mathbf{z}^{(l)})
\end{aligned}$$

- Sample from proposal distribution and re-weight samples
- Accepts all samples

Importance Sampling



$$r_l = \frac{p\left(z^{(l)}\right)}{q\left(z^{(l)}\right)}$$

Metropolis Sampling

1. start with state $\mathbf{z}^{(0)}$

Metropolis Sampling

1. start with state $\mathbf{z}^{(0)}$
2. sample from conditional proposal distribution $q(\mathbf{z}^*|\mathbf{z}^{(0)})$

Metropolis Sampling

1. start with state $\mathbf{z}^{(0)}$
2. sample from conditional proposal distribution $q(\mathbf{z}^*|\mathbf{z}^{(0)})$
3. compute acceptance probability

$$A(\mathbf{z}^*, \mathbf{z}^{(0)}) = \min\left(1, \frac{\tilde{p}(\mathbf{z}^*)}{\tilde{p}(\mathbf{z}^{(0)})}\right)$$

Metropolis Sampling

1. start with state $\mathbf{z}^{(0)}$
2. sample from conditional proposal distribution $q(\mathbf{z}^*|\mathbf{z}^{(0)})$
3. compute acceptance probability

$$A(\mathbf{z}^*, \mathbf{z}^{(0)}) = \min\left(1, \frac{\tilde{p}(\mathbf{z}^*)}{\tilde{p}(\mathbf{z}^{(0)})}\right)$$

4. Draw uniform random number $u \sim \text{Uniform}(0, 1)$

Metropolis Sampling

1. start with state $\mathbf{z}^{(0)}$
2. sample from conditional proposal distribution $q(\mathbf{z}^*|\mathbf{z}^{(0)})$
3. compute acceptance probability

$$A(\mathbf{z}^*, \mathbf{z}^{(0)}) = \min\left(1, \frac{\tilde{p}(\mathbf{z}^*)}{\tilde{p}(\mathbf{z}^{(0)})}\right)$$

4. Draw uniform random number $u \sim \text{Uniform}(0, 1)$
    - if $A(\mathbf{z}^*, \mathbf{z}^{(0)}) > u \rightarrow \mathbf{z}^{(1)} = \mathbf{z}^*$

Metropolis Sampling

1. start with state $\mathbf{z}^{(0)}$
2. sample from conditional proposal distribution $q(\mathbf{z}^*|\mathbf{z}^{(0)})$
3. compute acceptance probability

$$A(\mathbf{z}^*, \mathbf{z}^{(0)}) = \min\left(1, \frac{\tilde{p}(\mathbf{z}^*)}{\tilde{p}(\mathbf{z}^{(0)})}\right)$$

4. Draw uniform random number $u \sim \text{Uniform}(0,1)$
   - if $A(\mathbf{z}^*, \mathbf{z}^{(0)}) > u \rightarrow \mathbf{z}^{(1)} = \mathbf{z}^*$
   - otherwise reject $\mathbf{z}^*$ and start over

Gibbs Sampling [1]

1. Initialise $\mathbf{z}^{(0)}$

Gibbs Sampling [1]

1. Initialise $\mathbf{z}^{(0)}$
2. Pick single variable $z_i \in \mathbf{z}$

Gibbs Sampling [1]

1. Initialise $\mathbf{z}^{(0)}$
2. Pick single variable $z_i \in \mathbf{z}$
3. Formulate posterior $p(z_i|\mathbf{z}_{\neg i})$

Gibbs Sampling [1]

1. Initialise $\mathbf{z}^{(0)}$
2. Pick single variable $z_i \in \mathbf{z}$
3. Formulate posterior $p(z_i|\mathbf{z}_{\neg i})$
4. Sample from posterior

$$z_i^{(1)} \sim p(z_i|\mathbf{z}_{\neg i})$$

Gibbs Sampling [1]

1. Initialise $\mathbf{z}^{(0)}$
2. Pick single variable $z_i \in \mathbf{z}$
3. Formulate posterior $p(z_i|\mathbf{z}_{\neg i})$
4. Sample from posterior

$$z_i^{(1)} \sim p(z_i|\mathbf{z}_{\neg i})$$

5. cycle through variables

Deterministic Approximations

Introduction

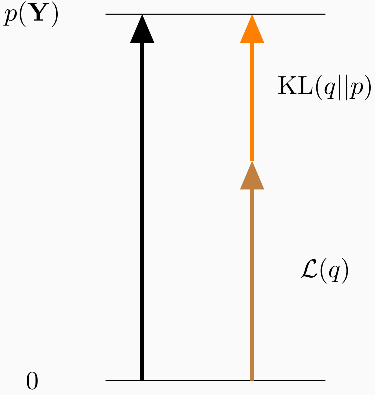
- Stochastic inference
  - approximate expectation with sum
  - works in the limit
  - hard to know how well we are doing
  - usually slow
- Idea
  - *can we reformulate inference as optimisation?*

Learning

$$p(\mathbf{Y})$$

- Given some observed data $\mathbf{Y}$
- Find a probabilistic model such that the probability of the data is maximised
- Idea: find an approximate model q that we can integrate

Deterministic Approximation



Variational Bound

$$p(\mathbf{Y})$$

Variational Bound

$$\log p(\mathbf{Y})$$

Variational Bound

$$\log p(\mathbf{Y}) = \log \int p(\mathbf{Y}, \mathbf{X}) \mathrm{d}\mathbf{X}$$

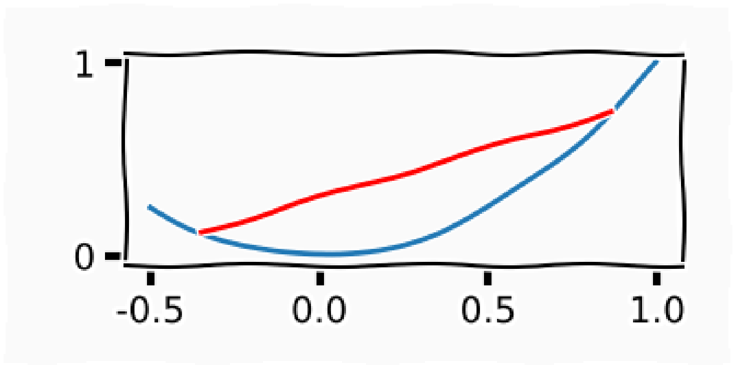
Variational Bound

$$\log\, p(\mathbf{Y}) = \log \int p(\mathbf{Y}, \mathbf{X}) \mathrm{d}\mathbf{X} = \log \int p(\mathbf{X}|\mathbf{Y}) p(\mathbf{Y}) \mathrm{d}\mathbf{X}$$

Variational Bound

$$\begin{aligned}\log\, p(\mathbf{Y}) &= \log \int p(\mathbf{Y}, \mathbf{X}) \mathrm{d}\mathbf{X} = \log \int p(\mathbf{X}|\mathbf{Y}) p(\mathbf{Y}) \mathrm{d}\mathbf{X} \\ &= \log \int \frac{q(\mathbf{X})}{q(\mathbf{X})} p(\mathbf{X}|\mathbf{Y}) p(\mathbf{Y}) \mathrm{d}\mathbf{X}\end{aligned}$$

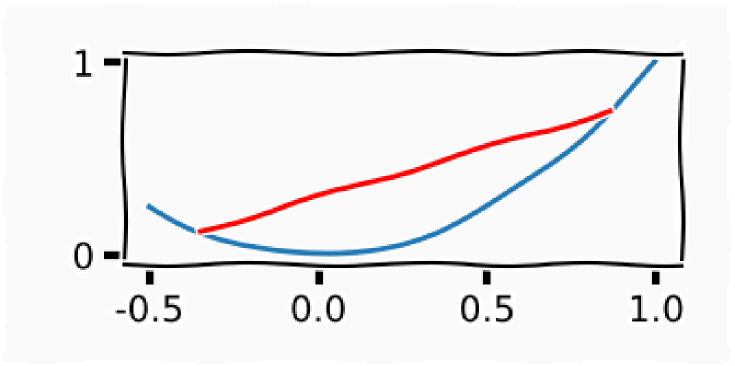
Jensen Inequality



Convex Function

$$\lambda f(x_0) + (1-\lambda) f(x_1) \geq f(\lambda x_0 + (1-\lambda) x_1)$$
$$x \in [x_{min}, x_{max}]$$
$$\lambda \in [0, 1]]$$

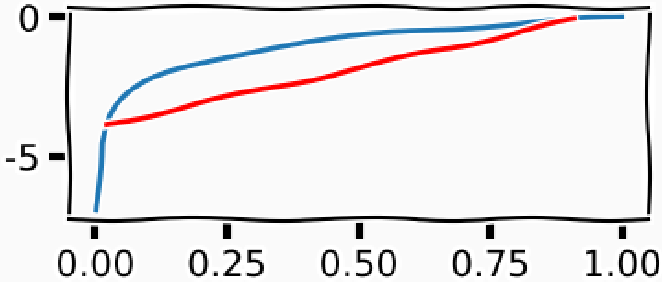
Jensen Inequality



$$\mathbb{E}[f(x)] \geq f(\mathbb{E}[x])$$

$$\int f(x)p(x)\mathrm{d}x \geq f\left(\int xp(x)\mathrm{d}x\right)$$

Jensen Inequality in Variational Bayes



$$\int \log(x)p(x)\mathrm{d}x \leq \log\left(\int xp(x)\mathrm{d}x\right)$$

moving the log inside the the integral is a lower-bound on the integral

Variational Bayes cont.

$$\log p(\mathbf{Y}) = \log \int \frac{q(\mathbf{X})}{q(\mathbf{X})} p(\mathbf{X}|\mathbf{Y}) p(\mathbf{Y}) \mathrm{d}\mathbf{X}$$

Variational Bayes cont.

$$\begin{aligned}\log p(\mathbf{Y}) &= \log \int \frac{q(\mathbf{X})}{q(\mathbf{X})} p(\mathbf{X}|\mathbf{Y}) p(\mathbf{Y}) \mathrm{d}\mathbf{X} \\ &\geq \int q(\mathbf{X}) \log \frac{p(\mathbf{X}|\mathbf{Y}) p(\mathbf{Y})}{q(\mathbf{X})} \mathrm{d}\mathbf{X}\end{aligned}$$

Variational Bayes cont.

$$
\begin{aligned}
\log p(\mathbf{Y}) &= \log \int \frac{q(\mathbf{X})}{q(\mathbf{X})} p(\mathbf{X}|\mathbf{Y}) p(\mathbf{Y}) \mathrm{d}\mathbf{X} \\
&\geq \int q(\mathbf{X}) \log \frac{p(\mathbf{X}|\mathbf{Y}) p(\mathbf{Y})}{q(\mathbf{X})} \mathrm{d}\mathbf{X} \\
&= \int q(\mathbf{X}) \log \frac{p(\mathbf{X}|\mathbf{Y})}{q(\mathbf{X})} \mathrm{d}\mathbf{X} + \int q(\mathbf{X}) \mathrm{d}\mathbf{X} \log\, p(\mathbf{Y})
\end{aligned}
$$

Variational Bayes cont.

$$
\begin{aligned}
\log p(\mathbf{Y}) &= \log \int \frac{q(\mathbf{X})}{q(\mathbf{X})} p(\mathbf{X}|\mathbf{Y}) p(\mathbf{Y}) \mathrm{d}\mathbf{X} \\
&\geq \int q(\mathbf{X}) \log \frac{p(\mathbf{X}|\mathbf{Y}) p(\mathbf{Y})}{q(\mathbf{X})} \mathrm{d}\mathbf{X} \\
&= \int q(\mathbf{X}) \log \frac{p(\mathbf{X}|\mathbf{Y})}{q(\mathbf{X})} \mathrm{d}\mathbf{X} + \int q(\mathbf{X}) \mathrm{d}\mathbf{X} \log\, p(\mathbf{Y}) \\
&= -\mathrm{KL}\left(q(\mathbf{X}) || p(\mathbf{X}|\mathbf{Y})\right) + \log\, p(\mathbf{Y})
\end{aligned}
$$

Kullback-Leibler Divergence

$$\mathrm{KL}\left(q(\mathbf{X})||p(\mathbf{X}|\mathbf{Y})\right) = \int q(\mathbf{X}) \log \frac{q(\mathbf{X})}{p(\mathbf{X}|\mathbf{Y})} \mathrm{d}\mathbf{X}$$

- Divergence measure between distributions
- Relative Shannon Entropy
- Not a metric, (not symmetric), 0 only if $p = q$, strictly positive
- $\mathrm{KL}(p(\mathbf{X}|\mathbf{Y})||p(\mathbf{X})$ information gain

Variational Bayes cont.

$$\log p(\mathbf{Y}) \geq -\mathrm{KL}\left(q(\mathbf{X}) || p(\mathbf{X}|\mathbf{Y})\right) + \log\, p(\mathbf{Y})$$

- if $q(\mathbf{X})$ is the true posterior we have an equality, therefore match the distributions
- i.e. $\mathrm{argmin}_q\ \mathrm{KL}\left(q(\mathbf{X}) || p(\mathbf{X}|\mathbf{Y})\right)$

  $\Rightarrow$ variational distributions are approximations to intractable posteriors

ELBO

$$\mathrm{KL}(q(\mathbf{X})||p(\mathbf{X}|\mathbf{Y}))$$

ELBO

$$\mathrm{KL}(q(\mathsf{X})||p(\mathsf{X}|\mathsf{Y})) = \int q(\mathsf{X}) log \frac{q(\mathsf{X})}{p(\mathsf{X}|\mathsf{Y})} \mathrm{d}\mathsf{X}$$

ELBO

$$
\begin{aligned}
\mathrm{KL}(q(\mathbf{X})||p(\mathbf{X}|\mathbf{Y})) &= \int q(\mathbf{X}) log \frac{q(\mathbf{X})}{p(\mathbf{X}|\mathbf{Y})} \mathrm{d}\mathbf{X} \\
&= \int q(\mathbf{X}) log \frac{q(\mathbf{X})}{p(\mathbf{X},\mathbf{Y})} \mathrm{d}\mathbf{X} + log \;\; p(\mathbf{Y})
\end{aligned}
$$

ELBO

$$
\begin{aligned}
\mathrm{KL}(q(\mathbf{X})||p(\mathbf{X}|\mathbf{Y})) &= \int q(\mathbf{X}) log \frac{q(\mathbf{X})}{p(\mathbf{X}|\mathbf{Y})} \mathrm{d}\mathbf{X} \\
&= \int q(\mathbf{X}) log \frac{q(\mathbf{X})}{p(\mathbf{X},\mathbf{Y})} \mathrm{d}\mathbf{X} + log \ p(\mathbf{Y}) \\
&= \int q(\mathbf{X}) log \ q(\mathbf{x}) \mathrm{d}\mathbf{X} - \int q(\mathbf{x}) \log \ p(\mathbf{X},\mathbf{Y}) \mathrm{d}\mathbf{X} + log \ p(\mathbf{Y})
\end{aligned}
$$

ELBO

$$
\begin{aligned}
\mathrm{KL}(q(\mathbf{X})||p(\mathbf{X}|\mathbf{Y})) &= \int q(\mathbf{X}) log \frac{q(\mathbf{X})}{p(\mathbf{X}|\mathbf{Y})} \mathrm{d}\mathbf{X} \\
&= \int q(\mathbf{X}) log \frac{q(\mathbf{X})}{p(\mathbf{X},\mathbf{Y})} \mathrm{d}\mathbf{X} + log \;\; p(\mathbf{Y}) \\
&= \int q(\mathbf{X}) log \;\; q(\mathbf{x}) \mathrm{d}\mathbf{X} - \int q(\mathbf{x}) \log \;\; p(\mathbf{X},\mathbf{Y}) \mathrm{d}\mathbf{X} + log \;\; p(\mathbf{Y}) \\
&= H(q(\mathbf{X})) - \mathbb{E}_{q(\mathbf{X})} \left[log \;\; p(\mathbf{X},\mathbf{Y}) \right] + log \;\; p(\mathbf{Y})
\end{aligned}
$$

ELBO

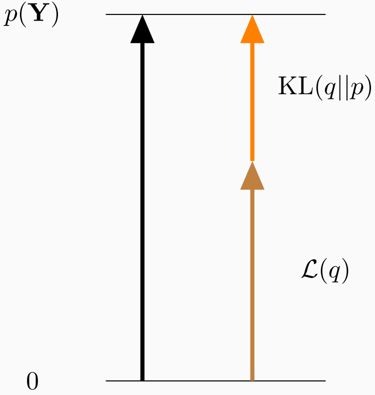
$$\mathit{log}\ \ p(\mathbf{Y}) = \mathrm{KL}(q(\mathbf{X})||p(\mathbf{X}|\mathbf{Y})) + \underbrace{\mathbb{E}_{q(\mathbf{X})}\left[\mathit{log}\ \ p(\mathbf{X},\mathbf{Y})\right] - H(q(\mathbf{X}))}_{\mathrm{ELBO}}$$

ELBO

$$log\ \ p(\mathbf{Y}) = \mathrm{KL}(q(\mathbf{X})||p(\mathbf{X}|\mathbf{Y})) + \underbrace{\mathbb{E}_{q(\mathbf{X})}\left[log\ \ p(\mathbf{X}, \mathbf{Y})\right] - H(q(\mathbf{X}))}_{\mathrm{ELBO}}$$

$$\geq \mathbb{E}_{q(\mathbf{X})}\left[log\ \ p(\mathbf{X}, \mathbf{Y})\right] - H(q(\mathbf{X})) = \mathcal{L}(q(\mathbf{X}))$$

Deterministic Approximation

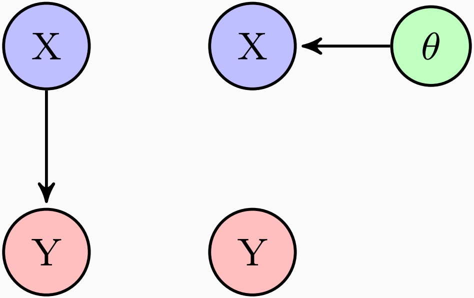


ELBO

$$log\ \ p(\mathbf{Y}) \geq \mathbb{E}_{q(\mathbf{X})}\left[log\ \ p(\mathbf{X},\mathbf{Y})\right] - H(q(\mathbf{X})) = \mathcal{L}(q(\mathbf{X}))$$

- if we maximise the ELBO we,
    - find an approximate posterior
    - get an approximation to the marginal likelihood
- *maximising* $p(\mathbf{Y})$ is learning
- finding $p(\mathbf{X}|\mathbf{Y}) \approx q(\mathbf{X})$ is prediction

ELBO



$$p(\mathbf{X}|\mathbf{Y}) \approx q(\mathbf{X}|\theta)$$

Why is this useful?

Why is this a sensible thing to do?

– Ryan Adams[1]

[1] `Talking Machines Season 2, Episode 5`

Why is this useful?

Why is this a sensible thing to do?

- If we can't formulate the joint distribution there isn't much we can do

– Ryan Adams[1]

[1] `Talking Machines Season 2, Episode 5`

Why is this useful?

Why is this a sensible thing to do?

- If we can't formulate the joint distribution there isn't much we can do
- Taking the expectation of a log is usually easier than the expectation

– Ryan Adams[1]

[1] Talking Machines Season 2, Episode 5

Why is this useful?

Why is this a sensible thing to do?

- If we can't formulate the joint distribution there isn't much we can do
- Taking the expectation of a log is usually easier than the expectation
- We are allowed to choose the distribution to take the expectation over

– Ryan Adams[1]

[1] `Talking Machines Season 2, Episode 5`

How to choose Q?

$$\mathcal{L}(q(\mathbf{X})) = \mathbb{E}_{q(\mathbf{X})}\left[log\ p(\mathbf{X},\mathbf{Y})\right] - H(q(\mathbf{X}))$$

- We have to be able to compute an expectation over the joint distribution
- The second term should be trivial

Mean Field Approximation

$$q(\mathbf{X}) = \prod_i q_i(\mathsf{x}_i)$$

$$\mathcal{L}(q_j) = \mathcal{L}(q_j) + \mathcal{L}(q_{\neg j}),$$

- Model originating if Physics
- We model marginals rather than the full distribution
- We can update each distribution in turn and cycle

Mean Field Approximation

$$\mathcal{L}(q) = \int q(\mathbf{X}) \log \frac{p(\mathbf{Y}, \mathbf{X})}{q(\mathbf{X})} \mathrm{d}\mathbf{X}$$

Mean Field Approximation

$$\mathcal{L}(q) = \int q(\mathbf{X}) \log \frac{p(\mathbf{Y}, \mathbf{X})}{q(\mathbf{X})} \mathrm{d}\mathbf{X} = \int \prod_i q_i(\mathbf{x}_i) \log \frac{p(\mathbf{Y}, \mathbf{X})}{\prod_k q_k(\mathbf{x}_k)} \mathrm{d}\mathbf{X}$$

Mean Field Approximation

$$\mathcal{L}(q) = \int q(\mathbf{X}) \log \frac{p(\mathbf{Y}, \mathbf{X})}{q(\mathbf{X})} \mathrm{d}\mathbf{X} = \int \prod_i q_i(\mathbf{x}_i) \log \frac{p(\mathbf{Y}, \mathbf{X})}{\prod_k q_k(\mathbf{x}_k)} \mathrm{d}\mathbf{X}$$

$$= \int \prod_i q_i(\mathbf{x}_i) \left(\log p(\mathbf{Y}, \mathbf{X}) - \sum_k \log q_k(\mathbf{x}_k) \right)$$

Mean Field Approximation

$$\mathcal{L}(q_j) = \int \prod_i q_i(\mathsf{x}_i) \left(\log p(\mathsf{Y}, \mathsf{X}) - \sum_k \log q_k(\mathsf{x}_k) \right)$$

Mean Field Approximation

$$\mathcal{L}(q_j) = \int \prod_i q_i(\mathsf{x}_i) \left(\log p(\mathsf{Y}, \mathsf{X}) - \sum_k \log q_k(\mathsf{x}_k) \right)$$

$$= \int_j \int_{\neg j} q_j(\mathsf{x}_j) \prod_{i \neq j} q_i(\mathsf{x}_i) \left(\log\, p(\mathsf{X}, \mathsf{Y}) - \sum_k \log\, q_k(\mathsf{x}_k) \right) \mathrm{d}\mathsf{x}_{\neg j} \mathrm{d}\mathsf{x}_j$$

Mean Field Approximation

$$
\begin{aligned}
\mathcal{L}(q_j) &= \int \prod_i q_i(\mathbf{x}_i) \left(\log p(\mathbf{Y}, \mathbf{X}) - \sum_k \log q_k(\mathbf{x}_k) \right) \\
&= \int_j \int_{\neg j} q_j(\mathbf{x}_j) \prod_{i \neq j} q_i(\mathbf{x}_i) \left(\log\ p(\mathbf{X}, \mathbf{Y}) - \sum_k \log\ q_k(\mathbf{x}_k) \right) \mathrm{d}\mathbf{x}_{\neg j} \mathrm{d}\mathbf{x}_j \\
&= \int_j q_j(\mathbf{x}_j) \underbrace{\int_{\neg j} \prod_{i \neq j} q(\mathbf{x}_i) \log\ p(\mathbf{Y}, \mathbf{X}) \mathrm{d}\mathbf{x}_{\neg j}}_{\log\ f_j(\mathbf{x}_j)} \mathrm{d}\mathbf{x}_j \\
&\quad - \int_j q_j(\mathbf{x}_j) \int_{\neg j} \prod_{i \neq j} q(\mathbf{x}_i) \left(\log\ q_j(\mathbf{x}_j) + \sum_{k \neq j} \log\ q_k(\mathbf{x}_k) \right) \mathrm{d}\mathbf{x}_{\neg j} \mathrm{d}\mathbf{x}_j
\end{aligned}
$$

Mean Field Approximation

$$= \int_j q_j(\mathsf{x}_j) \underbrace{\int_{\neg j} \prod_{i \neq j} q(\mathsf{x}_i) \log\ p(\mathbf{Y}, \mathbf{X}) \mathrm{d}\mathsf{x}_{\neg j}}_{\log\ f_j(\mathsf{x}_j)} \mathrm{d}\mathsf{x}_j$$

$$- \int_j q_j(\mathsf{x}_j) \int_{\neg j} \prod_{i \neq j} q(\mathsf{x}_i) \left(\log\ q_j(\mathsf{x}_j) + \sum_{k \neq j} \log\ q_k(\mathsf{x}_k) \right) \mathrm{d}\mathsf{x}_{\neg j} \mathrm{d}\mathsf{x}_j$$

Mean Field Approximation

$$= \int_j q_j(\mathbf{x}_j) \underbrace{\int_{\neg j} \prod_{i \neq j} q(\mathbf{x}_i) \log\, p(\mathbf{Y}, \mathbf{X}) \mathrm{d}\mathbf{x}_{\neg j}}_{\log\, f_j(\mathbf{x}_j)} \mathrm{d}\mathbf{x}_j$$

$$- \int_j q_j(\mathbf{x}_j) \int_{\neg j} \prod_{i \neq j} q(\mathbf{x}_i) \left(\log\, q_j(\mathbf{x}_j) + \sum_{k \neq j} \log\, q_k(\mathbf{x}_k) \right) \mathrm{d}\mathbf{x}_{\neg j} \mathrm{d}\mathbf{x}_j$$

$$= \int_j q_j(\mathbf{x}_j) \log\, f_j(\mathbf{x}_j) \mathrm{d}\mathbf{x}_j$$

$$- \int_j q_j(\mathbf{x}_j) \left(\log\, q_j(\mathbf{x}_j) \underbrace{\int_{\neg j} \prod_{i \neq j} q_i(\mathbf{x}_i) \mathrm{d}\mathbf{x}_{\neg j}}_{=1} + \underbrace{\int_{\neg j} \prod_{i \neq j} q_i(\mathbf{x}_i) \sum_{k \neq j} \log\, q_k(\mathbf{x}_k) \mathrm{d}\mathbf{x}_{\neg j}}_{\text{constant w.r.t. } q_j} \right)$$

Mean Field Approximation

$$= \int_j q_j(\mathsf{x}_j) \log f_j(\mathsf{x}_j) \mathrm{d}\mathsf{x}_j$$

$$- \int_j q_j(\mathsf{x}_j) \left(\log q_j(\mathsf{x}_j) \underbrace{\int_{\neg j} \prod_{i \neq j} q_i(\mathsf{x}_i) \mathrm{d}\mathsf{x}_{\neg j}}_{=1} + \underbrace{\int_{\neg j} \prod_{i \neq j} q_i(\mathsf{x}_i) \sum_{k \neq j} \log q_k(\mathsf{x}_k) \mathrm{d}\mathsf{x}_{\neg j}}_{\text{constant w.r.t. } q_j} \right) \mathrm{d}x_j$$

Mean Field Approximation

$$= \int_j q_j(\mathsf{x}_j) \log f_j(\mathsf{x}_j) \mathrm{d}\mathsf{x}_j$$

$$- \int_j q_j(\mathsf{x}_j) \left(\log q_j(\mathsf{x}_j) \underbrace{\int_{\neg j} \prod_{i \neq j} q_i(\mathsf{x}_i) \mathrm{d}\mathsf{x}_{\neg j}}_{=1} + \underbrace{\int_{\neg j} \prod_{i \neq j} q_i(\mathsf{x}_i) \sum_{k \neq j} \log q_k(\mathsf{x}_k) \mathrm{d}\mathsf{x}_{\neg j}}_{\text{constant w.r.t. } q_j} \right) \mathrm{d}x_j$$

$$= \int_j q_j(\mathsf{x}_j) \log f_j(\mathsf{x}_j) \mathrm{d}\mathsf{x}_j - \int_j q_j(\mathsf{x}_j) \log q_j(\mathsf{x}_j) \mathrm{d}\mathsf{x}_j + \text{const.} \cdot \underbrace{\int_j q_j(\mathsf{x}_j) \mathrm{d}\mathsf{x}_j}_{=1}$$

Mean Field Approximation

$$= \int_j q_j(\mathsf{x}_j) \log f_j(\mathsf{x}_j) \mathrm{d}\mathsf{x}_j - \int_j q_j(\mathsf{x}_j) \log q_j(\mathsf{x}_j) \mathrm{d}\mathsf{x}_j + \text{const.} \cdot \underbrace{\int_j q_j(\mathsf{x}_j) \mathrm{d}\mathsf{x}_j}_{=1}$$

Mean Field Approximation

$$= \int_j q_j(\mathsf{x}_j) \log f_j(\mathsf{x}_j) \mathrm{d}\mathsf{x}_j - \int_j q_j(\mathsf{x}_j) \log q_j(\mathsf{x}_j) \mathrm{d}\mathsf{x}_j + \text{const.} \cdot \underbrace{\int_j q_j(\mathsf{x}_j) \mathrm{d}\mathsf{x}_j}_{=1}$$

$$= \int_j q_j(\mathsf{x}_j) \log \frac{f_j(\mathsf{x}_j)}{q_j(\mathsf{x}_j)} \mathrm{d}\mathsf{x}_j + \text{const.}$$

Mean Field Approximation

$$= \int_j q_j(\mathsf{x}_j) \log\, f_j(\mathsf{x}_j) \mathrm{d}\mathsf{x}_j - \int_j q_j(\mathsf{x}_j) \log\, q_j(\mathsf{x}_j) \mathrm{d}\mathsf{x}_j + \text{const.} \cdot \underbrace{\int_j q_j(\mathsf{x}_j) \mathrm{d}\mathsf{x}_j}_{=1}$$

$$\begin{aligned} &= \int_j q_j(\mathsf{x}_j) \log\, \frac{f_j(\mathsf{x}_j)}{q_j(\mathsf{x}_j)} \mathrm{d}\mathsf{x}_j + \text{const.} \\ &= -\int_j q_j(\mathsf{x}_j) \log\, \frac{q_j(\mathsf{x}_j)}{f_j(\mathsf{x}_j)} \mathrm{d}\mathsf{x}_j + \text{const.} \end{aligned}$$

Mean Field Approximation

$$= \int_j q_j(\mathsf{x}_j) \log f_j(\mathsf{x}_j) \mathrm{d}\mathsf{x}_j - \int_j q_j(\mathsf{x}_j) \log q_j(\mathsf{x}_j) \mathrm{d}\mathsf{x}_j + \text{const.} \cdot \underbrace{\int_j q_j(\mathsf{x}_j) \mathrm{d}\mathsf{x}_j}_{=1}$$

$$\begin{aligned}
&= \int_j q_j(\mathsf{x}_j) \log \frac{f_j(\mathsf{x}_j)}{q_j(\mathsf{x}_j)} \mathrm{d}\mathsf{x}_j + \text{const.} \\
&= -\int_j q_j(\mathsf{x}_j) \log \frac{q_j(\mathsf{x}_j)}{f_j(\mathsf{x}_j)} \mathrm{d}\mathsf{x}_j + \text{const.} \\
&= -\mathsf{KL}(q_j(\mathsf{x}_j) || f_j(\mathsf{x}_j)) + \text{const.}
\end{aligned}$$

Mean Field Approximation

$$= \int_j q_j(\mathsf{x}_j) \log f_j(\mathsf{x}_j) \mathrm{d}\mathsf{x}_j - \int_j q_j(\mathsf{x}_j) \log q_j(\mathsf{x}_j) \mathrm{d}\mathsf{x}_j + \text{const.} \cdot \underbrace{\int_j q_j(\mathsf{x}_j) \mathrm{d}\mathsf{x}_j}_{=1}$$

$$\begin{aligned} &= \int_j q_j(\mathsf{x}_j) \log \frac{f_j(\mathsf{x}_j)}{q_j(\mathsf{x}_j)} \mathrm{d}\mathsf{x}_j + \text{const.} \\ &= -\int_j q_j(\mathsf{x}_j) \log \frac{q_j(\mathsf{x}_j)}{f_j(\mathsf{x}_j)} \mathrm{d}\mathsf{x}_j + \text{const.} \\ &= -\mathrm{KL}(q_j(\mathsf{x}_j) || f_j(\mathsf{x}_j)) + \text{const.} = \mathcal{L}(q_j) \end{aligned}$$

Mean Field Approximation

$$\mathcal{L}(q_j) = -\mathsf{KL}(q_j(\mathbf{x}_j)||f_j(\mathbf{x}_j)) + \text{const.}$$

- Want to maximise lower bound
- Negative KL $\rightarrow$ minimise KL term
- *we are free to choose the form of the distribution*

Mean Field Approximation

$$\log q_j(\mathbf{x}_j) = \log f_j(\mathbf{x}_j) = \int_{\neg j} \underbrace{\prod_{i \neq j} q_i(\mathbf{x}_i)}_{q_{\neg j}(\mathbf{x}_{\neg j})} \log p(\mathbf{Y}, \mathbf{X}) \mathrm{d}\mathbf{x}_{\neg j}$$

$$= \mathbb{E}_{q_{\neg j}(\mathbf{x}_{\neg j})} \left[\log p(\mathbf{Y}, \mathbf{X})\right]$$

- Choose the marginal distribution that makes the bound tight
- Will not make the bound tight in general though

Mean Field Variational Bayes

1. Formulate joint distribution over data and latent parameters

Mean Field Variational Bayes

1. Formulate joint distribution over data and latent parameters
2. Formulate fully factorised approximative posterior over latent variables

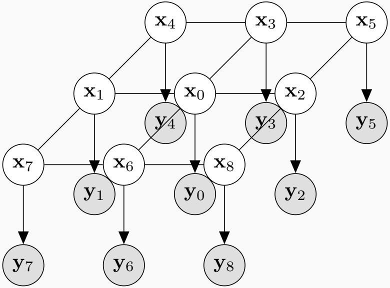
Mean Field Variational Bayes

1. Formulate joint distribution over data and latent parameters
2. Formulate fully factorised approximative posterior over latent variables
3. Fit marginal approximation by making bound tight

Mean Field Variational Bayes

1. Formulate joint distribution over data and latent parameters
2. Formulate fully factorised approximative posterior over latent variables
3. Fit marginal approximation by making bound tight
4. Iterate through variables

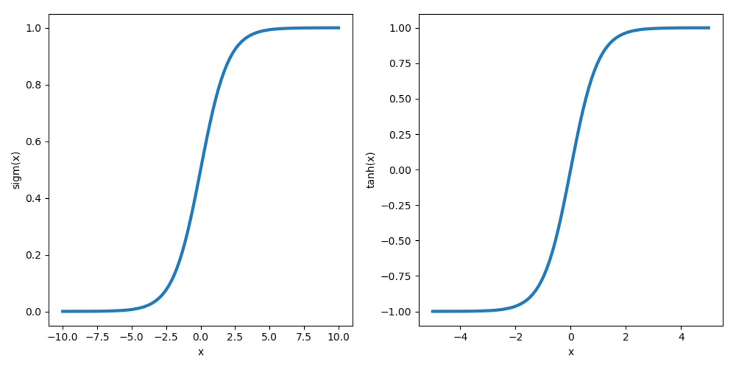
Coursework



$$q(\mathbf{x}, \boldsymbol{\mu}) = \prod_{i}^{N} q(x_i, \mu_i)$$

$$\mu_i = \mathbb{E}[x_i]$$

Coursework



sigm(x)
x
1.0
0.8
0.6
0.4
0.2
0.0
−10.0 −7.5 −5.0 −2.5 0.0 2.5 5.0 7.5 10.0
tanh(x)
x
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−4 −2 0 2 4

Summary

Summary

- Variational methods can be very efficient
  - really fun to work with
- Can be made black-box [2]
- Will never be correct
- Provides us with approximative posterior for inference

Explore Week

- No lectures next week
-

eof

References

Christopher M. Bishop.
***Pattern Recognition and Machine Learning (Information Science and Statistics)*.**
Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

Rajesh Ranganath, Sean Gerrish, and David Blei.
Black Box Variational Inference.
In Samuel Kaski and Jukka Corander, editors, *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 814–822, Reykjavik, Iceland, 22–25 Apr 2014. PMLR.