Machine Learning

Deterministic Approximative Inference

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

December 2, 2019

http://www.carlhenrik.com

Big Number

$$p(y) = \sum_{n=1}^{N} p(y, x_i)$$

- Number of atoms in the universe

$$10^{80} \approx (2^{\frac{10}{3}})^{80} \approx 2^{267}$$

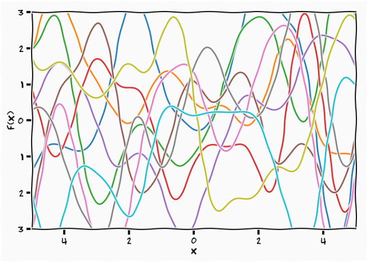
- Age of the universe in seconds

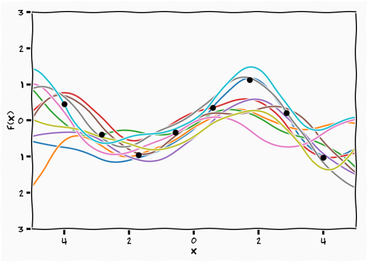
$$4.35 \cdot 10^{17} \approx 2^{59}$$

Big Number

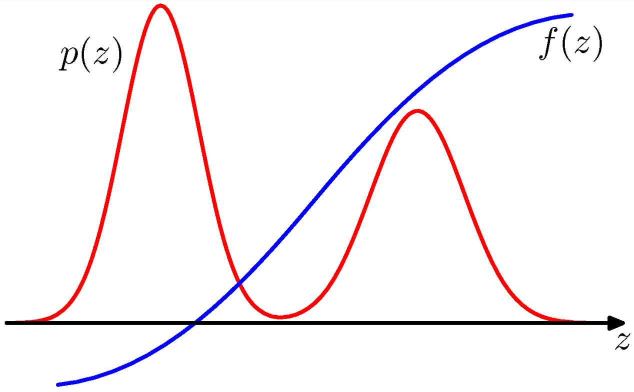
$$\infty$$

Big Number





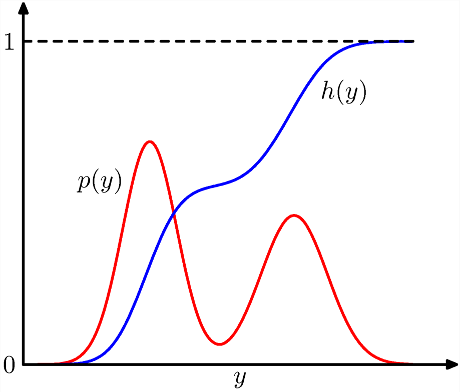
$$p(\mathbf{y}|\mathbf{x}) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{x})\mathrm{d}\mathbf{f}$$



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

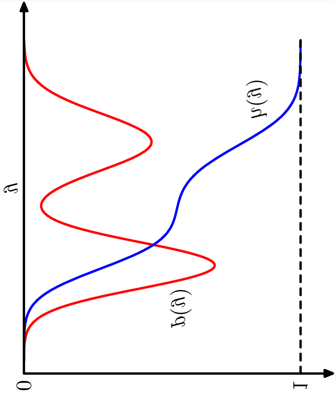
$$\mathbf{z}^{(l)} \sim p(\mathbf{z})$$

Basic Probabilities



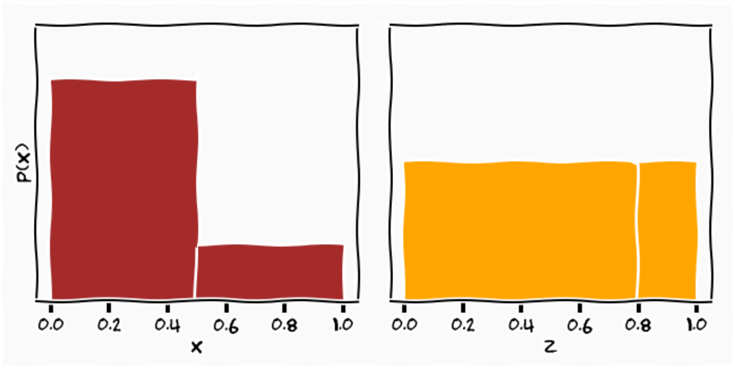
$$z = h(y) = \int_{-\infty}^{y} p(y)\mathrm{d}y$$

Basic Probabilities



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

Change of Variables



Change of Variables

$$\int_{y \in \mathcal{Y}} p(y) \mathrm{d}y = \int_{x \in \mathcal{X}} p(x) \mathrm{d}x$$

Change of Variables

$$\int_{y \in \mathcal{Y}} p(y)\mathrm{d}y = \int_{x \in \mathcal{X}} p(x)\mathrm{d}x$$

$$x = x(y)$$

Change of Variables

$$\int_{y \in \mathcal{Y}} p(y)\mathrm{d}y = \int_{x \in \mathcal{X}} p(x)\mathrm{d}x$$

$$x = x(y)$$

$$\int_{y \in \mathcal{Y}} p(y)\mathrm{d}y = \int_{y \in \mathcal{Y}} p(x(y))\frac{\mathrm{d}x}{\mathrm{d}y}\mathrm{d}y$$

Change of Variables

$$\int_{y \in \mathcal{Y}} p(y) \mathrm{d}y = \int_{x \in \mathcal{X}} p(x) \mathrm{d}x$$

$$x = x(y)$$

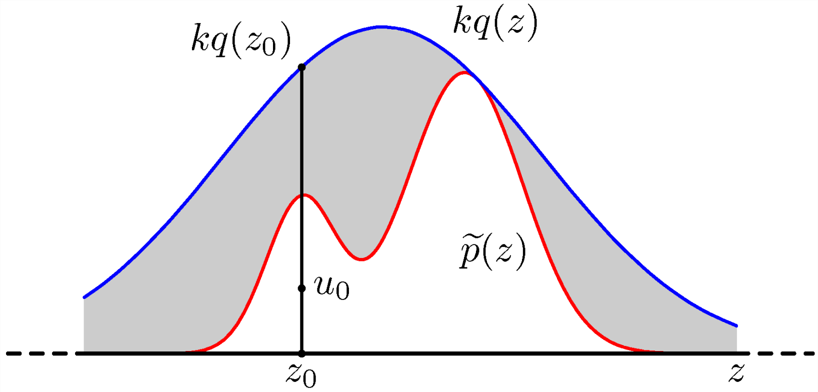
$$\int_{y \in \mathcal{Y}} p(y) \mathrm{d}y = \int_{y \in \mathcal{Y}} p(x(y)) \frac{\mathrm{d}x}{\mathrm{d}y} \mathrm{d}y$$

$$p(y) = p(x(y)) \frac{\mathrm{d}x}{\mathrm{d}y}$$

Starting point

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

Rejection Sampling



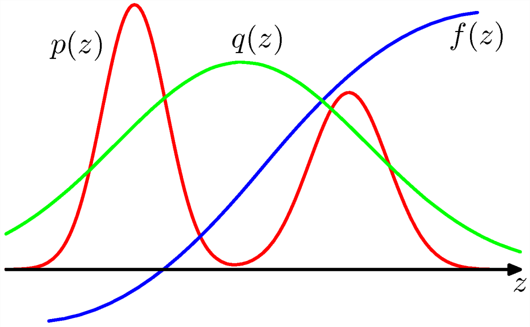
Importance Sampling Ch 11.1.4 [1]

$$\mathbb{E}_{p(\mathbf{z})}[f] = \int f(\mathbf{z}) p(\mathbf{z}) \mathrm{d}\mathbf{z} = \int f(\mathbf{z}) \frac{p(\mathbf{z})}{q(\mathbf{z})} q(\mathbf{z}) \mathrm{d}\mathbf{z} = \mathbb{E}_{q(\mathbf{z})}\left[f(\mathbf{z}) \frac{p(\mathbf{z})}{q(\mathbf{z})} \right]$$

$$\approx \frac{1}{L} \sum_{l=1}^{L} \frac{p(z^{(l)})}{q(z^{(l)})} \cdot f(\mathbf{z}^{(l)})$$

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

Importance Sampling



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

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

Summary

- We can sample from distributions whos explicit form we know

Summary

- We can sample from distributions whos explicit form we know
- Formulate different proxies using these distributions to the general case

Summary

- We can sample from distributions whos explicit form we know
- Formulate different proxies using these distributions to the general case
    - Rejection sampling

Summary

- We can sample from distributions whos explicit form we know
- Formulate different proxies using these distributions to the general case
  - Rejection sampling
  - Importance sampling

Summary

- We can sample from distributions whos explicit form we know
- Formulate different proxies using these distributions to the general case
  - Rejection sampling
  - Importance sampling
- Create a random walk to encourage exploration

Summary

- We can sample from distributions whos explicit form we know
- Formulate different proxies using these distributions to the general case
  - Rejection sampling
  - Importance sampling
- Create a random walk to encourage exploration
  - Gibbs sampling

Deterministic Approximations

Introduction

- Stochastic inference
  - approximate expectation with sum
  - works in the limit
  - hard to know how well we are doing
  - usually slow

Introduction

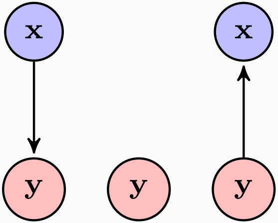
- Stochastic inference
  - approximate expectation with sum
  - works in the limit
  - hard to know how well we are doing
  - usually slow
- Idea
  - *can we reformulate inference as optimisation?*

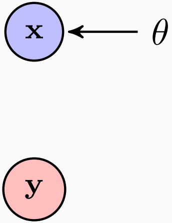
Machine Learning

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

- 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

Lower Bound

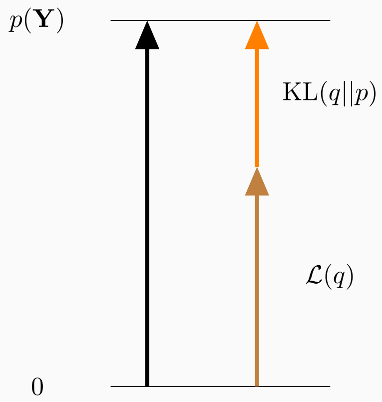


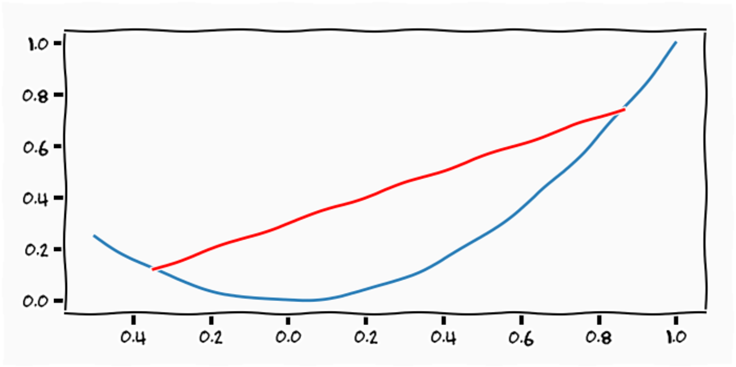


$$p(\mathbf{y}) = \int_{\mathbf{x}} p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{x}|\mathbf{y})}$$

$$q_\theta(\mathbf{x}) \approx p(\mathbf{x}|\mathbf{y})$$

Deterministic Approximation





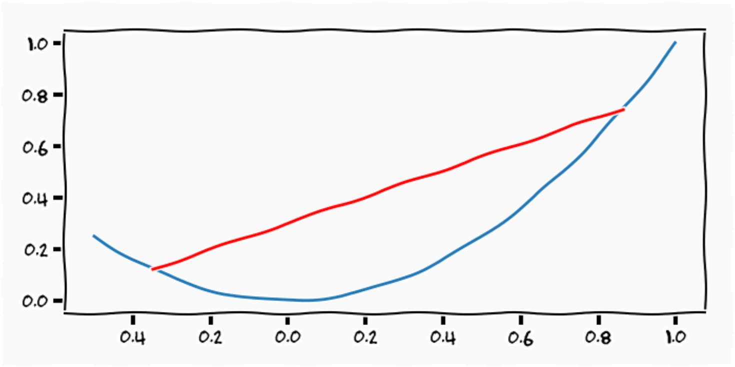
Convex Function

$$\lambda f(x_0) + (1-\lambda)f(x_1) \geq f(\lambda x_0 + (1-\lambda)x_1)$$

$$x \in [x_{min}, x_{max}]$$

$$\lambda \in [0,1]]$$

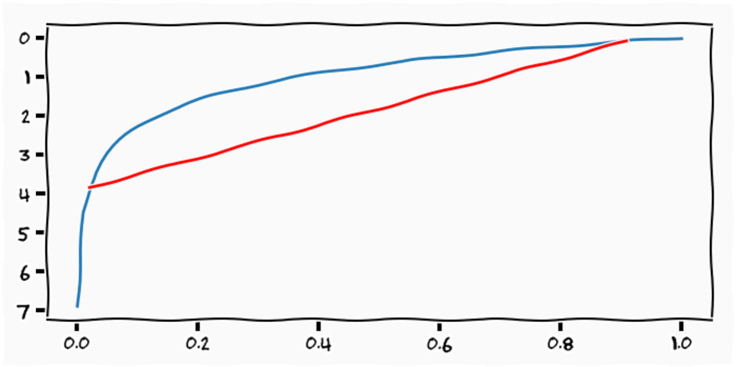
Jensen Inequality



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

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

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

$$p(y)$$

Variational Bayes

$$\log p(y)$$

Variational Bayes

$$\log\, p(y) = \log\, p(y) + \int \log\, \frac{p(x|y)}{p(x|y)} \mathrm{d}x$$

Variational Bayes

$$\log\, p(y) = \log\, p(y) + \int \log\, \frac{p(x|y)}{p(x|y)} \mathrm{d}x$$
$$= \int q(x) \log\, p(y) \mathrm{d}x + \int q(x) \log\, \frac{p(x|y)}{p(x|y)} \mathrm{d}x$$

Variational Bayes

$$
\begin{aligned}
\log\, p(y) &= \log\, p(y) + \int \log\, \frac{p(x|y)}{p(x|y)} \mathrm{d}x \\
&= \int q(x) \log\, p(y) \mathrm{d}x + \int q(x) \log\, \frac{p(x|y)}{p(x|y)} \mathrm{d}x \\
&= \int q(x) \log\, \frac{p(x|y)p(y)}{p(x|y)} \mathrm{d}x
\end{aligned}
$$

Variational Bayes

$$\log\, p(y) = \log\, p(y) + \int \log\, \frac{p(x|y)}{p(x|y)} \mathrm{d}x$$

$$= \int q(x) \log\, p(y) \mathrm{d}x + \int q(x) \log\, \frac{p(x|y)}{p(x|y)} \mathrm{d}x$$

$$= \int q(x) \log\, \frac{p(x|y)p(y)}{p(x|y)} \mathrm{d}x$$

$$= \int q(x) \log\, \frac{q(x)}{q(x)} \mathrm{d}x + \int q(x) \log\, p(x, y) \mathrm{d}x + \int q(x)\, \log\, \frac{1}{p(x|y)} \mathrm{d}x$$

$$
\begin{aligned}
\log p(y) &= \log p(y) + \int \log \frac{p(x|y)}{p(x|y)} \mathrm{d}x \\
&= \int q(x) \log p(y) \mathrm{d}x + \int q(x) \log \frac{p(x|y)}{p(x|y)} \mathrm{d}x \\
&= \int q(x) \log \frac{p(x|y)p(y)}{p(x|y)} \mathrm{d}x \\
&= \int q(x) \log \frac{q(x)}{q(x)} \mathrm{d}x + \int q(x) \log p(x,y) \mathrm{d}x + \int q(x) \log \frac{1}{p(x|y)} \mathrm{d}x \\
&= {\color{blue}\int q(x) \log \frac{1}{q(x)} \mathrm{d}x + \int q(x) \log p(x,y) \mathrm{d}x} + {\color{red}\int q(x) \log \frac{q(x)}{p(x|y)} \mathrm{d}x}
\end{aligned}
$$

The "posterior" term

$$KL(q(x)||q(x|y)) = \int q(x) \log \frac{q(x)}{p(x|y)} \mathrm{d}x$$

The "posterior" term

$$
\begin{aligned}
\color{red}{KL(q(x)||q(x|y))} &= \int q(x) \log \frac{q(x)}{p(x|y)} \mathrm{d}x \\
&= -\int q(x) \log \frac{p(x|y)}{q(x)} \mathrm{d}x
\end{aligned}
$$

The "posterior" term

$$
\begin{aligned}
KL(q(x)||q(x|y)) &= \int q(x)\,\log\frac{q(x)}{p(x|y)}\mathrm{d}x \\
&= -\int q(x)\,\log\frac{p(x|y)}{q(x)}\mathrm{d}x \\
&\geq -\log\int p(x|y)\mathrm{d}x = -\log 1 = 0
\end{aligned}
$$

The "posterior" term

$$KL(q(x)||p(x|y)) = \int q(x) \log \frac{q(x)}{p(x|y)} \mathrm{d}x$$

The "posterior" term

$$
\begin{aligned}
KL(q(x)||p(x|y)) &= \int q(x) \log \frac{q(x)}{p(x|y)} \mathrm{d}x \\
&= \{\text{Lets assume that } q(x) = p(x|y)\}
\end{aligned}
$$

The "posterior" term

$$
\begin{aligned}
\color{red}{KL(q(x)||p(x|y))} &= \int q(x) \, \log \frac{q(x)}{p(x|y)} \mathrm{d}x \\
&= \{\text{Lets assume that } q(x) = p(x|y)\} \\
&= \int p(x|y) \, \log \underbrace{\frac{p(x|y)}{p(x|y)}}_{=1} \mathrm{d}x
\end{aligned}
$$

The "posterior" term

$$
\begin{aligned}
\color{red}{KL(q(x)||p(x|y))} &= \int q(x)\, \log \frac{q(x)}{p(x|y)} \mathrm{d}x \\
&= \{\text{Lets assume that } q(x) = p(x|y)\} \\
&= \int p(x|y)\, \log \underbrace{\frac{p(x|y)}{p(x|y)}}_{=1} \mathrm{d}x \\
&= 0
\end{aligned}
$$

Kullback-Leibler Divergence

$$KL(q(x)||p(x|y)) = \int q(x) \log \frac{q(x)}{p(x|y)} \mathrm{d}x$$

- Measure of divergence between distributions
- Not a metric (not symmetric)
- $KL(q(x)||p(x|y)) = 0 \Leftrightarrow q(x) = p(x|y)$
- $KL(q(x)||p(x|y)) \geq 0$

The "other terms"

$$\int q(x) \log \frac{1}{q(x)} \mathrm{d}x + \int q(x) \log p(x, y) \mathrm{d}x =$$

The "other terms"

$$\int q(x) \log \frac{1}{q(x)} \mathrm{d}x + \int q(x) \log p(x, y) \mathrm{d}x =$$

$$= \int q(x) \log \frac{p(x, y)}{q(x)} \mathrm{d}x$$

The "other terms"

$$\int q(x)\log \frac{1}{q(x)}\mathrm{d}x + \int q(x)\log p(x,y)\mathrm{d}x =$$

$$= \int q(x)\log \frac{p(x,y)}{q(x)}\mathrm{d}x$$

$$= \{\text{Lets assume that } q(x) = p(x|y)\}$$

The "other terms"

$$
\begin{aligned}
&\int q(x)\log \frac{1}{q(x)}\mathrm{d}x + \int q(x)\log p(x,y)\mathrm{d}x = \\
&= \int q(x)\log \frac{p(x,y)}{q(x)}\mathrm{d}x \\
&= \{\text{Lets assume that } q(x) = p(x|y)\} \\
&= \int p(x|y)\log \frac{p(x,y)}{p(x|y)}\mathrm{d}x
\end{aligned}
$$

The "other terms"

$$\int q(x)\log \frac{1}{q(x)}\mathrm{d}x + \int q(x)\log p(x,y)\mathrm{d}x =$$

$$= \int q(x)\log \frac{p(x,y)}{q(x)}\mathrm{d}x$$

$$= \{\text{Lets assume that } q(x) = p(x|y)\}$$

$$= \int p(x|y)\log \frac{p(x,y)}{p(x|y)}\mathrm{d}x = \int p(x|y)\log \frac{p(x|y)p(y)}{p(x|y)}\mathrm{d}x$$

The "other terms"

$$\int q(x)\log \frac{1}{q(x)}\mathrm{d}x + \int q(x)\log p(x,y)\mathrm{d}x =$$

$$= \int q(x)\log \frac{p(x,y)}{q(x)}\mathrm{d}x$$

$$= \{\text{Lets assume that } q(x) = p(x|y)\}$$

$$= \int p(x|y)\log \frac{p(x,y)}{p(x|y)}\mathrm{d}x = \int p(x|y)\log \frac{p(x|y)p(y)}{p(x|y)}\mathrm{d}x$$

$$= \int p(x|y)\log \underbrace{\frac{p(x|y)}{p(x|y)}\mathrm{d}x}_{=1} + \int p(x|y)\log p(y)\mathrm{d}x$$

The "other terms"

$$\int q(x)\log\frac{1}{q(x)}\mathrm{d}x + \int q(x)\log p(x,y)\mathrm{d}x =$$

$$= \int q(x)\log\frac{p(x,y)}{q(x)}\mathrm{d}x$$

$$= \{\text{Lets assume that } q(x) = p(x|y)\}$$

$$= \int p(x|y)\log\frac{p(x,y)}{p(x|y)}\mathrm{d}x = \int p(x|y)\log\frac{p(x|y)p(y)}{p(x|y)}\mathrm{d}x$$

$$= \int p(x|y)\log\underbrace{\frac{p(x|y)}{p(x|y)}\mathrm{d}x}_{=1} + \int p(x|y)\log p(y)\mathrm{d}x$$

$$= \underbrace{\int p(x|y)\mathrm{d}x}_{=1}\log p(y)$$

The "other terms"

$$\int q(x)\log \frac{1}{q(x)}\mathrm{d}x + \int q(x)\log p(x,y)\mathrm{d}x =$$

$$= \int q(x)\log \frac{p(x,y)}{q(x)}\mathrm{d}x$$

$$= \{\text{Lets assume that } q(x) = p(x|y)\}$$

$$= \int p(x|y)\log \frac{p(x,y)}{p(x|y)}\mathrm{d}x = \int p(x|y)\log \frac{p(x|y)p(y)}{p(x|y)}\mathrm{d}x$$

$$= \int p(x|y)\log \underbrace{\frac{p(x|y)}{p(x|y)}\mathrm{d}x}_{=1} + \int p(x|y)\log p(y)\mathrm{d}x$$

$$= \underbrace{\int p(x|y)\mathrm{d}x}_{=1} \log p(y) = \log p(y)$$

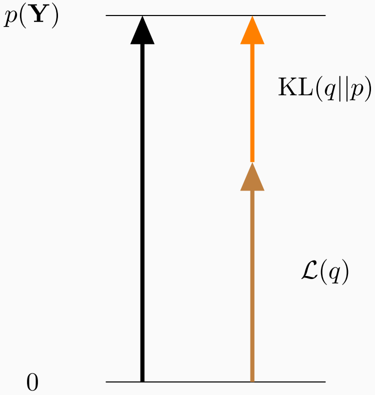
Variational Bayes

$$\log\, p(y) = \int q(x)\log\,\frac{1}{q(x)}\mathrm{d}x + \int q(x)\log\, p(x,y)\mathrm{d}x + \int q(x)\ \log\,\frac{q(x)}{p(x|y)}\mathrm{d}x$$

$$\geq -\int q(x)\log\, q(x)\mathrm{d}x + \int q(x)\log\, p(x,y)\mathrm{d}x$$

- The Evidence Lower BOnd
- Tight if $q(x) = p(x|y)$

Deterministic Approximation

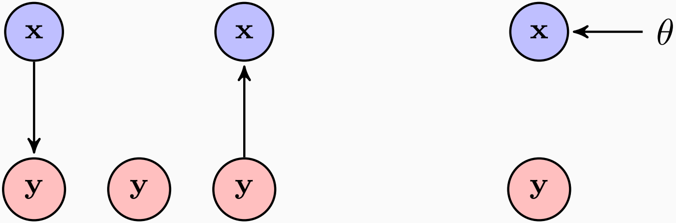


ELBO

$$\log p(y) \geq -\int q(x)\log q(x)\mathrm{d}x + \int q(x)\log p(x,y)\mathrm{d}x$$
$$= \mathbb{E}_{q(x)}\left[\log p(x,y)\right] - H(q(x)) = \mathcal{L}(q(x))$$

- if we maximise the ELBO we,
    - find an approximate posterior
    - lower bound the marginal likelihood
- *maximising* $p(y)$ is learning
- finding $q(x) \approx p(x|y)$ is prediction

Lower Bound



$$p(\mathbf{y}) = \int_{\mathbf{x}} p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{x}|\mathbf{y})}$$

$$q_\theta(\mathbf{x}) \approx p(\mathbf{x}|\mathbf{y})$$

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?

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

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

Mean Field Approximation

$$q(\mathbf{X}) = \prod_i q_i(\mathbf{x}_i)$$
$$\mathcal{L}(q_j) = \mathcal{L}_j(q_j) + \mathcal{L}_{\neg j}(q_{\neg j}),$$

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

Mean Field 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

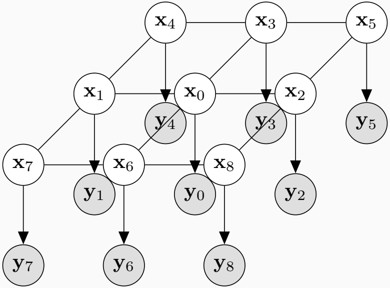
Mean Field Variational Bayes

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

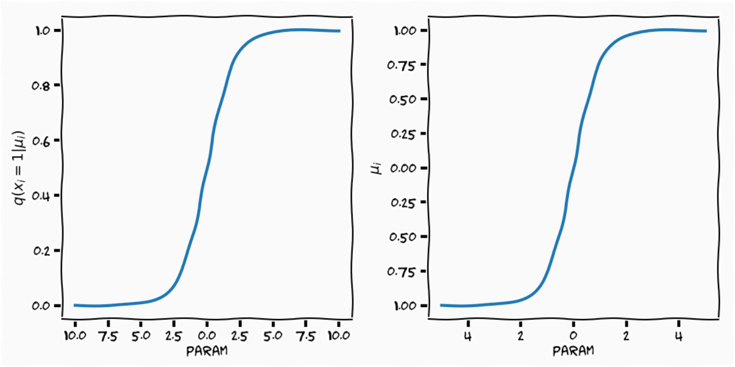
Lab



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

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

Lab



$q(x_i = 1|\mu_i)$
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
PARAM
μ_i
1.00
0.75
0.50
0.25
0.00
0.25
0.50
0.75
1.00
4
2
0
2
4
PARAM

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 predictions

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.