Machine Learning

Stochastic Approximative Inference

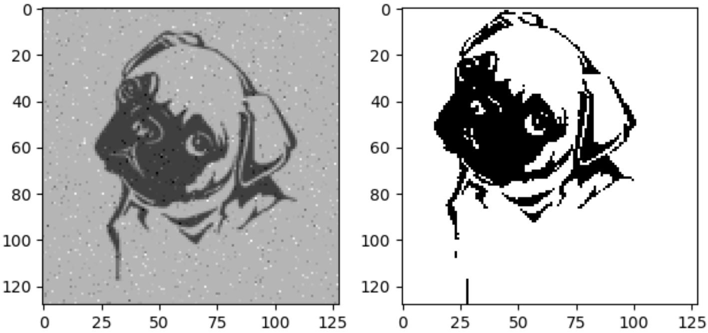
Carl Henrik Ek - carlhenrik.ek@bristol.ac.uk

November 12, 2018

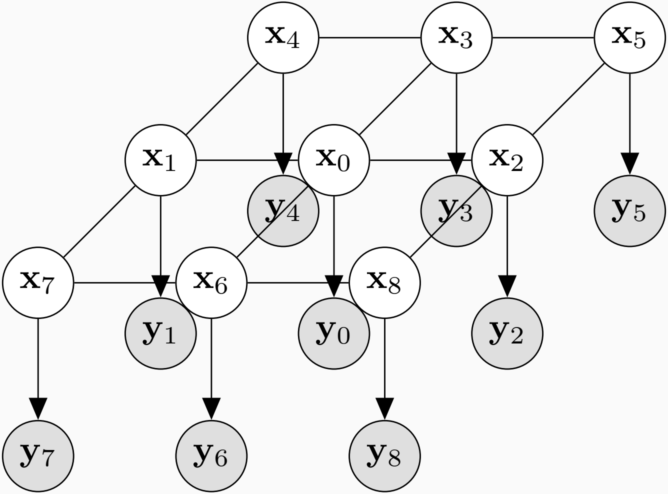
http://www.carlhenrik.com

Introduction

Coursework II



Coursework II



Posterior: Markov Random Field

- Posterior

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

Posterior: Markov Random Field

- Posterior

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

- For the MRF the marginal likelihood/evidence can be computed as

$$p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) = \sum_{i}^{N} p(\mathbf{y}|\mathbf{x}_i)p(\mathbf{x}_i)$$

Posterior: Markov Random Field

- Posterior

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

- For the MRF the marginal likelihood/evidence can be computed as

$$p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) = \sum_{i}^{N} p(\mathbf{y}|\mathbf{x}_i)p(\mathbf{x}_i)$$

- $\mathbf{x}_i$ is a specific binary image


3G
9:42 AM
SMS
Text
26
Calendar
Photos
Camera
YouTube
Stocks
Maps
Clock
Calculator
Weather
Notes
Settings
iTunes
App Store
Remote
Remote
Phone
Mail
Safari
iPod

Number of terms I

2290593203500326442498254071102
8779924646158308390547680551234
5054431338510774037915738775865
8057318635099533562444284837656
6408900340661545734126916095393
4651531316272895970961099648619
5486636741656944283948869330648
4701733713508133208092688099524
0707971539803921050200955733579
43662055566676730638553849508752
9677470990968153918788613785751
3890052212385415364000233552517

Number of terms II

9230941551480812783648467474496
15787812522261713953420063416790
7552057630497077601674681891226
1453204962575441115371836944715
6895505073882545721273943517481
6507334054019330445298798029650
8746618030728963410359112463410
9184832439049686890853942279882
9655406361370980789697504759416
7461331023628146001054998291892
8850448033966038407878196527044
7157474368533868315778800203562
1474121034155871572968019805251

Number of terms III

8982409725023084881200238736500

2027283572275248844963488736471

3943526031912848227248826190464

8476965948928382396693052519124

1687725175533908692952453783598

2837023543516588536916371046489

4220310701508827933380526429979

2599815801920922903898158871712

8926097153382729134531621865313

9786085815417055159827515344471

3326325034781836776513703100360

9793889758575377908303501066776

6548311999605347475370343426743

Number of terms IV

8253400053810997864187276609708

2093090380663944422789696913654

8900202322285082544979530967870

6304437009833849217731493021674

2550624871750833859476679189509

5680602732346712939153259990811

4893913032842065037601973054196

1524092173016464047938013691439

6671843203605981118777513627755

7250792266837423597968228683403

4089138475154767372727122932222

8878852083218796660305975797728

8778298768646815994259957325408

Number of terms V

8749600987758158350339985951647
5121708697580746029473842801833
8592485796034133919973077413533
6869491956368516611377674237208
1780419191068702807890339161440
9912666138730775266005780452422
5302437317858452782485229505751
3761093944464722805553911771716
4315059230286413698788578331540
1782239495790781650110059887274
5959467831004471989549305375741
9073809906471822251882514747849
0657161167548497523333968812279

Number of terms VI

4911475119965635459462447339289
7828672753085721621023943443062
0144907278084466853892944205719
8697060107876495003418069047901
8142025673307261276950347320181
6461274039931292984401423199725
4340930170763466037725337419662
9143599599348813527131013125346
3508530232037816302115328138866
8643014293963947674718567131663
5043595580465472543695170605663
2361702749907044372801683830358
6991365299464326205642839343150

Number of terms VII

4053504888101754720253838078891

9253939272110382634932825138554

3816977282386956487514065578882

3474751813846542682825520838131

0069117625217360239526199430454

3464350338428593031654513507976

7510717638042435127189839307791

2093765743451201386745554882022

4148073627378623609980111113076

0640189547044207203761774747082

0243516866198003957569584101060

8046613562965001201466456771415

5778664863093617634553900426210

Number of terms VIII

9110167208910075825348801584001
7224071067971558665492397885347
6607256313817084019127947685341
8537351879721277733449450507730
3189505040470344922506903873556
9656865708529073446623478695245
6543122517479114466613670208736
0842313671545657762822696089905
6802168279902278674508669673834
7816102210900054189076993778672
7705964820658607375143364171301
1744511704016132334906338900377
1777472580944833242545989973822

Number of terms IX

5646744609738390155521757096422

2619375692340966923479020630115

9076383049447801135255878205328

2752643299087648267991015324907

4963538068771014944040060242262

3804497742682401904233153226013

9373317250133351983527123955504

2292211010517136771541981666250

0131430427440349387764312765762

4870317305687566284108475166000

1324414350620739304183073837766

8972502903711649967733818943578

9237255328232566165426546313829 1135999395862937 6

Numbers

- Possible black and white 3 Megapixel images

$$2^{3145728}$$

Numbers

- Possible black and white 3 Megapixel images

$$2^{3145728}$$

- Number of atoms in the universe

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

Numbers

- Possible black and white 3 Megapixel images

$$2^{3145728}$$

- Number of atoms in the universe

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

- Age of the universe in seconds

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

Numbers

- Possible black and white 3 Megapixel images

$$2^{3145728}$$

- Number of atoms in the universe

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

- Age of the universe in seconds

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

- *Lets agree that this for loop is intractable*

$$\hat{\boldsymbol{\theta}} = \mathrm{argmax}_{\theta} p(\mathcal{D}|\boldsymbol{\theta})$$

MCMC Man

Bio

- Takes his time
- Works on his own and is hard to understand
- Will in the limit always catch the right guy
- Works a lot of cold cases

"To catch the right guy we need to consider every avenue, every possibility"

Variational Bayes Woman

Bio

- Gets the job done
- Reports to central command
- Believes smoke implies fire
- Sometimes catches the wrong guy

"Shoot first, ask questions later"

Coursework

- Stochastic approximation (today)
    - Iterative Conditional Modes (ICM)
    - Markov Chain Monte Carlo, Gibbs Sampler

Coursework

- Stochastic approximation (today)
    - Iterative Conditional Modes (ICM)
    - Markov Chain Monte Carlo, Gibbs Sampler
- Deterministic approximation (tomorrow)
    - Mean-field Variational Bayes
    - Mean-field Ising model VB update equations
        - derivation tomorrow lecture 2

Coursework

- Stochastic approximation (today)
  - Iterative Conditional Modes (ICM)
  - Markov Chain Monte Carlo, Gibbs Sampler
- Deterministic approximation (tomorrow)
  - Mean-field Variational Bayes
  - Mean-field Ising model VB update equations
    - derivation tomorrow lecture 2
- Extra [Kingma and Welling, 2014]
  - Amortised Inference in tensorflow

Coursework

- Availible on GitHub
- 10 Questions
- Deadline Friday 7th of December 12:00
  - extended one week from previous date
- Groups of two submits report

Laplace Approximation

Analytical Intractability

$$\log\, p(\mathbf{w}|\mathbf{t}) = \log\left(\prod_i^N \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i)^{t_i} \cdot (1-\sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i))^{1-t_i}\right) - \frac{1}{2}(\mathbf{w}-\mathbf{m}_0)^{\mathrm{T}}\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0) - \log(Z)$$

- Sometimes conjugacy does not make sense
- The prior and the likelihood makes sense by themselves
- Classification is the typical example

Laplace Approximation

$$p(z) = \frac{1}{Z} f(z) = \frac{f(z)}{\int f(z) dz}$$

- $p(z)$ is unknown as we cannot compute Z
- $f(z)$ is possible to compute if we have likelihood and prior

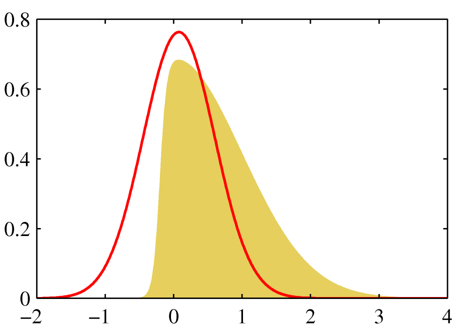
$$f(z) = p(x|z)p(z)$$

Laplace Approximation

$$\log\, p(z) = \log\, (\frac{1}{Z} f(z)) = \log(f(z)) + \text{const w.r.t. } z$$

- $p(z)$ and $f(z)$ will have the same modes
- Idea: we can approximate each mode with a distribution we can normalise

Laplace Approximation Ch. 4.4 [Bishop, 2006]



- Find the mode of the posterior
- Fit Gaussian to this mode

Taylor Expansion

$$f(x) = f(x_0) + \frac{\partial}{\partial x} f(x_0)(x - x_0) + \frac{1}{2}\frac{\partial^2}{\partial x^2} f(x_0)(x - x_0)^2 + \mathcal{O}((x - x0)^3)$$

- A Taylor expansion is an approximation of a function around a specific value
- If we expand around a maxima x_0

$$\frac{\partial}{\partial x} f(x_0) = 0$$

- This leads to

$$f(x) = f(x_0) - \frac{1}{2}\left|\frac{\partial^2}{\partial x^2} f(x_0)\right|(x - x_0)^2 + \mathcal{O}((x - x0)^3)$$

Laplace Approximation

1. Find mode of $p(z)$

$$\frac{\partial}{\partial z} p(z_0) = \frac{\partial}{\partial z} f(z_0) = 0$$

2. Make Taylor Expansion around mode

$$\log f(z) \approx \log f(z_0) - \frac{1}{2} \frac{\partial^2}{\partial^2} \log(f(z_0))(z - z_0)^2$$

3. Take exponential to get function

$$f(z) \approx f(z_0) e^{-\frac{1}{2} \underbrace{\frac{\partial^2}{\partial^2} \log(f(z_0))}_{A} (z-z_0)^2} = f(z_0) e^{-\frac{1}{2} A (z-z_0)^2}$$

Laplace Approximation

$$f(z) \approx f(z_0) e^{-\frac{1}{2}A(z-z_0)^2}$$

- we want to find an approximation, to $p(z)$ so we need to normalise to a distribution

$$p(z) = \frac{1}{Z} f(z) \approx q(z)$$

- assume that $q(z)$ is Gaussian, i.e. $f(z_0) = p(\text{mean})$

$$q(z) = \left(\frac{A}{2\pi}\right)^{\frac{1}{2}} e^{-\frac{A}{2}(z-z_0)^2}$$

Laplace Approximation

- Compute a mode of the posterior distribution, i.e MAP estimate

Laplace Approximation

- Compute a mode of the posterior distribution, i.e MAP estimate
- Perform Taylor expansion around mode to quadratic term

Laplace Approximation

- Compute a mode of the posterior distribution, i.e MAP estimate
- Perform Taylor expansion around mode to quadratic term
- Identify elements in expansion as parameters of a Gaussian

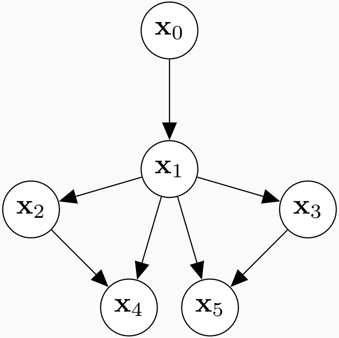
Laplace Approximation

- Compute a mode of the posterior distribution, i.e MAP estimate
- Perform Taylor expansion around mode to quadratic term
- Identify elements in expansion as parameters of a Gaussian
- Normalise to a distribution

Stochastic Approximative Inference

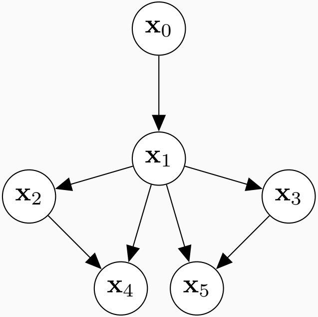
Cookbook

How to sample?



$$p(\mathbf{x}) = \prod_i p(x_i | \mathrm{pa}_i)$$

Ancestral Sampling

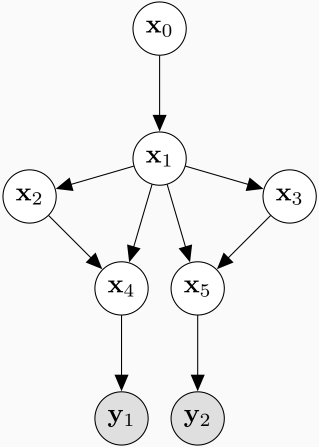


Sample from $p(\mathbf{x})$

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

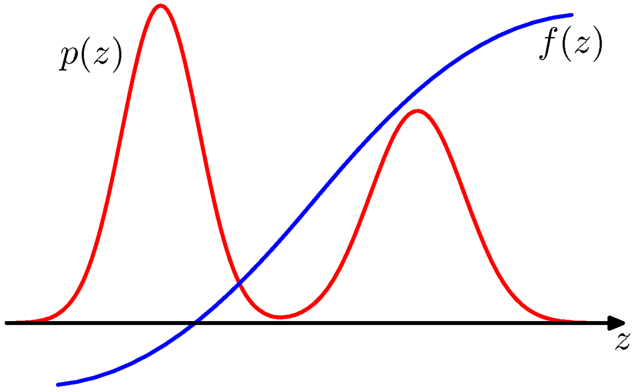
$$p(\mathbf{x}) = p(x_5|x_3, x_1)p(x_4|x_2, x_1)p(x_3|x_1)p(x_2|x_1)p(x_1|x_0)p(x_0)$$

Observed Data



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

1. Ancestral sampling for all latent variables
2. When latent variables child is observed
    - sample from conditional
    - if sample agrees with observation $\mathbf{x}$ comes from posterior
    - if not discard sample and restart



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

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

Sampling

$$\hat{f} = \frac{1}{L}\sum_{l=1}^{L} f(z^{(l)})$$

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

$$\mathbb{E}[\hat{f}] = \mathbb{E}[f]$$

$$\text{var}[\hat{f}] = \frac{1}{L}\mathbb{E}\left[(f(z) - \mathbb{E}[f])^2\right]$$

- Approximation not dependent on dimensionality of z
- Variance of estimator shrinks with number of samples

Basic Sampling Ch 11.1 [Bishop, 2006]

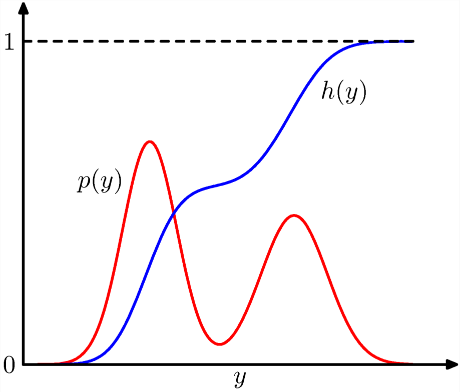
$$z^{(l)} \sim p(z)$$

- Lets assume that we can get uniformly random numbers $z \sim \text{Uniform}(0, 1)$
- A computer cannot, but lets assume it could
- Idea: can we transform this uniform distribution to something interesting
- If we could then we could use samples from z

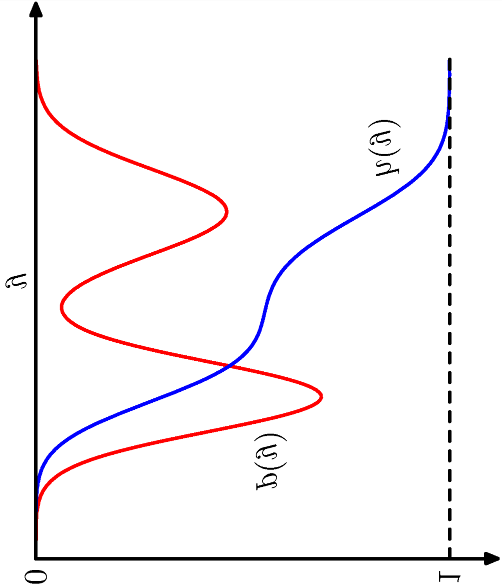
$$z \sim \text{Uniform}(0,1)$$

- We have access to a uniformly distributed variable z
- Change of variable
$$y = f(z)$$
- Idea: can we find $f(z)$ such that it induces $p(y)$ to be the distribution that we want?

Basic Probabilities

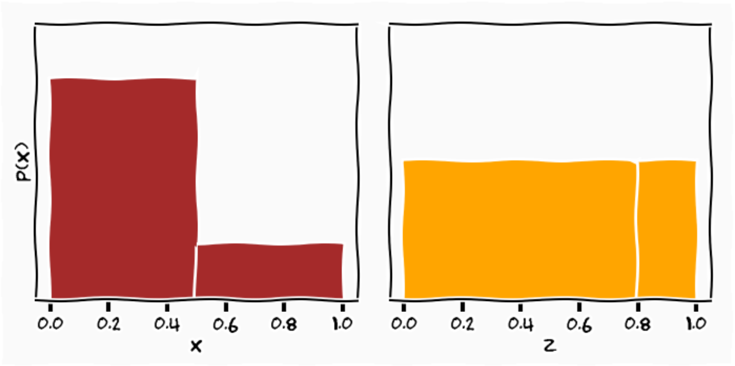


$$z = f^{-1}(y) = \int_{-\infty}^{y} p(y)\mathrm{d}y$$



y
$p(y)$
$h(y)$
0
1

Change of Variables



Sampling

- We know how to transform samples from uniform to any distribution we can formulate the cummulative distribution
- Can we sample from distrubitons we do not know the form of?

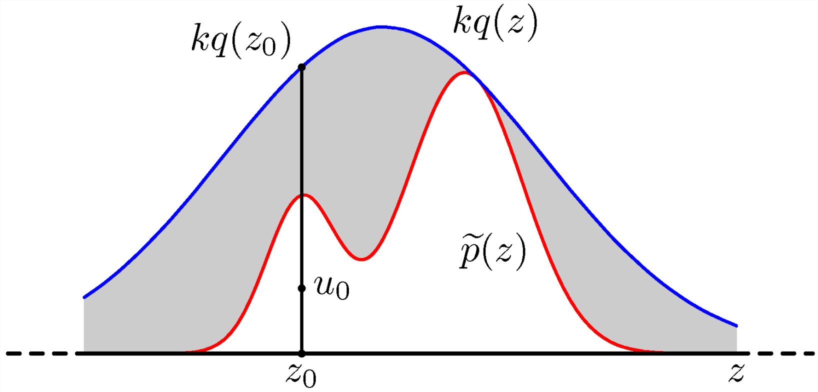
    1. Rejection Sampling
    2. Importance Sampling
    3. Markov Chain Monte Carlo

Rejection Sampling Ch 11.1.2 [Bishop, 2006]

$$p(\mathbf{z}) = \frac{1}{Z}\tilde{p}(\mathbf{z})$$

- $p(\mathbf{z})$ is a distribution of unknown form
- We can evaluate $\tilde{p}(\mathbf{z})$
- Can we draw samples from a simpler distribution and transform them?

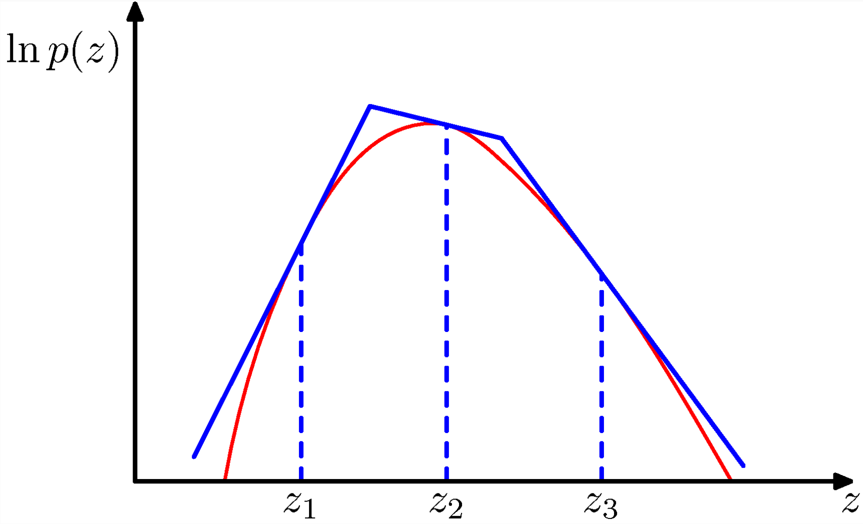
Rejection Sampling



Rejection Sampling

1. Pick approximate distribution $q(\mathbf{z})$
2. Pick constant k such that $k \cdot q(\mathbf{z}) \geq \tilde{p}(\mathbf{z})$
3. Pick random location $\mathbf{z}_0 \sim q(\mathbf{z})$
4. Pick random number $u_0 \sim \text{Uniform}(0, k \cdot q(\mathbf{z}_0))$
5. If $u_0 > \tilde{p}(\mathbf{z}_0)$ reject z_0 otherwise retain

Adaptive Rejection Sampling



Rejection Sampling

- Basic sampling allows us to draw samples from known distributions
- We can use these distributions as *proposal distributions*
- If bound is small we will get an efficient sampler
- Generally works well in few dimensions but do not scale
- We reject too many samples

Importance Sampling Ch 11.1.4 [Bishop, 2006]

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

Importance Sampling Ch 11.1.4 [Bishop, 2006]

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

Importance Sampling Ch 11.1.4 [Bishop, 2006]

$$
\begin{aligned}
\mathbb{E}_{p(\mathbf{z})}[f] &= \int f(\mathbf{z})p(\mathbf{z})\mathrm{d}\mathbf{z} = \int f(\mathbf{z})p(\mathbf{z})\frac{q(\mathbf{z})}{q(\mathbf{z})}\mathrm{d}\mathbf{z} \\
&= \int f(\mathbf{z})\frac{p(\mathbf{z})}{q(\mathbf{z})}q(\mathbf{z})\mathrm{d}\mathbf{z}
\end{aligned}
$$

Importance Sampling Ch 11.1.4 [Bishop, 2006]

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

Importance Sampling Ch 11.1.4 [Bishop, 2006]

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

Importance Sampling Ch 11.1.4 [Bishop, 2006]

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

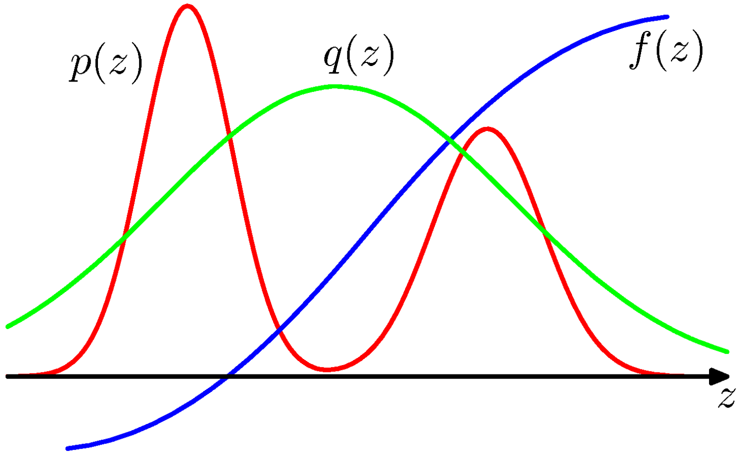
Importance Sampling

$$\mathbb{E}_{p(\mathbf{z})}[f] \approx \frac{1}{L}\sum_{l=1}^{L} r_l \cdot f(\mathbf{z}^{(l)})$$

$$\mathbf{z}^{(l)} \sim q(\mathbf{z}), \qquad r_l = \frac{p(\mathbf{z}^{(l)})}{q(\mathbf{z}^{(l)})}$$

- Directly approximate expectation
- Accepts all samples
- r_l corrects bias in sampling from wrong distribution

Importance Sampling



Importance Sampling

$$p(\mathbf{z}) = \frac{1}{Z_p}\tilde{p}(\mathbf{z}), \qquad\qquad q(\mathbf{z}) = \frac{1}{Z_q}\tilde{q}(\mathbf{z})$$

- Often it will not be possible to evaluate $p(\mathbf{z})$ and maybe not even $q(\mathbf{z})$

$$\mathbb{E}[f] = \frac{Z_q}{Z_p}\int f(\mathbf{z})\frac{\tilde{p}(\mathbf{z})}{\tilde{q}(\mathbf{z})}q(\mathbf{z})\mathrm{d}\mathbf{z} \approx \frac{Z_q}{Z_p}\frac{1}{L}\sum_{l=1}^{L}\tilde{r}_l \cdot f(\mathbf{z}^{(l)})$$

Importance Sampling

$$\frac{Z_p}{Z_q} = \frac{1}{Z_q} \int \tilde{p}(\mathbf{z}) \mathrm{d}\mathbf{z}$$

Importance Sampling

$$\frac{Z_p}{Z_q} = \frac{1}{Z_q}\int \tilde{p}(\mathbf{z})\mathrm{d}\mathbf{z} = \frac{1}{Z_q}\int \tilde{p}(\mathbf{z})\frac{q(\mathbf{z})}{q(\mathbf{z})}\mathrm{d}\mathbf{z}$$

Importance Sampling

$$
\begin{aligned}
\frac{Z_p}{Z_q} &= \frac{1}{Z_q}\int \tilde{p}(\mathbf{z})\mathrm{d}\mathbf{z} = \frac{1}{Z_q}\int \tilde{p}(\mathbf{z})\frac{q(\mathbf{z})}{q(\mathbf{z})}\mathrm{d}\mathbf{z} \\
&= \frac{1}{Z_q}\int \tilde{p}(\mathbf{z})\frac{q(\mathbf{z})}{\frac{1}{Z_q}\tilde{q}(\mathbf{z})}\mathrm{d}\mathbf{z}
\end{aligned}
$$

Importance Sampling

$$
\begin{aligned}
\frac{Z_p}{Z_q} &= \frac{1}{Z_q}\int \tilde{p}(\mathbf{z})\mathrm{d}\mathbf{z} = \frac{1}{Z_q}\int \tilde{p}(\mathbf{z})\frac{q(\mathbf{z})}{q(\mathbf{z})}\mathrm{d}\mathbf{z} \\
&= \frac{1}{Z_q}\int \tilde{p}(\mathbf{z})\frac{q(\mathbf{z})}{\frac{1}{Z_q}\tilde{q}(\mathbf{z})}\mathrm{d}\mathbf{z} = \int \frac{\tilde{p}(\mathbf{z})}{\tilde{q}(\mathbf{z})}q(\mathbf{z})\mathrm{d}\mathbf{z}
\end{aligned}
$$

Importance Sampling

$$\begin{aligned}\frac{Z_p}{Z_q} &= \frac{1}{Z_q}\int \tilde{p}(\mathbf{z})\mathrm{d}\mathbf{z} = \frac{1}{Z_q}\int \tilde{p}(\mathbf{z})\frac{q(\mathbf{z})}{q(\mathbf{z})}\mathrm{d}\mathbf{z} \\ &= \frac{1}{Z_q}\int \tilde{p}(\mathbf{z})\frac{q(\mathbf{z})}{\frac{1}{Z_q}\tilde{q}(\mathbf{z})}\mathrm{d}\mathbf{z} = \int \frac{\tilde{p}(\mathbf{z})}{\tilde{q}(\mathbf{z})}q(\mathbf{z})\mathrm{d}\mathbf{z} \\ &\approx \frac{1}{L}\sum_{l=1}^{L}\frac{\tilde{p}(\mathbf{z}^{(l)})}{\tilde{q}(\mathbf{z}^{(l)})}\end{aligned}$$

Importance Sampling

$$
\begin{aligned}
\frac{Z_p}{Z_q} &= \frac{1}{Z_q}\int \tilde{p}(\mathbf{z})\mathrm{d}\mathbf{z} = \frac{1}{Z_q}\int \tilde{p}(\mathbf{z})\frac{q(\mathbf{z})}{q(\mathbf{z})}\mathrm{d}\mathbf{z} \\
&= \frac{1}{Z_q}\int \tilde{p}(\mathbf{z})\frac{q(\mathbf{z})}{\frac{1}{Z_q}\tilde{q}(\mathbf{z})}\mathrm{d}\mathbf{z} = \int \frac{\tilde{p}(\mathbf{z})}{\tilde{q}(\mathbf{z})}q(\mathbf{z})\mathrm{d}\mathbf{z} \\
&\approx \frac{1}{L}\sum_{l=1}^{L}\frac{\tilde{p}(\mathbf{z}^{(l)})}{\tilde{q}(\mathbf{z}^{(l)})} = \frac{1}{L}\sum_{l=1}^{L} r_l
\end{aligned}
$$

Importance Sampling

$$
\begin{aligned}
\frac{Z_p}{Z_q} &= \frac{1}{Z_q}\int \tilde{p}(\mathbf{z})\mathrm{d}\mathbf{z} = \frac{1}{Z_q}\int \tilde{p}(\mathbf{z})\frac{q(\mathbf{z})}{q(\mathbf{z})}\mathrm{d}\mathbf{z} \\
&= \frac{1}{Z_q}\int \tilde{p}(\mathbf{z})\frac{q(\mathbf{z})}{\frac{1}{Z_q}\tilde{q}(\mathbf{z})}\mathrm{d}\mathbf{z} = \int \frac{\tilde{p}(\mathbf{z})}{\tilde{q}(\mathbf{z})}q(\mathbf{z})\mathrm{d}\mathbf{z} \\
&\approx \frac{1}{L}\sum_{l=1}^{L}\frac{\tilde{p}(\mathbf{z}^{(l)})}{\tilde{q}(\mathbf{z}^{(l)})} = \frac{1}{L}\sum_{l=1}^{L} r_l
\end{aligned}
$$

- Not very surprising can we take the average ratio between the unormalised functions to get the normalisers
- We can use the same samples

Importance Sampling

$$\mathbb{E}[f] \approx \frac{Z_q}{Z_p} \frac{1}{L} \sum_{l=1}^{L} r_l f(\mathbf{z}^{(l)})$$

Importance Sampling

$$\mathbb{E}[f] \approx \frac{Z_q}{Z_p} \frac{1}{L} \sum_{l=1}^{L} r_l f(\mathbf{z}^{(l)}) = \frac{1}{\frac{1}{L}\sum_{l=1}^{L} r_l} \frac{1}{L} \sum_{l=1}^{L} r_l f(\mathbf{z}^{(l)})$$

Importance Sampling

$$\mathbb{E}[f] \approx \frac{Z_q}{Z_p}\frac{1}{L}\sum_{l=1}^{L} r_l f(\mathbf{z}^{(l)}) = \frac{1}{\frac{1}{L}\sum_{l=1}^{L} r_l}\frac{1}{L}\sum_{l=1}^{L} r_l f(\mathbf{z}^{(l)})$$

$$= \sum_{l=1}^{L} \frac{r_l}{\sum_{k=1}^{L} r_k} f(\mathbf{z}^{(l)}) = \sum_{l=1}^{L} w_l f(\mathbf{z}^{(l)})$$

Importance Sampling

- More efficient compared to Rejection sampling as it uses all samples
- Hard to know how well you are doing
- We want to make sure that the importance weights are of small variance
  - $q(\mathbf{z})$ should not be small where $p(\mathbf{z})$ is large
- Will work wonders if $q(\mathbf{z})$ is good

Markov Chain Monte Carlo

Markov Chain Monte Carlo Ch 11.2 [Bishop, 2006]

- Sample from a proposal distribution
- Remembers the state and samples from a conditional
- Can lead to much better exploration of the space

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 $A(\mathbf{z}^*, \mathbf{z}^{(0)}) > u \rightarrow \mathbf{z}^{(1)} = \mathbf{z}^*$

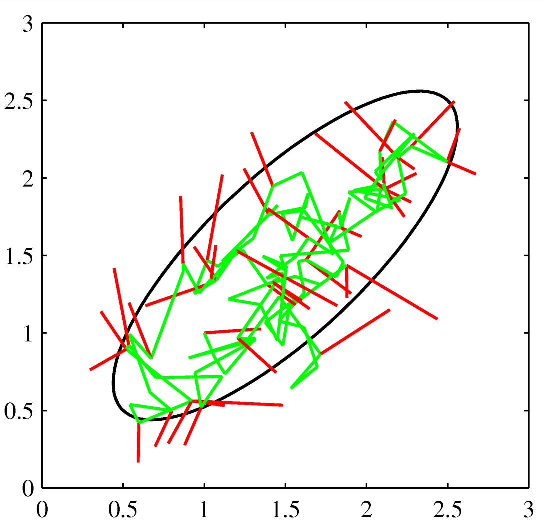
Metropolis Sampling

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

Metropolis Gaussian



Gibbs Sampling Ch 11.3 [Bishop, 2006]

- Often 1D samples are easy to get
- Gibbs sampling exploits this to create a very simple Markov Chain
- Sample each variable in turn conditioned on the others and cycle through
- Each variable depends only on its Markov blanket so conditionals can be very simple

Gibbs Sampling

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

Gibbs Sampling

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

Gibbs Sampling

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. 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. 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

Why is this easier?

Multivariate case

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}, \mathbf{x})}{p(\mathbf{y})}$$
$$p(\mathbf{y}) = \sum_i p(\mathbf{y}|\mathbf{x}^{(i)})p(\mathbf{x}^{(i)})$$

Why is this easier?

Multivariate case

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

$$p(\mathbf{y}) = \sum_i p(\mathbf{y}|\mathbf{x}^{(i)})p(\mathbf{x}^{(i)})$$

1D case

$$p(x_i|\mathbf{x}_{\neg i},\mathbf{y}) = \frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{x}_{\neg i},\mathbf{y})}$$

Why is this easier?

Multivariate case

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

$$p(\mathbf{y}) = \sum_i p(\mathbf{y}|\mathbf{x}^{(i)})p(\mathbf{x}^{(i)})$$

1D case

$$p(x_i|\mathbf{x}_{\neg i}, \mathbf{y}) = \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x}_{\neg i}, \mathbf{y})}$$

$$p(\mathbf{x}_{\neg i}, \mathbf{y}) = \int p(\mathbf{x}, \mathbf{y})\mathrm{d}x_i = \sum_{x_i \in [1, -1]} p(x_i, \mathbf{x}_{\neg i}, \mathbf{y})$$

$$= p(x_i = 1, \mathbf{x}_{\neg i}, \mathbf{y}) + p(x_i = -1, \mathbf{x}_{\neg i}, \mathbf{y})$$

Summary

Summary

- Using sampling we can approximate tricky integrals by computing samples from distributions we do not know
- Sampling is a bit of a black-art
- Often exact given infinite time
- Generally works but often time consuming

eof

References

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

Kingma, D. P. and Welling, M. (2014).
Auto-encoding variational Bayes.
In *Proceedings of the International Conference on Learning Representations*.