https://xkcd.com/780/

Sampling can be hard to get right ...

Class update:
Exam info posted. Friday "week 13". 3 June.

Week 12 Monday - no lecture, finish your video assignment + good luck with all end-of-S1 deadlines!
Week 12 Wednesday - Q&A, Talk about selected questions in 2021 Exam.
"Week 13" – drop-in sessions will be announced.

Sampling methods

Motivation, why? Sampling and ML

Basic sampling algorithms

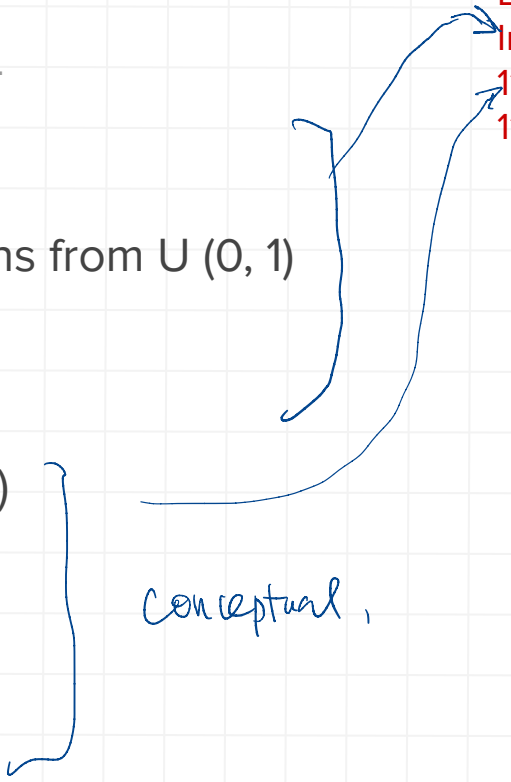
- Sampling standard distributions from U (0, 1)
- Rejection sampling
- Importance sampling

Markov chain Monte Carlo (MCMC)

- Markov chains
- Metropolis-Hasting

Gibbs sampling 11.3

Bishop Chap 11
Intro, 11.1.1, 11.1.2, 11.1.4, 11.1.6
11.2, 11.2.1, 11.2.2, 11.2.3
11.3

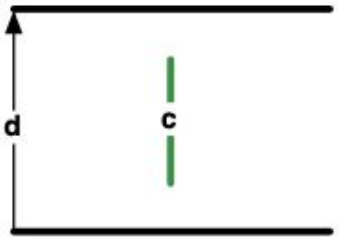


Estimating π – Buffon's needle

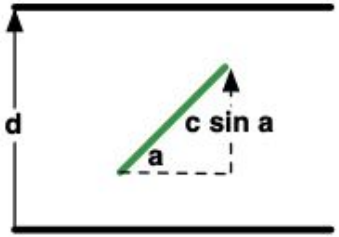
- Drop length c needle on parallel lines distance d apart
- Needle falls perpendicular:
  Probability of crossing the line is c/d.
- Needle falls at an arbitrary angle a:
  Probability of crossing the line $c\sin(a)/d$.
- Every angle is equally probable. Calculate the mean:

$$p(\text{crossing}) = \frac{c}{d}\int_0^{\pi} \sin(a)\, dp(a) = \frac{1}{\pi}\frac{c}{d}\int_0^{\pi} \sin(a)\, da = \frac{2}{\pi}\frac{c}{d}$$

(iv) n crossings in N experiments results in $\frac{n}{N} \approx \frac{2}{\pi}\frac{c}{d}$



(i) Needle falls perpendicular ($a = \pi/2$).



(ii) Needle falls at arbitrary angle a.

- $N \to \infty$ estimate "correct"
- how fast?

https://mathworld.wolfram.com/BuffonsNeedleProblem.html

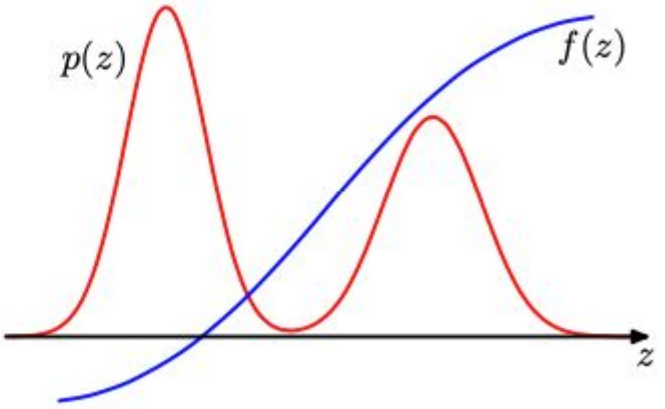
Why do we need sampling?

Distributions can be quite complex, e.g. posterior distributions, mixture distributions, graphical models ~~(coming next)~~.

Many ML tasks can be stated as estimating the expectation of functions under a distribution. e.g. posterior mean and variance, moments.

$$\mathbb{E}[f] = \int f(\mathbf{z})p(\mathbf{z})\,\mathrm{d}\mathbf{z} \qquad (11.1)$$

Figure 11.1 Schematic illustration of a function $f(z)$ whose expectation is to be evaluated with respect to a distribution $p(z)$.



Bayesian logistic regression

(GP2)

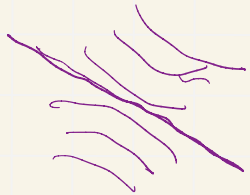
$$p(\mathcal{C}_1|\boldsymbol{\phi}) = y(\boldsymbol{\phi}) = \sigma\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}\right) \tag{4.87}$$

LR.

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}) = -\sum_{n=1}^{N}\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\} \tag{4.90}$$

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0) \tag{4.140}$$

→ prior



$$p(\mathbf{w}|\mathbf{t}) \propto p(\mathbf{w})p(\mathbf{t}|\mathbf{w}) \tag{4.141}$$

$$\begin{aligned}\ln p(\mathbf{w}|\mathbf{t}) &= -\frac{1}{2}(\mathbf{w}-\mathbf{m}_0)^{\mathrm{T}}\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0) \\ &\quad + \sum_{n=1}^{N}\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\} + \text{const}\end{aligned} \tag{4.142}$$

$$\mathbf{S}_N = -\nabla\nabla \ln p(\mathbf{w}|\mathbf{t}) = \mathbf{S}_0^{-1} + \sum_{n=1}^{N} y_n(1-y_n)\boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\mathrm{T}}. \tag{4.143}$$

Laplace approximation

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{w}_{\mathrm{MAP}}, \mathbf{S}_N). \tag{4.144}$$

$$p(\mathcal{C}_1|\boldsymbol{\phi}, \mathbf{t}) = \int p(\mathcal{C}_1|\boldsymbol{\phi}, \mathbf{w})p(\mathbf{w}|\mathbf{t})\,\mathrm{d}\mathbf{w} \simeq \int \sigma(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi})q(\mathbf{w})\,\mathrm{d}\mathbf{w} \tag{4.145}$$

$$= \int \sigma(a)\mathcal{N}(a|\mu_a, \sigma_a^2)\,\mathrm{d}a.$$

sample.

$$\mathbb{E}[f] = \int f(\mathbf{z})p(\mathbf{z})\,\mathrm{d}\mathbf{z} \tag{11.1}$$

Sampling 101

Goal: estimate $$\mathbb{E}[f] = \int f(\mathbf{z}) p(\mathbf{z}) \, \mathrm{d}\mathbf{z} \tag{11.1}$$

sample from $p(z)$ i.i.d.

Simply

$$\widehat{f} = \frac{1}{L} \sum_{l=1}^{L} f(\mathbf{z}^{(l)}). \tag{11.2}$$

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

$$\mathrm{var}[\widehat{f}] = \frac{1}{L} \mathbb{E}\left[(f - \mathbb{E}[f])^2 \right] \tag{11.3}$$

"effective # of samples"

The good:

- the accuracy of the estimator does not depend on the dimensionality of z
- in principle, high accuracy may be achievable with a relatively small number of samples

The bad:

- samples $\{z^{(l)}\}$ might not be independent, and so the effective sample size might be much smaller than the apparent sample size.

The good:

- the accuracy of the estimator does not depend on the dimensionality of z
- in principle, high accuracy may be achievable with a relatively small number of samples

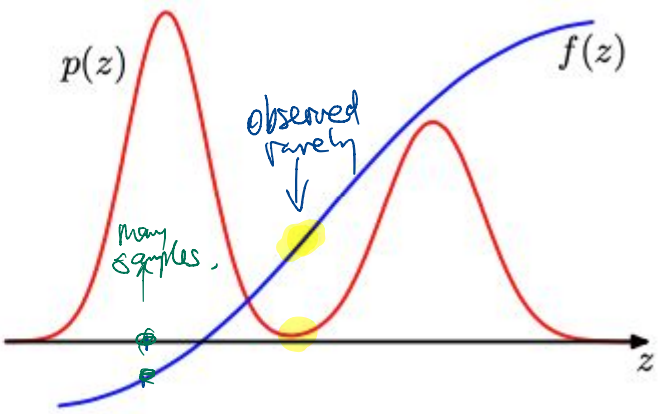
The bad:

- samples $\{z^{(l)}\}$ might not be independent, and so the effective sample size might be much smaller than the apparent sample size.

The ugly:

f(z) is small in regions where p(z) is large, and vice versa, then the expectation may be dominated by regions of small probability, implying that relatively large sample sizes will be required to achieve sufficient accuracy.

Figure 11.1 Schematic illustration of a function $f(z)$ whose expectation is to be evaluated with respect to a distribution $p(z)$.



Sampling from uniform distributions

- In a computer usually via pseudorandom number generator : an algorithm generating a sequence of numbers that approximates the properties of random numbers.
- Use a mathematically well crafted pseudorandom number generator.
- From now on we will assume that we have a good pseudorandom number generator for uniformly distributed data available.
- If you don't trust any algorithm :
  Three carefully adjusted radio receivers picking up atmospheric noise to provide real random numbers at `http://www.random.org/`

For this class: assume we can generate $z \sim U(0, 1)$

Q: given $z \sim U(0, 1)$
How do you obtain $z \sim U(a, b)$?

Sampling a distribution p(y)

$$p(y) = p(z)\left|\frac{dz}{dy}\right| \tag{11.5}$$

$$z = h(y) \equiv \int_{-\infty}^{y} p(\widehat{y})\, d\widehat{y} \tag{11.6}$$

Steps:

- Generate z ~ U(0, 1)
- Use inverse CDF $y=h^{-1}(z)$ to obtain y

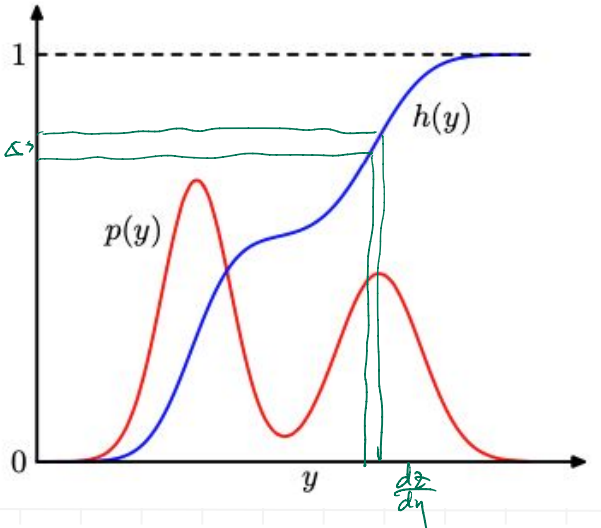
Multiple variables

$$p(y_1, \ldots, y_M) = p(z_1, \ldots, z_M)\left|\frac{\partial(z_1, \ldots, z_M)}{\partial(y_1, \ldots, y_M)}\right|. \tag{11.9}$$

iid U(0,1)

determinant of Jacobian $M \times M$ ij $\frac{\partial z_i}{\partial y_j}$

Figure 11.2 Geometrical interpretation of the transformation method for generating nonuniformly distributed random numbers. $h(y)$ is the indefinite integral of the desired distribution $p(y)$. If a uniformly distributed random variable z is transformed using $y = h^{-1}(z)$, then y will be distributed according to $p(y)$.



Example: exponential distribution

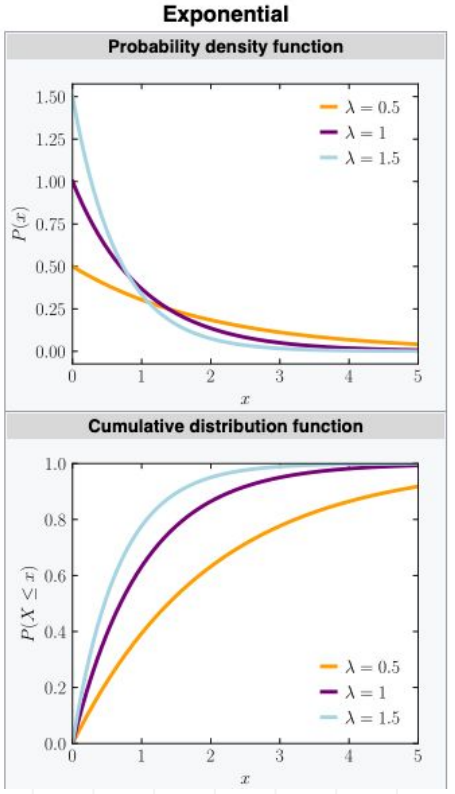
$$p(y) = \lambda \exp(-\lambda y) \qquad (11.7) \qquad 0 \leqslant y < \infty.$$

cdf $\quad z = h(y) = 1 - \exp(-\lambda y)$

$1 - z = \exp^{-\lambda y}$

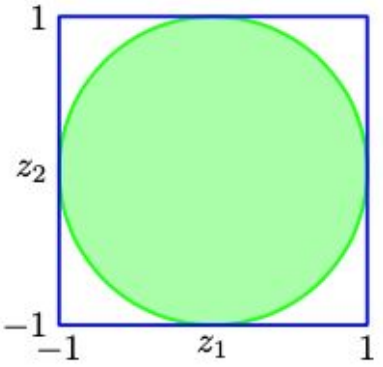
$\log(1-z) = -\lambda y.$

$$y = -\lambda^{-1} \ln(1 - z)$$



https://en.wikipedia.org/wiki/Exponential_distribution

Figure 11.3 The Box-Muller method for generating Gaussian distributed random numbers starts by generating samples from a uniform distribution inside the unit circle.



$r^2 = z_1^2 + z_2^2$

Typo in the book (online pdf version, fixed in the latest print version)

$$y_1 = z_1 \left(\frac{-2 \ln z_1}{r^2} \right)^{1/2} \qquad (11.10)$$

$$y_2 = z_2 \left(\frac{-2 \ln z_2}{r^2} \right)^{1/2} \qquad (11.11)$$

$$y_1 = z_1 \left(\frac{-2 \ln r^2}{r^2} \right)^{1/2}$$

$$y_2 = z_2 \left(\frac{-2 \ln r^2}{r^2} \right)^{1/2}$$

$$p(y_1, y_2) = p(z_1, z_2) \left| \frac{\partial(z_1, z_2)}{\partial(y_1, y_2)} \right|$$

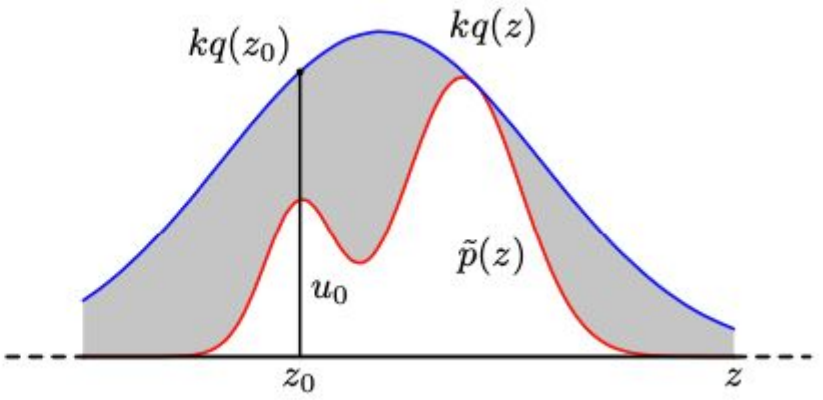
$$= \left[\frac{1}{\sqrt{2\pi}} \exp(-y_1^2/2) \right] \left[\frac{1}{\sqrt{2\pi}} \exp(-y_2^2/2) \right] \qquad (11.12)$$

Full derivation: through change of variable to the polar coordinates (r, **θ**), somewhat involved

$r \sim U(0,1)$, $\theta \sim U(0, 2\pi)$

Rejection sampling

Figure 11.4 In the rejection sampling method, samples are drawn from a simple distribution $q(z)$ and rejected if they fall in the grey area between the unnormalized distribution $\widetilde{p}(z)$ and the scaled distribution $kq(z)$. The resulting samples are distributed according to $p(z)$, which is the normalized version of $\widetilde{p}(z)$.



Assumption 1 : Sampling directly from $p(z)$ is difficult, but we can evaluate $p(z)$ up to some unknown normalisation constant Z_p

$$p(z) = \frac{1}{Z_p}\widetilde{p}(z)$$

Assumption 2 : We can draw samples from a simpler distribution $q(z)$ and for some constant k and all z holds

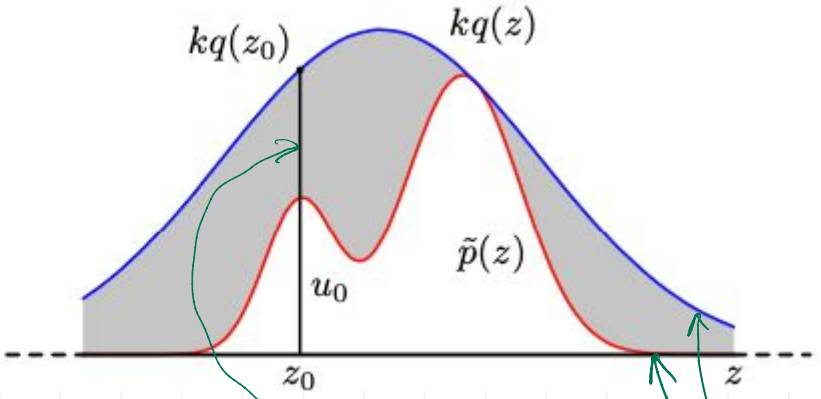
$$kq(z) \geq \widetilde{p}(z)$$

Impossible:

$q(z) \geq p(z) \quad \forall z$

Rejection sampling

Figure 11.4 In the rejection sampling method, samples are drawn from a simple distribution $q(z)$ and rejected if they fall in the grey area between the unnormalized distribution $\widetilde{p}(z)$ and the scaled distribution $kq(z)$. The resulting samples are distributed according to $p(z)$, which is the normalized version of $\widetilde{p}(z)$.



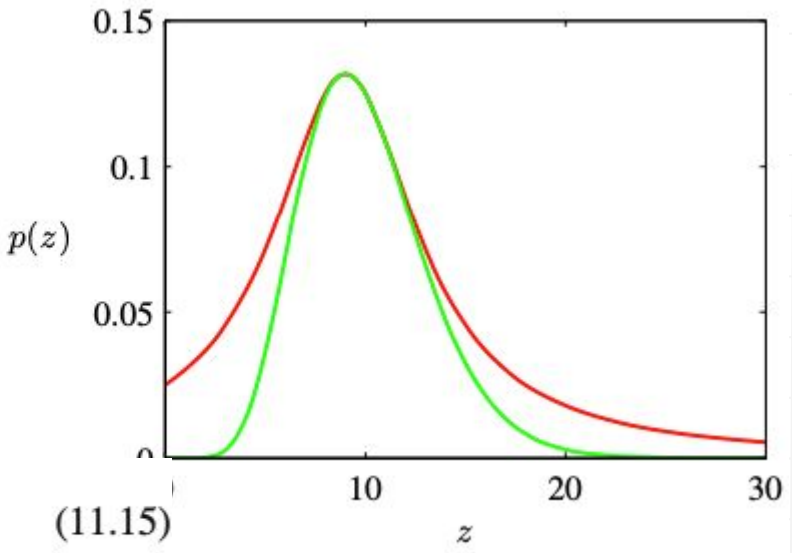
1. Generate a random number z_0 from the distribution $q(z)$.
2. Generate a number from the u_0 from the uniform distribution over $[0, k\,q(z_0)]$.
3. If $u_0 > \widetilde{p}(z_0)$ then reject the pair (z_0, u_0).
4. The remaining pairs have uniform distribution under the curve $\widetilde{p}(z)$.
5. The z values are distributed according to $p(z)$.

reject

$q(z)$ should have "heavier" tails than $p(z)$

Rejection sampling example

Figure 11.5 Plot showing the gamma distribution given by (11.15) as the green curve, with a scaled Cauchy proposal distribution shown by the red curve. Samples from the gamma distribution can be obtained by sampling from the Cauchy and then applying the rejection sampling criterion.



Sample Gamma distribution with a>1

$$\mathrm{Gam}(z|a,b) = \frac{b^a z^{a-1} \exp(-bz)}{\Gamma(a)} \tag{11.15}$$

Proposal distribution: Cauchy $p(y) = \frac{1}{\pi}\frac{1}{1+y^2}.$ (11.8)

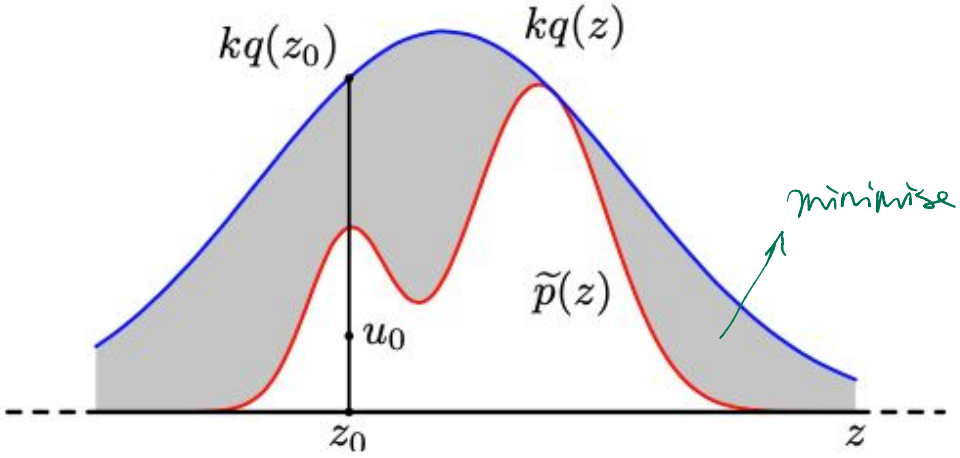
Or, $$q(z) = \frac{k}{1 + (z-c)^2/b^2}$$

u~ U[0, 1]
z = b tan *u +c*

Rejection sampling - challenges

- Need to find a proposal distribution $q(z)$ which is a close upper bound to $p(z)$; otherwise many samples are rejected.
- Curse of dimensionality for multivariate distributions.

Proposal distribution need to have heavier tails than the target distribution.



Acceptance rate *can* decreases exponentially as dimensionality increases.

minimise area in green

Importance sampling

Desired: $$\mathbb{E}[f] \simeq \sum_{l=1}^{L} p(\mathbf{z}^{(l)}) f(\mathbf{z}^{(l)}). \tag{11.18}$$

- Directly calculate $\mathbb{E}_p[f(z)]$ w.r.t. some $p(z)$.
- Does not sample $p(z)$ as an intermediate step.
- Again use samples $z^{(l)}$ from some proposal $q(z)$.

$$\begin{aligned}
\mathbb{E}[f] &= \int f(\mathbf{z}) p(\mathbf{z}) \, \mathrm{d}\mathbf{z} \\
&= \int f(\mathbf{z}) \frac{p(\mathbf{z})}{q(\mathbf{z})} q(\mathbf{z}) \, \mathrm{d}\mathbf{z} \\
&\simeq \frac{1}{L} \sum_{l=1}^{L} \frac{p(\mathbf{z}^{(l)})}{q(\mathbf{z}^{(l)})} f(\mathbf{z}^{(l)}).
\end{aligned} \tag{11.19}$$

All samples retained.

Importance weights

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

r_l can be > 1 :)

Estimating the (ratio of) normalization constants

Easy to evaluate distribution up to a normalising constant, but hard to know what the constant is. $p(\mathbf{z}) = \widetilde{p}(\mathbf{z})/Z_p$ $q(\mathbf{z}) = \widetilde{q}(\mathbf{z})/Z_q$

simple.

$$\begin{aligned}\mathbb{E}[f] &= \int f(\mathbf{z})p(\mathbf{z})\,\mathrm{d}\mathbf{z} \\ &= \frac{Z_q}{Z_p}\int f(\mathbf{z})\frac{\widetilde{p}(\mathbf{z})}{\widetilde{q}(\mathbf{z})}q(\mathbf{z})\,\mathrm{d}\mathbf{z} \\ &\simeq \frac{Z_q}{Z_p}\frac{1}{L}\sum_{l=1}^{L}\widetilde{r}_l f(\mathbf{z}^{(l)}). \qquad (11.20)\end{aligned}$$

$\widetilde{r}_l = r_l = \frac{\widetilde{p}(z_l)}{\widetilde{q}(z_l)}$

$\rightarrow \frac{Z_p}{Z_q} r_l$

$$\begin{aligned}\frac{Z_p}{Z_q} &= \frac{1}{Z_q}\int \widetilde{p}(\mathbf{z})\,\mathrm{d}\mathbf{z} = \int \frac{\widetilde{p}(\mathbf{z})}{\widetilde{q}(\mathbf{z})}q(\mathbf{z})\,\mathrm{d}\mathbf{z} \\ &\simeq \frac{1}{L}\sum_{l=1}^{L}\widetilde{r}_l \qquad (11.21)\end{aligned}$$

→ average of unnormalized importance weights!

Using the same samples!

Importance sampling - comments

- Importance weights r_l correct the bias introduced by sampling from the proposal distribution $q(z)$ instead of the wanted distribution $p(z)$.
- Success depends on how well $q(z)$ approximates $p(z)$.
- If $p(z) > 0$ in same region, then $q(z) > 0$ necessary.

! support of $q(z)$ is a superset of support of $p(z)$

effective sample size can be much smaller than the apparent sample size L. The problem is even more severe if none of the samples falls in the regions where $p(\mathbf{z})f(\mathbf{z})$ is large. In that case, the apparent variances of r_l and $r_l f(\mathbf{z}^{(l)})$ may be small even though the estimate of the expectation may be severely wrong. Hence a major drawback of the importance sampling method is the potential to produce results that are arbitrarily in error and with no diagnostic indication. This also highlights a key requirement for the sampling distribution $q(\mathbf{z})$, namely that it should not be small or zero in regions where $p(\mathbf{z})$ may be significant.

Sampling and the EM algorithm

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \int p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{Z}, \mathbf{X}|\boldsymbol{\theta}) \, \mathrm{d}\mathbf{Z}. \tag{11.28}$$

Sample this

Monte Carlo EM

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) \simeq \frac{1}{L} \sum_{l=1}^{L} \ln p(\mathbf{Z}^{(l)}, \mathbf{X}|\boldsymbol{\theta}). \tag{11.29}$$

M-step: as usual,

Being Bayesian in EM

IP Algorithm

I-step. We wish to sample from $p(\mathbf{Z}|\mathbf{X})$ but we cannot do this directly. We therefore note the relation

$$p(\mathbf{Z}|\mathbf{X}) = \int p(\mathbf{Z}|\boldsymbol{\theta}, \mathbf{X}) p(\boldsymbol{\theta}|\mathbf{X}) \, \mathrm{d}\boldsymbol{\theta} \tag{11.30}$$

and hence for $l = 1, \ldots, L$ we first draw a sample $\boldsymbol{\theta}^{(l)}$ from the current estimate for $p(\boldsymbol{\theta}|\mathbf{X})$, and then use this to draw a sample $\mathbf{Z}^{(l)}$ from $p(\mathbf{Z}|\boldsymbol{\theta}^{(l)}, \mathbf{X})$.

P-step. Given the relation

$$p(\boldsymbol{\theta}|\mathbf{X}) = \int p(\boldsymbol{\theta}|\mathbf{Z}, \mathbf{X}) p(\mathbf{Z}|\mathbf{X}) \, \mathrm{d}\mathbf{Z} \tag{11.31}$$

we use the samples $\{\mathbf{Z}^{(l)}\}$ obtained from the I-step to compute a revised estimate of the posterior distribution over $\boldsymbol{\theta}$ given by

$$p(\boldsymbol{\theta}|\mathbf{X}) \simeq \frac{1}{L} \sum_{l=1}^{L} p(\boldsymbol{\theta}|\mathbf{Z}^{(l)}, \mathbf{X}). \tag{11.32}$$

By assumption, it will be feasible to sample from this approximation in the I-step.

Sampling methods

Motivation, why? Sampling and ML

Basic sampling algorithms

- Sampling standard distributions from U (0, 1)
- Rejection sampling
- Importance sampling

Markov chain Monte Carlo (MCMC)

- Markov chains
- Metropolis-Hasting

Gibbs sampling

assume: ① can evaluate $p(z)$ for any given z.

② evaluate $p(z)|_{z=z_0}$ is easier than sampling $z \sim p(z)$ ↑ involves inverse CDF

Markov Chain Monte Carlo (MCMC)

$E_p[f(z)]$

Goal: sample from $p(z)$.

Generate a sequence using a Markov chain.

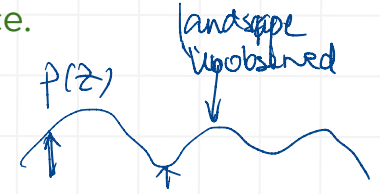
1. Generate a new sample $z^\star \sim q(z \mid z^{(l)})$, conditional on the previous sample $z^{(l)}$.
2. Accept or reject the new sample according to some appropriate criterion.

$$z^{(l+1)} = \begin{cases} z^\star & \text{if accepted} \\ z^{(l)} & \text{if rejected} \end{cases}$$

— copy of the prev value.

3. For an appropriate proposal and corresponding acceptance criterion, as $l \to \infty$, $z^{(l)}$ approaches an independent sample of $p(z)$.

Motivation:
Cover high-probability regions of p(z) -- can we move towards it?
Scale better with the dimensionality of the sample space.



Metropolis Algorithm

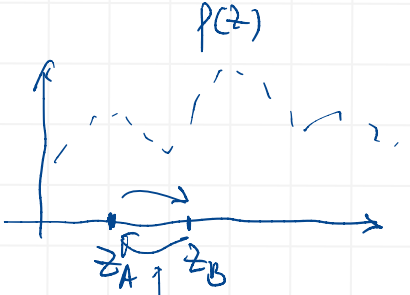
1. Choose a symmetric proposal distribution $q(z_A \mid z_B) = q(z_B \mid z_A)$.
2. Accept the new sample $z^\star$ with probability

$$A(z^\star, z^{(r)}) = \min\left(1, \frac{\widetilde{p}(z^\star)}{\widetilde{p}(z^{(r)})}\right)$$

e.g., let $u \sim \text{Uniform}(0,1)$ and accept if $\frac{\widetilde{p}(z^\star)}{\widetilde{p}(z^{(r)})} > u$.

3. Unlike rejection sampling we include the previous sample on rejection of the proposal:

$$z^{(r+1)} = \begin{cases} z^\star & \text{if accepted} \\ z^{(r)} & \text{if rejected} \end{cases}$$



when $\widetilde{p}(z^\star) > \widetilde{p}(z^{(r)})$ then accept

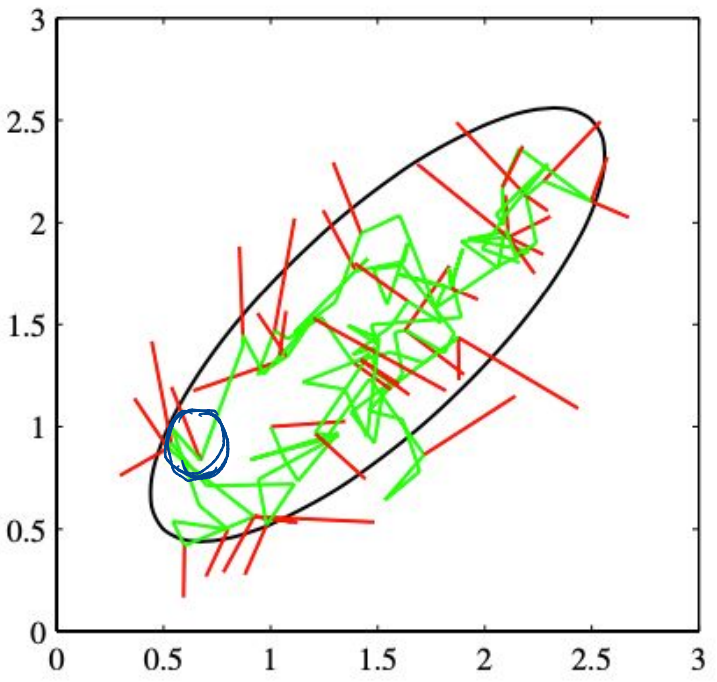
when $\widetilde{p}(z^\star) < \widetilde{p}(z^{(r)})$ accept w. $\frac{\widetilde{p}(z^\star)}{\widetilde{p}(z^{(r)})} < 1$

$q(z_A \mid z_B)$

a copy of $z^{(r)}$

Metropolis Algorithm - illustration

Figure 11.9 A simple illustration using Metropolis algorithm to sample from a Gaussian distribution whose one standard-deviation contour is shown by the ellipse. The proposal distribution is an isotropic Gaussian distribution whose standard deviation is 0.2. Steps that are accepted are shown as green lines, and rejected steps are shown in red. A total of 150 candidate samples are generated, of which 43 are rejected.



Random walk as a Markov Chain

Consider a state space z consisting of the integers, with probabilities

$$p(z^{(\tau+1)} = z^{(\tau)}) = 0.5 \quad (11.34)$$

$$p(z^{(\tau+1)} = z^{(\tau)} + 1) = 0.25 \quad (11.35)$$

$$p(z^{(\tau+1)} = z^{(\tau)} - 1) = 0.25 \quad (11.36)$$

if $z^{(1)} = 0$,

then $\mathbb{E}[z^{(\tau)}] = 0$

$$\mathbb{E}[(z^{(\tau)})^2] = \tau/2.$$

after τ steps, the random walk has only travelled a distance that on average is proportional to the square root of τ

➜ Random walks are very inefficient in exploring the state space.

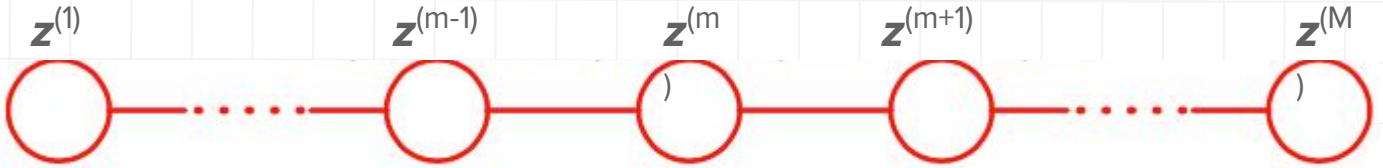
Design goal of MCMC algorithms: avoid random-walk behaviour.

First order Markov Chain

A series of random variable $\mathbf{z}^{(1)}$, … , $\mathbf{z}^{(M)}$ such that

$$p(\mathbf{z}^{(m+1)}|\mathbf{z}^{(1)},\ldots,\mathbf{z}^{(m)}) = p(\mathbf{z}^{(m+1)}|\mathbf{z}^{(m)}). \qquad (11.37)$$

→ chain shaped graphical model



Marginal probability

$$p(z^{(m+1)}) = \sum_{z^{(m)}} p(z^{(m+1)} \mid z^{(m)})\, p(z^{(m)})$$

$$= \sum_{z^{(m)}} T_m(z^{(m)} \mid z^{(m+1)})\, p(z^{(m)})$$

where $T_m(z^{(m)} \mid z^{(m+1)})$ are the transition probabilities.

MCMC - why it works

- Marginal probability

$$p(z^{(m+1)}) = \sum_{z^{(m)}} T_m(z^{(m)} \mid z^{(m+1)})\, p(z^{(m)})$$

drop cons

- A Markov chain is called homogeneous if the transition probabilities are the same for all m, denoted by $T(z', z)$.
- A distribution is invariant, or stationary, with respect to a Markov chain if each step leaves the distribution invariant.
- For a homogeneous Markov chain, the distribution $p^\star(z)$ is invariant if

$$p^\star(\mathbf{z}) = \sum_{\mathbf{z}'} T(\mathbf{z}', \mathbf{z}) p^\star(\mathbf{z}'). \qquad (11.39)$$

(Note: There can be many. If T is the identity matrix, every distribution is invariant.)

↑ we stay put at all values of z.

MCMC - why it works

We want

$$p^\star(\mathbf{z}) = \sum_{\mathbf{z}'} T(\mathbf{z}', \mathbf{z}) p^\star(\mathbf{z}'). \tag{11.39}$$

- Detailed balance

$p(z \to z')$ $p(z' \to z)$

$$p^\star(\mathbf{z}) T(\mathbf{z}, \mathbf{z}') = p^\star(\mathbf{z}') T(\mathbf{z}', \mathbf{z}) \tag{11.40}$$

$\forall z$.

is sufficient (but not necessary) for $p^\star(z)$ to be invariant (to check, put (2) into (1)). (A Markov chain that respects the detailed balance is called reversible.)

- A Markov chain is ergodic if it converges to the invariant distribution irrespective of the choice of the initial conditions. The invariant distribution is then called equilibrum.
- An ergodic Markov chain can have only one equilibrium distribution.
- Why is it working? Choose the transition probabilities T to satisfy the detailed balance for our goal distribution $p(z)$.

repeat many steps.

$p^*(z) \to p(z)$.

The Metropolis-Hastings Algorithm

Desired: Avoid rejection

1. Choose a symmetric proposal distribution $q(z_A \mid z_B) = q(z_B \mid z_A)$.

- Generalisation of the Metropolis algorithm for nonsymmetric proposal distributions q_k.
- At step τ, draw a sample $z^\star$ from the distribution $q_k(z \mid z^{(\tau)})$ where k labels the set of possible transitions.
- Accept with probability

Metropolis requires $q_k(z^{(\tau)} \mid z^\star) = q_k(z^\star \mid z^{(\tau)})$

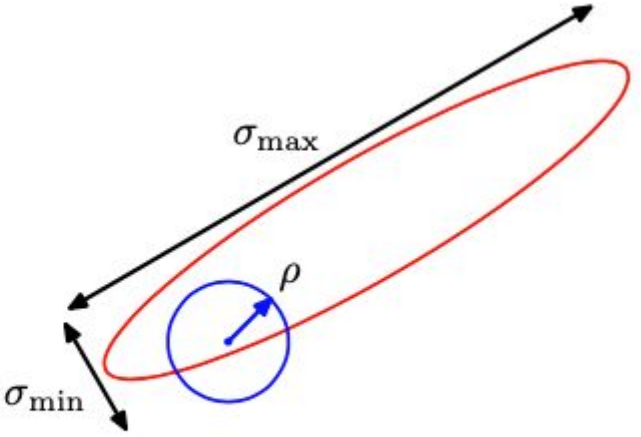
2. Accept the new sample $z^\star$ with probability

$$A(z^\star, z^{(\tau)}) = \min\left(1, \frac{\widetilde{p}(z^\star)}{\widetilde{p}(z^{(\tau)})}\right)$$

$$A_k(\mathbf{z}^\star, \mathbf{z}^{(\tau)}) = \min\left(1, \frac{\widetilde{p}(\mathbf{z}^\star) q_k(\mathbf{z}^{(\tau)} \mid \mathbf{z}^\star)}{\widetilde{p}(\mathbf{z}^{(\tau)}) q_k(\mathbf{z}^\star \mid \mathbf{z}^{(\tau)})}\right). \tag{11.44}$$

- Choice of proposal distribution critical.
- Common choice : Gaussian centered on the current state.
  - small variance → high acceptance rate, but slow walk through the state space; samples not independent
  - large variance → high rejection rate

Figure 11.10 Schematic illustration of the use of an isotropic Gaussian proposal distribution (blue circle) to sample from a correlated multivariate Gaussian distribution (red ellipse) having very different standard deviations in different directions, using the Metropolis-Hastings algorithm. In order to keep the rejection rate low, the scale ρ of the proposal distribution should be on the order of the smallest standard deviation $\sigma_{\min}$, which leads to random walk behaviour in which the number of steps separating states that are approximately independent is of order $(\sigma_{\max}/\sigma_{\min})^2$ where $\sigma_{\max}$ is the largest standard deviation.



Metropolis-Hastings: why does it work?

$$A_k(\mathbf{z}^\star, \mathbf{z}^{(\tau)}) = \min\left(1, \frac{\widetilde{p}(\mathbf{z}^\star) q_k(\mathbf{z}^{(\tau)}|\mathbf{z}^\star)}{\widetilde{p}(\mathbf{z}^{(\tau)}) q_k(\mathbf{z}^\star|\mathbf{z}^{(\tau)})}\right)$$

- Transition probability of this Markov chain is

$$T(z, z') = q_k(z' \,|\, z)\, A_k(z', z)$$

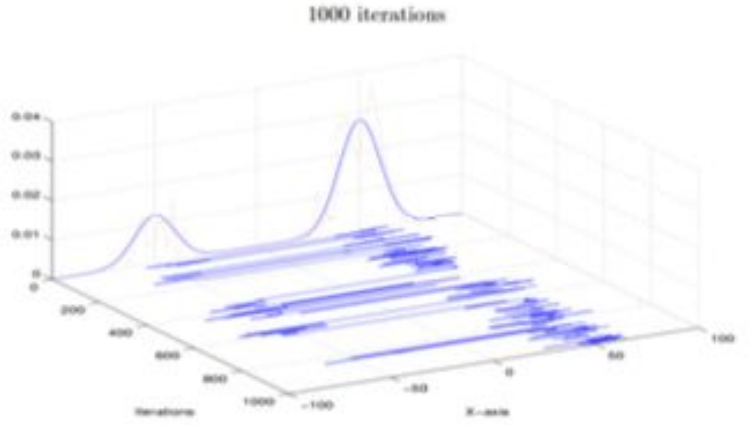
- Prove that $p(z)$ is the invariant distribution if the detailed balance holds

$$p(z)\, T(z, z') = T(z', z)\, p(z').$$

- Using the symmetry $\min(a, b) = \min(b, a)$ it can be shown that the detailed balance holds

$$\begin{aligned} p(\mathbf{z}) q_k(\mathbf{z}|\mathbf{z}') A_k(\mathbf{z}', \mathbf{z}) &= \min\left(p(\mathbf{z}) q_k(\mathbf{z}|\mathbf{z}'), p(\mathbf{z}') q_k(\mathbf{z}'|\mathbf{z})\right) \\ &= \min\left(p(\mathbf{z}') q_k(\mathbf{z}'|\mathbf{z}), p(\mathbf{z}) q_k(\mathbf{z}|\mathbf{z}')\right) \\ &= p(\mathbf{z}') q_k(\mathbf{z}'|\mathbf{z}) A_k(\mathbf{z}, \mathbf{z}') \end{aligned} \tag{11.45}$$

Metropolis-Hastings for sampling from mixture of Gaussians:



http://www.cs.ubc.ca/~arnaud/stat535/slides10.pdf

- With a random walk q we may get stuck in one mode.
- We could have proposal be mixture between random walk and "mode jumping".

Gibbs Sampling

- Goal: sample from a joint distribution $p(\mathbf{z}) = p(z_1, \ldots, z_M)$
- How? sample one variable from the distribution conditioned on all the other variable
- Example : given $p(z_1, z_2, z_3)$
- At step τ we have samples $z_1^{(\tau)}$, $z_2^{(\tau)}$ and $z_3^{(\tau)}$.
- Get samples for the next step $\tau + 1$

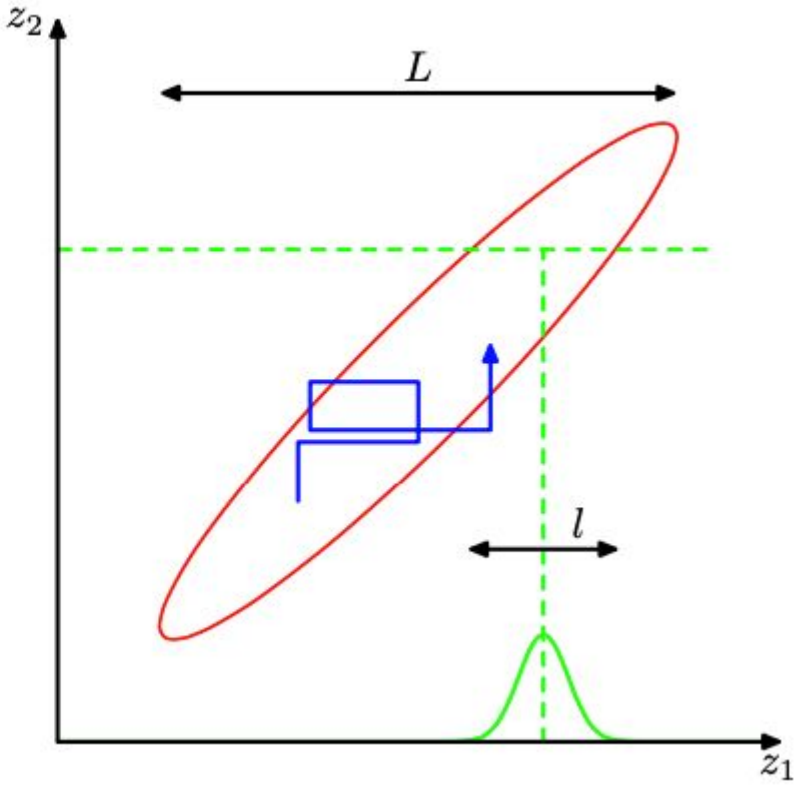
$$
\begin{aligned}
z_1^{(\tau+1)} &\sim p(z_1^{(\tau)} \mid z_2^{(\tau)}, z_3^{(\tau)}) \\
z_2^{(\tau+1)} &\sim p(z_2^{(\tau)} \mid z_1^{(\tau+1)}, z_3^{(\tau)}) \\
z_3^{(\tau+1)} &\sim p(z_3^{(\tau)} \mid z_1^{(\tau+1)}, z_2^{(\tau+1)})
\end{aligned}
$$

Gibbs sampling - why does it work?

1. $p(\mathbf{z})$ is invariant of each of the Gibbs sampling steps and hence of the whole Markov chain : a) $p(z_i \mid \{\mathbf{z}_{\setminus i}\})$ is invariant because the marginal distribution $p(\mathbf{z}_{\setminus i})$ does not change, and b) by definition each step samples from $p(z_i \mid \{\mathbf{z}_{\setminus i}\})$.
2. Ergodicity: sufficient that none of the conditional distribution is zero anywhere.
3. Gibbs sampling is a Metropolis-Hastings sampling in which each step is accepted.

$$A(\mathbf{z}^\star, \mathbf{z}) = \frac{p(\mathbf{z}^\star) q_k(\mathbf{z}|\mathbf{z}^\star)}{p(\mathbf{z}) q_k(\mathbf{z}^\star|\mathbf{z})} = \frac{p(z_k^\star|\mathbf{z}_{\setminus k}^\star) p(\mathbf{z}_{\setminus k}^\star) p(z_k|\mathbf{z}_{\setminus k}^\star)}{p(z_k|\mathbf{z}_{\setminus k}) p(\mathbf{z}_{\setminus k}) p(z_k^\star|\mathbf{z}_{\setminus k})} = 1 \qquad (11.49)$$

Figure 11.11 Illustration of Gibbs sampling by alternate updates of two variables whose distribution is a correlated Gaussian. The step size is governed by the standard deviation of the conditional distribution (green curve), and is $O(l)$, leading to slow progress in the direction of elongation of the joint distribution (red ellipse). The number of steps needed to obtain an independent sample from the distribution is $O((L/l)^2)$.



initial goal $E_p[f(x)]$

practically: keep one sample every L' samples

Gibbs Sampling

1. Initialize $\{z_i : i = 1, \ldots, M\}$
2. For $\tau = 1, \ldots, T$:
   - Sample $z_1^{(\tau+1)} \sim p(z_1 | z_2^{(\tau)}, z_3^{(\tau)}, \ldots, z_M^{(\tau)})$.
   - Sample $z_2^{(\tau+1)} \sim p(z_2 | z_1^{(\tau+1)}, z_3^{(\tau)}, \ldots, z_M^{(\tau)})$.
   - $\vdots$
   - Sample $z_j^{(\tau+1)} \sim p(z_j | z_1^{(\tau+1)}, \ldots, z_{j-1}^{(\tau+1)}, z_{j+1}^{(\tau)}, \ldots, z_M^{(\tau)})$.
   - $\vdots$
   - Sample $z_M^{(\tau+1)} \sim p(z_M | z_1^{(\tau+1)}, z_2^{(\tau+1)}, \ldots, z_{M-1}^{(\tau+1)})$.

Sampling methods

Motivation, why? Sampling and ML

Basic sampling algorithms

- Sampling standard distributions from U (0, 1)
- Rejection sampling
- Importance sampling

Markov chain Monte Carlo (MCMC)

- Markov chains
- Metropolis-Hasting

Gibbs sampling