

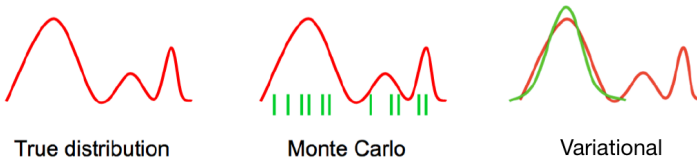


# Outline

- Introduction
- Rejection sampling
- Importance sampling
- Metropolis-Hastings
- Hamiltonian Monte Carlo
- Gibbs sampling

# Previously, in MBML...

- Representation (Weeks 1-4)
- Modelling toolbox (Weeks 5-8, +12)
- Inference (Weeks 9-11)
  - Exact inference
    - Analytical (Bayes' rule)
    - Variable elimination
    - Belief propagation
  - Approximate inference
    - Stochastic methods - Markov chain Monte Carlo (MCMC)
    - Deterministic methods - Variational inference (VI)



# Approximate inference

- Suppose we wish to compute the posterior distribution of the latent variables  $\mathbf{z}$  given some observed data  $\mathbf{x}$ :  $p(\mathbf{z}|\mathbf{x})$
- So far, we discussed various **exact** algorithms for posterior inference
- But, for many problems of interest exact posterior inference is **intractable**
  - Cannot determine the posterior distribution analytically
  - Cannot even compute expectations with respect to the posterior
- Reasons for intractability:
  - Dimensionality of the latent space is too high to work with directly
  - Posterior distribution has a highly complex form for which expectations are not analytically tractable
- We must resort to **approximate inference** methods!

## Monte Carlo methods

- In most situations, the posterior distribution  $p(\mathbf{z}|\mathbf{x})$  is required only for evaluating **expectations** of some function  $f(\mathbf{z})$  (e.g. to make predictions)

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

- (Note: we are using  $\mathbf{z}$  to denote the variables whose posterior we wish to infer!)
- For example, the mean of  $\mathbf{z}$  is given by

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

- The idea behind sampling methods is to obtain a set of **samples**  $\mathbf{z}^{(s)}$ , for  $s \in \{1, \dots, S\}$ , drawn independently from the distribution  $p(\mathbf{z}|\mathbf{x})$
- Allows to approximate expectations as **finite sums**!

$$\mathbb{E}_{p(\mathbf{z}|\mathbf{x})}[f(\mathbf{z})] \approx \frac{1}{S} \sum_{s=1}^S f(\mathbf{z}^{(s)})$$

## Example: making predictions

- Let  $\mathbf{z}$  denote the parameters of our model (e.g. Bayesian logistic regression)
- Suppose we wish to make a prediction  $y_*$  for a new observation  $\mathbf{x}_*$

$$\begin{aligned}\mathbb{E}_{p(\mathbf{z}|\mathbf{x})}[y_*|\mathbf{x}_*] &= \int p(y_*|\mathbf{x}_*, \mathbf{z}) p(\mathbf{z}|\mathbf{x}) d\mathbf{z} \\ &\approx \frac{1}{S} \sum_{s=1}^S p(y_*|\mathbf{x}_*, \mathbf{z}^{(s)})\end{aligned}$$

where  $\mathbf{z}^{(s)} \sim p(\mathbf{z}|\mathbf{x})$

- Notice that

$$\frac{1}{S} \sum_{s=1}^S p(y_*|\mathbf{x}_*, \mathbf{z}^{(s)}) \neq p\left(y_* \middle| \mathbf{x}_*, \frac{1}{S} \sum_{s=1}^S \mathbf{z}^{(s)}\right)$$

- Even if the posterior  $p(\mathbf{z}|\mathbf{x})$  is intractable to compute, we can still make predictions as long as we can obtain samples from the posterior!
- But if  $p(\mathbf{z}|\mathbf{x})$  is intractable, how can we sample from it?

# Properties of Monte Carlo

- Monte Carlo estimator:

$$\begin{aligned}\mathbb{E}_{p(\mathbf{z}|\mathbf{x})}[f(\mathbf{z})] &= \int f(\mathbf{z}) p(\mathbf{z}|\mathbf{x}) d\mathbf{z} \\ &\approx \hat{f} \triangleq \frac{1}{S} \sum_{s=1}^S f(\mathbf{z}^{(s)}), \quad \mathbf{z}^{(s)} \sim p(\mathbf{z}|\mathbf{x})\end{aligned}$$

- Estimator is unbiased:

$$\mathbb{E}_{p(\{\mathbf{z}^{(s)}\}|\mathbf{x})}[\hat{f}] = \frac{1}{S} \sum_{s=1}^S \mathbb{E}_{p(\mathbf{z}|\mathbf{x})}[f(\mathbf{z})] = \mathbb{E}_{p(\mathbf{z}|\mathbf{x})}[f(\mathbf{z})]$$

- Variance shrinks proportionally to  $1/S$ :

$$\mathbb{V}_{p(\{\mathbf{z}^{(s)}\}|\mathbf{x})}[\hat{f}] = \frac{1}{S^2} \sum_{s=1}^S \mathbb{V}_{p(\mathbf{z}|\mathbf{x})}[f(\mathbf{z})] = \frac{1}{S} \mathbb{V}_{p(\mathbf{z}|\mathbf{x})}[f(\mathbf{z})]$$

## Sampling from distributions

- Draw mass to the left of point:  $u \sim \text{Uniform}(0, 1)$
- Use inverse CDF to obtain sample  $y(u) = h^{-1}(u)$

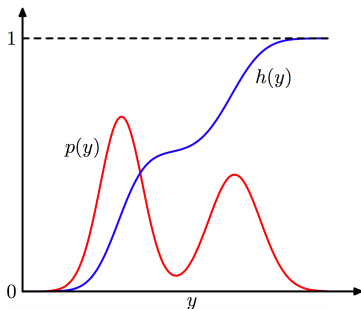


Figure: from PRML book, Bishop (2006)

- However, keep in mind that we can't always compute and invert  $h(y)$



## Rejection sampling

- Suppose we wish to sample from the unnormalized density  $\tilde{p}(z) \propto p(z)$
- Construct **proposal distribution**  $kq(z) \geq \tilde{p}(z)$

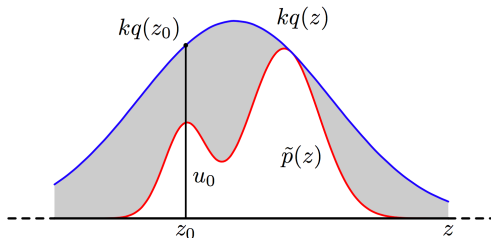


Figure: from PRML book, Bishop (2006)

- Draw sample  $z \sim q(z)$
- Draw height  $u \sim \text{Uniform}(0, kq(z))$
- **Reject** sample  $z$  if  $u > \tilde{p}(z)$

# Playtime!

- Rejection sampling
  - See “Part 1” of “11 - Markov chain Monte Carlo methods.ipynb” notebook
  - Expected duration: 30 minutes

## Importance sampling

- As previously mentioned, the posterior distribution  $p(z)$  is often required only for evaluating **expectations** of some function  $f(z)$

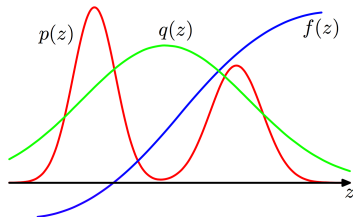


Figure: from PRML book, Bishop (2006)

- Importance sampling allows for just that!
- It does **not** provide a mechanism for drawing samples from the posterior  $p(z)$ , but it allows to compute expectations with respect to it:

$$\mathbb{E}_{p(z)}[f(z)] \approx \frac{1}{S} \sum_{s=1}^S f(z^{(s)}), \quad z^{(s)} \sim p(z)$$

# Importance sampling

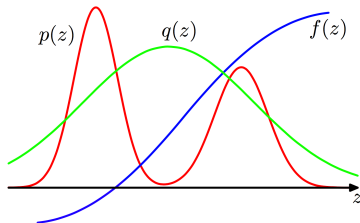


Figure: from PRML book, Bishop (2006)

- Rewrite the integral as an expectation under  $q$ :

$$\begin{aligned}\mathbb{E}_{p(z)}[f(z)] &= \int f(z) p(z) dz = \int f(z) \frac{p(z)}{q(z)} q(z) dz \\ &\approx \frac{1}{S} \sum_{s=1}^S f(z^{(s)}) \frac{p(z^{(s)})}{q(z^{(s)})}, \quad z^{(s)} \sim q(z)\end{aligned}$$

- The ratio  $\frac{p(z^{(s)})}{q(z^{(s)})}$  is the **importance/weights** of each sample!
- But, what if we cannot evaluate  $p(z)$  but only  $\tilde{p}(z) \propto p(z)$ ?

## Importance sampling

- But, what if we cannot evaluate  $p(z)$  but only  $\tilde{p}(z) \propto p(z)$ ?
- Let  $p(z) = \frac{1}{Z_p} \tilde{p}(z)$  and  $q(z) = \frac{1}{Z_q} \tilde{q}(z)$

$$\mathbb{E}_{p(z)}[f(z)] = \int f(z) p(z) dz \approx \frac{Z_q}{Z_p} \frac{1}{S} \sum_{s=1}^S f(z^{(s)}) \underbrace{\frac{\tilde{p}(z^{(s)})}{\tilde{q}(z^{(s)})}}_{\tilde{r}^{(s)}}, \quad z^{(s)} \sim q(z)$$

# Importance sampling

- But, what if we cannot evaluate  $p(z)$  but only  $\tilde{p}(z) \propto p(z)$ ?
- Let  $p(z) = \frac{1}{Z_p} \tilde{p}(z)$  and  $q(z) = \frac{1}{Z_q} \tilde{q}(z)$

$$\begin{aligned} \mathbb{E}_{p(z)}[f(z)] &= \int f(z) p(z) dz \approx \frac{Z_q}{Z_p} \frac{1}{S} \sum_{s=1}^S f(z^{(s)}) \underbrace{\frac{\tilde{p}(z^{(s)})}{\tilde{q}(z^{(s)})}}_{\tilde{r}^{(s)}}, \quad z^{(s)} \sim q(z) \\ &\approx \frac{1}{S} \sum_{s=1}^S f(z^{(s)}) \underbrace{\frac{\tilde{r}^{(s)}}{\frac{1}{S} \sum_{j=1}^S \tilde{r}^{(j)}}}_{w^{(s)}} = \sum_{s=1}^S f(z^{(s)}) w^{(s)} \end{aligned}$$

since  $\frac{Z_p}{Z_q} \approx \frac{1}{S} \sum_s \tilde{r}^{(s)}$ .

- All we need to do is compute the importance/weight  $w^{(s)}$  of each sample  $s$  as the normalized ratio  $\tilde{r}^{(s)} = \frac{\tilde{p}(z^{(s)})}{\tilde{q}(z^{(s)})}$

# Playtime!

- Importance sampling
  - See “Part 2” of “11 - Markov chain Monte Carlo methods.ipynb” notebook
  - Expected duration: 30 minutes

# Importance sampling

- Rejection and importance sampling scale badly with dimensionality
- Also, the whole procedure still feels naive... A 2-dimensional example:

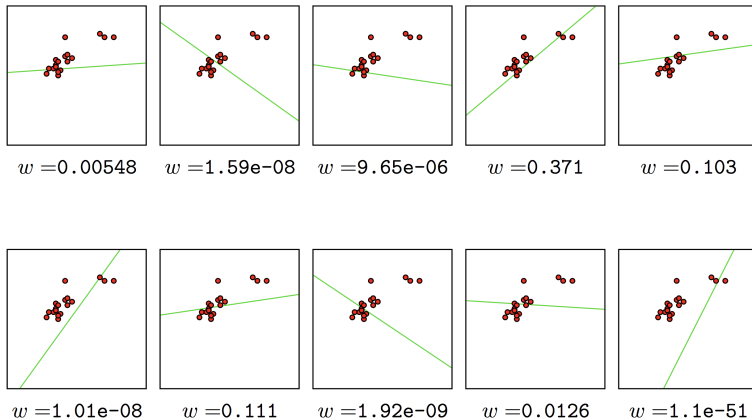


Figure: Samples from posterior over linear regression coefficients  $\beta = (\beta_0, \beta_1)^T$  using importance sampling. Figure from Iain Murray (2009)



# Metropolis algorithm

- Start with some initial state  $\mathbf{z}$  and iterate the following steps:
- Perturb variables using proposal distribution  $q(\mathbf{z}'|\mathbf{z})$ , e.g.  $q(\mathbf{z}'|\mathbf{z}) = \mathcal{N}(\mathbf{z}'|\mathbf{z}, \sigma^2\mathbf{I})$
- Accept proposal with probability:  $\min\left(1, \frac{\tilde{p}(\mathbf{z}')}{\tilde{p}(\mathbf{z})}\right)$
- In other words: always accept if new state  $\mathbf{z}'$  has higher probability, otherwise go to new state  $\mathbf{z}'$  anyway with probability  $\frac{\tilde{p}(\mathbf{z}')}{\tilde{p}(\mathbf{z})}$

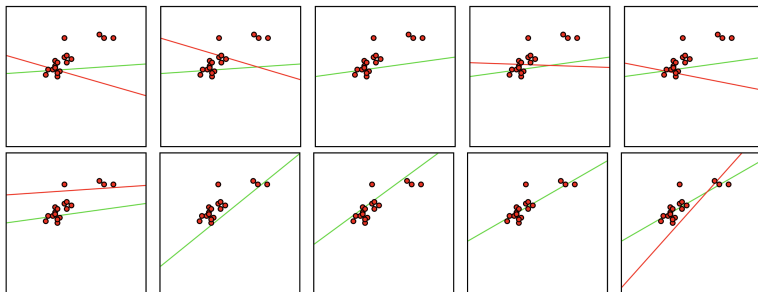


Figure: Samples from posterior over linear regression coefficients  $\beta = (\beta_0, \beta_1)^T$  using the Metropolis algorithm. Figure from Iain Murray (2009)

## Metropolis algorithm

- Start with some initial state  $\mathbf{z}$  and iterate the following steps:
- Perturb variables using proposal distribution  $q(\mathbf{z}'|\mathbf{z})$ , e.g.  $q(\mathbf{z}'|\mathbf{z}) = \mathcal{N}(\mathbf{z}'|\mathbf{z}, \sigma^2 \mathbf{I})$
- Accept proposal with probability:  $\min\left(1, \frac{\tilde{p}(\mathbf{z}')}{\tilde{p}(\mathbf{z})}\right)$

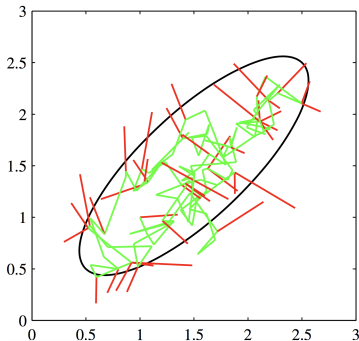


Figure: from PRML book, Bishop (2006)

- Green = accepted steps, Red = rejected steps

# Markov chain Monte Carlo

- Construct a **biased** random walk that explores the posterior distribution  $\tilde{p}(\mathbf{z})$
- Markov steps at each time step  $t$ :  $\mathbf{z}^{(t)} \sim p(\mathbf{z}^{(t)}|\mathbf{z}^{(t-1)})$

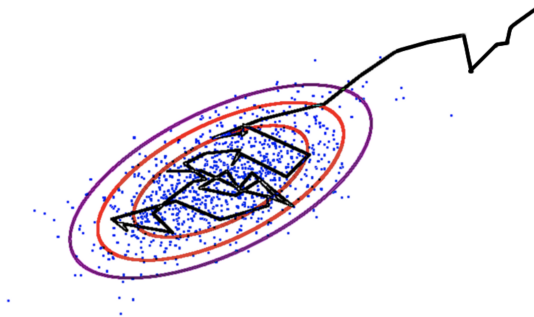


Figure: from Iain Murray (2009)

- MCMC gives approximate (but **correlated!**) samples from  $\tilde{p}(\mathbf{z})$
- Can we choose any **transition operator**  $T(\mathbf{z}^{(t)} \leftarrow \mathbf{z}^{(t-1)}) \triangleq p(\mathbf{z}^{(t)}|\mathbf{z}^{(t-1)})$ ?

## Some definitions...

- Consider a Markov chain with transition probabilities:  $T_t(\mathbf{z}^{(t)} \leftarrow \mathbf{z}^{(t-1)})$
- A Markov chain is called **homogeneous** if the transition probabilities are the same for all  $t$
- A distribution  $p^*(\mathbf{z})$  is said to be **invariant**, or **stationary**, with respect to a Markov chain if each step in the chain leaves that distribution invariant
- For a homogeneous Markov chain with transition probabilities  $T(\mathbf{z} \leftarrow \mathbf{z}')$ , the distribution  $p^*(\mathbf{z})$  is invariant if

$$p^*(\mathbf{z}) = \sum_{\mathbf{z}'} T(\mathbf{z} \leftarrow \mathbf{z}') p^*(\mathbf{z}')$$

- Our goal is to use Markov chains to sample from a given distribution
- We can achieve this if we set up a Markov chain such that the desired distribution is invariant!

# MCMC properties

- A sufficient (but not necessary) condition for ensuring that the required distribution  $p(\mathbf{z})$  is invariant is to choose the transition probabilities to satisfy the property of **detailed balance**, defined by

$$p^*(\mathbf{z}') T(\mathbf{z} \leftarrow \mathbf{z}') = p^*(\mathbf{z}) T(\mathbf{z}' \leftarrow \mathbf{z})$$

- A Markov chain that respects detailed balance is said to be **reversible**
- Lastly, we must also require that for  $t \rightarrow \infty$ ,  $p(\mathbf{z}^{(t)})$  converges to the required invariant distribution  $p^*(\mathbf{z})$ , irrespective of the choice of initial distribution  $p(\mathbf{z}^{(0)})$
- This property is called **ergodicity**
- The invariant distribution is then called the **equilibrium distribution**

# Metropolis-Hastings

- A more general version of the Metropolis algorithm:
  - Propose a move from the current state:  $\mathbf{z}' \sim q(\mathbf{z}'|\mathbf{z})$
  - Accept proposal with probability:

$$\min \left( 1, \frac{\tilde{p}(\mathbf{z}') q(\mathbf{z}|\mathbf{z}')}{\tilde{p}(\mathbf{z}) q(\mathbf{z}'|\mathbf{z})} \right)$$

- Satisfies detailed balance! (see Bishop (2006) for proof)
- Since the Gaussian is a symmetric function, i.e.  $\mathcal{N}(\mathbf{z}'|\mathbf{z}, \sigma^2) = \mathcal{N}(\mathbf{z}|\mathbf{z}', \sigma^2)$ , when  $q(\mathbf{z}'|\mathbf{z}) = \mathcal{N}(\mathbf{z}'|\mathbf{z}, \sigma^2\mathbf{I})$  we get back the Metropolis algorithm:

$$\min \left( 1, \frac{\tilde{p}(\mathbf{z}') \cancel{q(\mathbf{z}|\mathbf{z}')}}{\tilde{p}(\mathbf{z}) \cancel{q(\mathbf{z}'|\mathbf{z})}} \right) = \min \left( 1, \frac{\tilde{p}(\mathbf{z}')}{\tilde{p}(\mathbf{z})} \right)$$

# Metropolis-Hastings

- Be careful how you choose the step size  $\sigma^2$

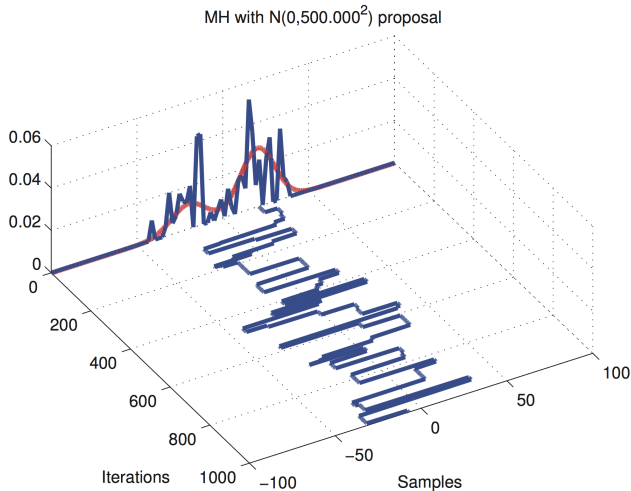


Figure: from ML: A prob. perspective book, Murphy (2011)

# Metropolis-Hastings

- Be careful how you choose the step size  $\sigma^2$

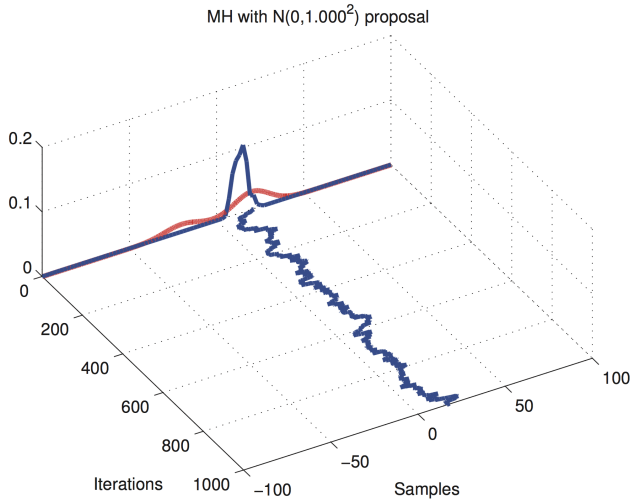


Figure: from ML: A prob. perspective book, Murphy (2011)



# Metropolis-Hastings

- Careful how you choose the step size  $\sigma^2$

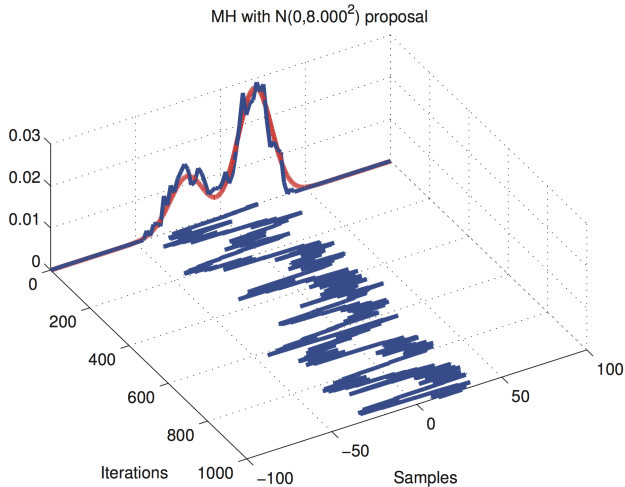


Figure: from ML: A prob. perspective book, Murphy (2011)

# MCMC in practice

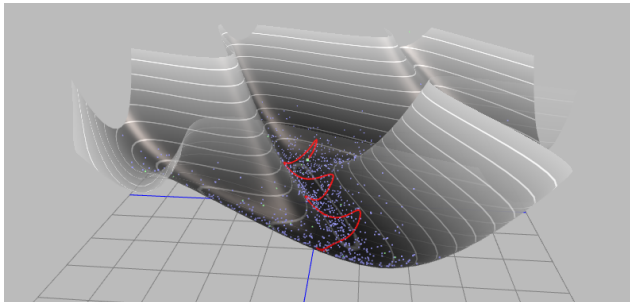
- The samples are correlated! We should **thin** - only keep every  $n^{th}$  sample
- Arbitrary initialization means early samples are bad - discard a **burn-in** period
- A good idea is to run **multiple chains** (e.g. in parallel - multicore)
- How do you know when to stop?
- Make sure to check diagnostics!  
(e.g. “Rhat” and number of effective samples in STAN)

# Playtime!

- Metropolis-Hastings
  - See “Part 3’ of “11 - Markov chain Monte Carlo methods.ipynb” notebook
  - Expected duration: 30 minutes

# Hamiltonian Monte Carlo: basic idea

- Include **Hamiltonian dynamics**
  - Construct a landscape with gravitational potential energy  $E(\mathbf{z}) = -\log p^*(\mathbf{z})$
  - Introduce velocity  $v$  (auxiliary variable) carrying kinetic energy  $K(v)$
- Define joint distribution:  $p(\mathbf{z}, v) \propto e^{-E(\mathbf{z})} e^{-K(v)} = e^{-H(\mathbf{z}, v)}$
- Velocity  $v$  is independent of position and Gaussian distributed



# Hamiltonian Monte Carlo: basic idea

- Procedure:
  - Gibbs sample velocity  $v$  (we will see Gibbs sampling next!)
  - Simulate Hamiltonian dynamics then flip sign of velocity
  - Accept new position with probability:  $\min[1, e^{H(\mathbf{z}, v) - H(\mathbf{z}', v')}]$
- Hamiltonian dynamics are simulated approximately
- Distances between successive generated points are typically large, so we need less iterations to get representative sampling!
- Each iteration is more computationally expensive, but sampling is more efficient
- For a detailed explanation, see:  
[http://arogozhnikov.github.io/2016/12/19/markov\\_chain\\_monte\\_carlo.html](http://arogozhnikov.github.io/2016/12/19/markov_chain_monte_carlo.html)

# Hamiltonian Monte Carlo: basic idea

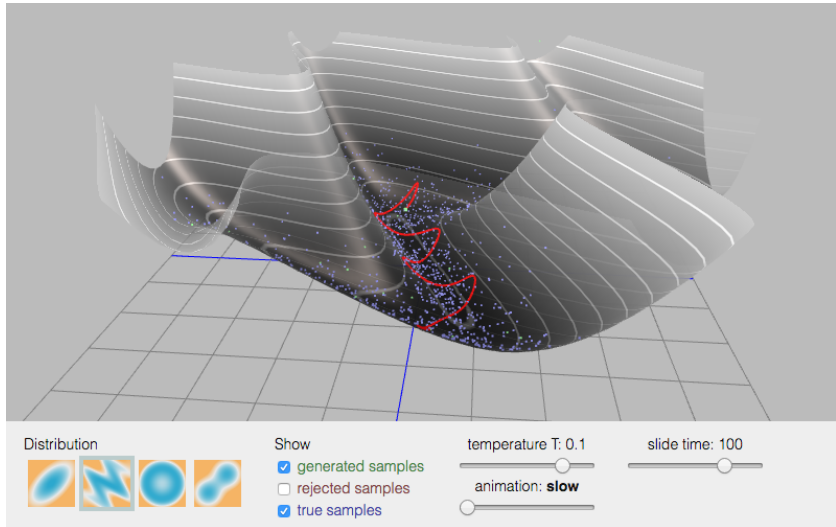
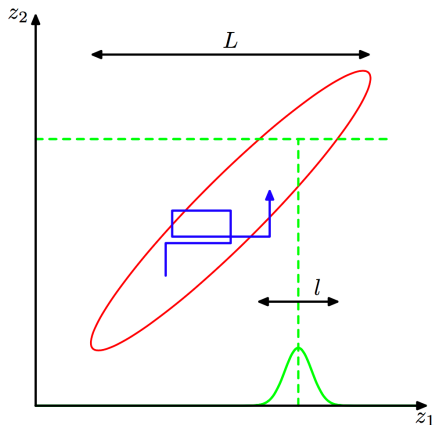


Figure: [http://arogozhnikov.github.io/2016/12/19/markov\\_chain\\_monte\\_carlo.html](http://arogozhnikov.github.io/2016/12/19/markov_chain_monte_carlo.html)

# Gibbs sampling



- Suppose we wish to sample from the posterior  $p(\mathbf{z}) = p(z_1, \dots, z_M)$
- Start with initial state  $\mathbf{z}^{(0)}$
- Pick each variable  $z_m$  in turn (or randomly) and resample it from the conditional:

$$z_m \sim p(z_m | \mathbf{z}_{i \neq m})$$

# Gibbs sampling

- Simple and widely applicable MCMC algorithm
- Gibbs sampling has no rejection! (see next slide)
- Must be able to sample from conditional distributions  $p(z_m | \mathbf{z}_{i \neq m})$ 
  - Discrete conditionals with a few possible settings can be explicitly normalized

$$p(z_m | \mathbf{z}_{i \neq m}) = \frac{p(z_m, \mathbf{z}_{i \neq m})}{\sum_{z_m} p(z_m, \mathbf{z}_{i \neq m})}$$

- Continuous conditionals require either:
  - Conjugacy - to allow for analytical solutions
  - Or efficient approximate solutions (e.g. standard sampling methods)



## Gibbs sampling

- Gibbs sampling satisfies detailed balance (see e.g. Bishop (2006) for proof)
- It is a particular instance of the Metropolis-Hastings algorithm where  $q(\mathbf{z}'|\mathbf{z}) = p(z'_m|\mathbf{z}_{i \neq m})$
- Probability of acceptance:

$$\min \left( 1, \frac{p(\mathbf{z}') q(\mathbf{z}|\mathbf{z}')}{p(\mathbf{z}) q(\mathbf{z}'|\mathbf{z})} \right) = \min \left( 1, \underbrace{\frac{p(z'_m|\mathbf{z}'_{i \neq m}) p(\mathbf{z}'_{i \neq m}) p(z_m|\mathbf{z}'_{i \neq m})}{p(z_m|\mathbf{z}_{i \neq m}) p(\mathbf{z}_{i \neq m}) p(z'_m|\mathbf{z}_{i \neq m})}}_{=1} \right) = 1$$

where we used the fact that  $p(\mathbf{z}) = p(z_m|\mathbf{z}_{i \neq m}) p(\mathbf{z}_{i \neq m})$   
and  $\mathbf{z}'_{i \neq m} = \mathbf{z}_{i \neq m}$  (because they were left unchanged!)

- Gibbs sampling has **no rejection!**
- A practical example (discrete variables): see slides 18-25 from David Sontag (<http://people.csail.mit.edu/dsontag/courses/pgm13/slides/lecture9.pdf>)

# Summary

- We need approximate methods to compute posteriors (sums/integrals)
- Monte Carlo does not explicitly depend on dimension, but simple methods work only in low dimensions
- MCMC methods can make local moves - important for higher dimensions!
- General and often easy to implement (simple computations), but hard to diagnose

# Playtime!

- Gibbs sampling
  - See “Part 4’ of “11 - Markov chain Monte Carlo methods.ipynb” notebook
  - Expected duration: 30 minutes