



# Sampling methods and approximations

Generative and Deep Learning (GDL)

Daide Bacciu ([davide.bacciu@unipi.it](mailto:davide.bacciu@unipi.it))



UNIVERSITÀ DI PISA



# Objectives

- ◇ Sampling
  - ◇ What is it?
  - ◇ Why do we need it?
  - ◇ Properties of samplers
- ◇ Sampling from **univariate distributions**
- ◇ Sampling from **multivariate distributions**
  - ◇ Ancestor sampling
  - ◇ Gibbs Sampling
  - ◇ (Elements of) Markov Chain Monte Carlo (MCMC)
- ◇ Sampling for **approximated LDA inference**

# Sampling fundamentals

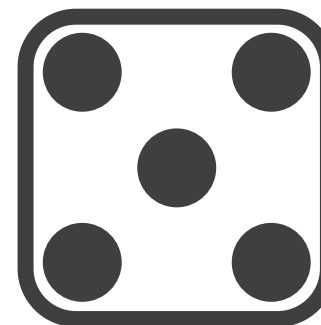
# What is sampling?

**Sampling** consists in drawing a set of **realizations**  $X = \{x_1, \dots, x_L\}$  of a random variable  $x$  with distribution  $p(x)$ .

**Example:**

We would like to sample a dice:  $p(x = i) = 1/6$ ,  $i \in [1, 6]$ .

$l$	$x^l$
1	5



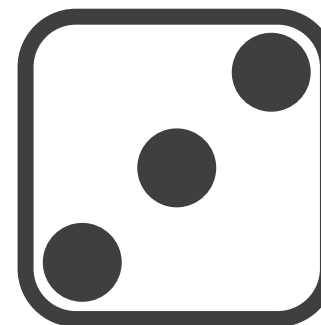
# What is sampling?

**Sampling** consists in drawing a set of **realizations**  $X = \{x_1, \dots, x_L\}$  of a random variable  $x$  with distribution  $p(x)$ .

**Example:**

We would like to sample a dice:  $p(x = i) = 1/6$ ,  $i \in [1, 6]$ .

$l$	$x^l$
1	5
2	3



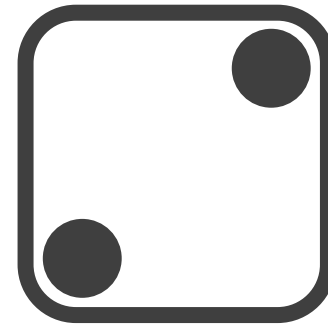
# What is sampling?

**Sampling** consists in drawing a set of **realizations**  $X = \{x_1, \dots, x_L\}$  of a random variable  $x$  with distribution  $p(x)$ .

**Example:**

We would like to sample a dice:  $p(x = i) = 1/6$ ,  $i \in [1, 6]$ .

$l$	$x^l$
1	5
2	3
3	2



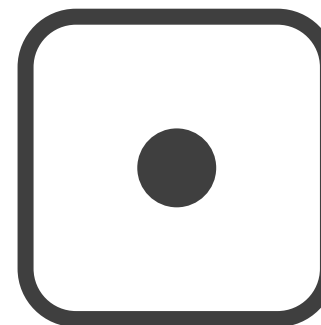
# What is sampling?

**Sampling** consists in drawing a set of **realizations**  $X = \{x_1, \dots, x_L\}$  of a random variable  $x$  with distribution  $p(x)$ .

**Example:**

We would like to sample a dice:  $p(x = i) = 1/6$ ,  $i \in [1, 6]$ .

$l$	$x^l$
1	5
2	3
3	2
4	1



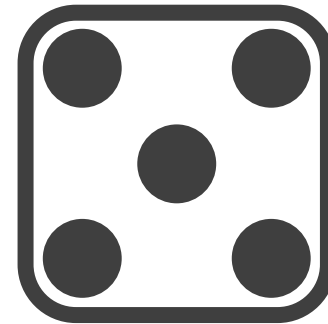
# What is sampling?

**Sampling** consists in drawing a set of **realizations**  $X = \{x_1, \dots, x_L\}$  of a random variable  $x$  with distribution  $p(x)$ .

**Example:**

We would like to sample a dice:  $p(x = i) = 1/6$ ,  $i \in [1, 6]$ .

$l$	$x^l$
1	5
2	3
3	2
4	1
5	5



The set  $X = \{5, 3, 2, 1, 5\}$  contains  $L = 5$  samples.

# Why do we need sampling?

## Approximating expectations

Suppose that we want to compute the expectation  $\mathbb{E}_{p(x)}[f(x)]$ .

If  $p(x)$  is **intractable**, we cannot compute it enumerating all the states of  $x$ .

If we have a sample set  $X = \{x_1, \dots, x_L\}$ , then we can approximate the expectation as:

$$\mathbb{E}_{p(x)}[f(x)] \approx \frac{1}{L} \sum_{l=1}^L f(x^l) \equiv \hat{f}_X$$

There are many cases where  $p(x)$  is **intractable**:

- ◇ the posterior  $P(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta)$  in **LDA**;
- ◇ posteriors when **non-conjugate priors** are used
- ◇ the distribution of a **Boltzmann Machine** (next lecture)

# Why do we need sampling?

## Learning parameters

In Bayesian models, **parameters** are **random variables**.

We can learn the model parameters by **sampling their posteriors!**

In **LDA**, we can learn the model parameters by sampling:

$$\theta, \mathbf{z}, \beta \sim P(\theta, \mathbf{z}, \beta | \mathbf{w}, \alpha)$$

**Sampling** from the posterior is also useful to **classify** new instances!

In this case, we sample:

$$\theta^*, \mathbf{z}^* \sim P(\theta, \mathbf{z} | \mathbf{w}^*, \alpha, \beta),$$

where  $\mathbf{w}^*$  are the words in the new documents.

# The Challenge

## Defining methods that can sample from unknown distributions!

A **sampler  $S$**  is a method that can generate a sample set  $X$  of a **target distribution  $p(x)$**

The **quality** of the sampling approximation depends on the **properties** of  $\tilde{p}(X)$ , i.e. the probability to obtain a **sample set  $X$  using sampler  $S$**

$$p(X) \neq p(x)$$

$p(x)$  → The distribution **to** sample → **Intractable**: does not depend on  $S$

$\tilde{p}(X)$  → The distribution **of** the samples → **Tractable**: depends on the **sampler  $S$**

# Properties of sampling

The most important properties of sampling are:

- ◇ the **empirical distribution converges** almost surely to the **true distribution**:

$$\lim_{L \rightarrow \infty} \frac{1}{L} \sum_{l=1}^L \mathbb{I}[x^l = i] = p(x = i), \quad x^l \sim p(x)$$

where  $\mathbb{I}[c] = 1$  if and only if  $c$  is true;

- ◇ the **sampling approximation**  $\hat{f}_X$  of the expectation can be an **unbiased estimator**;
- ◇ the **sampling approximation**  $\hat{f}_X$  of the expectation can have **low variance**;

The last two properties are **desirable but difficult** to ensure!

# Unbiased Sampler

Quick refresher:

Unbiased estimator  $\hat{\theta}$  of the unknown  $\theta$ .



the approximation computes a correct estimation on average.

Let  $\tilde{p}(X)$  the distribution over all possible realizations of the sampling set  $X$  generated by sampler  $S$ , then  $\hat{f}_X$  is an unbiased estimator if:

$$\mathbb{E}_{\tilde{p}(X)}[\hat{f}_X] = \mathbb{E}_{p(x)}[f(x)].$$

This is true provided that  $\tilde{p}(x^l) = p(x)$ !

The equality ensure us that we are **sampling the desired distribution**, i.e. we are using a **valid sampler S**!

# Sampler Variance - Definition

The variance of  $\hat{f}(X)$  tell us how much we can rely on the approximation computed using the sampling set  $X$ .

Let

$$\Delta \hat{f}_X = \hat{f}_X - \mathbb{E}_{\tilde{p}(X)}[\hat{f}_X],$$

the variance of  $\hat{f}(X)$  is given by:

$$\mathbb{E}_{\tilde{p}_X} [(\Delta \hat{f}_X)^2]$$

If the variance is **low**,  $\hat{f}(X)$  is (quite) always **close** to its expected value, i.e  $\mathbb{E}_{p(x)}[f(x)]!$

# Sampler Variance - Approximation

If we assume:

- ◇  $\tilde{p}(x^l) = p(x^l)$  (same marginals);
- ◇  $\tilde{p}(x^l, x^{l'}) = \tilde{p}(x^l)\tilde{p}(x^{l'})$  (samples independence);

we obtain

$$\mathbb{E}_{\tilde{p}_X} \left[ (\Delta \hat{f}_X)^2 \right] = \frac{1}{L} \text{Var}_{p(x)} [f(x)].$$

We can use a small number of samples to accurately estimate the expectation!

Provided that  $\text{Var}_{p(x)} [f(x)]$  is finite.

# Small recap

So far, we have shown that:

- ◇ we need **sampling**:
  - ◇ to **approximate expectations**;
  - ◇ to do **inference** in **Bayesian models**.
- ◇ properties of the **sampling procedure** depends on  $\tilde{p}(X)$ :
  - ◇  $\tilde{p}(x^l) = p(x^l) \Rightarrow$  **valid sampler**;
  - ◇  $\tilde{p}(x^l, x^{l'}) = \tilde{p}(x^l)\tilde{p}(x^{l'}) \Rightarrow$  **low approximation variance**

In the **next slides**, we introduce examples of **sampling procedures**:

- ◇ sampling from **univariate distributions**;
- ◇ sampling from **multivariate distributions**:
  - ◇ **naive approaches**;
  - ◇ **exact procedures**;
  - ◇ **approximated procedures**.

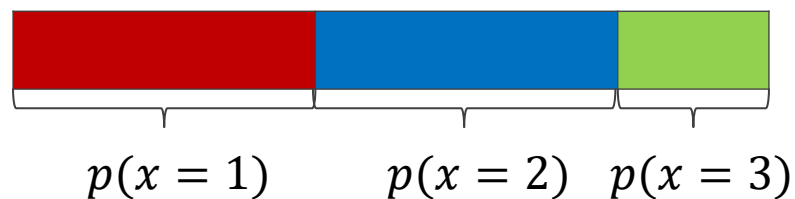
# Valid Samplers

# Univariate Sampling

Drawing samples from a univariate distribution is easy!

We only need a random number generator  $R$  which produces a value uniformly at random in  $[0, 1]$ .

$$p(x) = \begin{cases} 0.4 & x = 1 \\ 0.4 & x = 2 \\ 0.2 & x = 3 \end{cases}$$



$R$	$x$
0.19	1
0.24	1
0.47	2
0.88	3
0.73	2
0.63	2
0.52	2
0.96	3

# Multivariate Sampling

In the **multivariate** case,  $p(x)$  represents the **joint distribution** of a set of variables  $\{s_1, \dots, s_n\}$ , where each  $s_i$  is a **discrete variable** with  $C$  states. Hence, each sample  $x^l$  contains  $n$  values.

$X$	$S_1$	$S_2$	$S_3$	$S_4$	$S_5$
$x^1$	1	1	2	4	5
$x^2$	4	3	2	1	2
$x^3$	5	2	5	3	4
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x^L$	3	5	6	6	1

How can we sample from  $p(x)$ ?

# Naive Multivariate Sampling - 1

We build a **univariate distribution**  $p(S)$ , where  $S$  is a discrete variable with  $C^n$  states (i.e. **all possible combination** of  $s_i$  variable states).

$S$	$s_1$	$s_2$	$s_3$	$s_4$	$s_5$	$p(S)$
1	1	1	1	1	1	$p(1, 1, 1, 1, 1)$
2	1	1	1	1	2	$p(1, 1, 1, 1, 2)$
3	1	1	1	1	3	$p(1, 1, 1, 1, 3)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$C^n$	$C$	$C$	$C$	$C$	$C$	$p(C, C, C, C, C)$

We can sample from  $p(S)$  using the **univariate schema!**

$S$  has  $O(C^n)$  states! **Computationally infeasible!**

# Naive Multivariate Sampling - 2

Using the **chain rule**, we can rewrite the **joint distribution** as:

$$p(s_1, \dots, s_n) = p(s_1)p(s_2|s_1)p(s_3|s_1, s_2) \dots p(s_n|s_1, \dots, s_{n-1})$$

Then, we sample the variables in the following order:

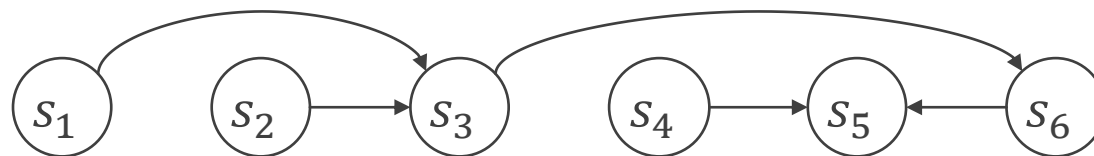
1. sample  $\tilde{s}_1 \sim p(s_1)$ ;
  2. sample  $\tilde{s}_2 \sim p(s_2|\tilde{s}_1)$ ;
  3. sample  $\tilde{s}_3 \sim p(s_3|\tilde{s}_1, \tilde{s}_2)$ ;
  - ⋮
  - n. sample  $\tilde{s}_n \sim p(s_n|\tilde{s}_1, \dots, \tilde{s}_{n-1})$ .
- Easy because **univariate!**

Unfortunately, computing the distribution  $p(s_i|s_{j<i})$  easily becomes **exponential w.r.t. number of states!**

# Ancestral Sampling

The approach used in the previous slide is called **Ancestral Sampling (AS)**.

If the distribution  $p(s_1, \dots, s_n)$  is **already represented** as a **Belief Network (BN)**, we can apply it directly!



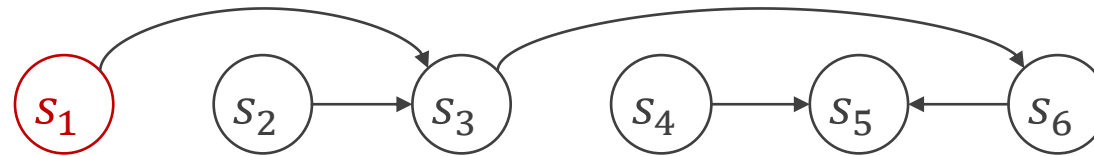
The **BN ancestral order** tell us the sampling order.

$$\{s_1, s_2, s_4\} < \{s_3\} < \{s_6\} < \{s_5\}$$

# Ancestral Sampling Example

$$\{s_1, s_2, s_4\} < \{s_3\} < \{s_6\} < \{s_5\}$$

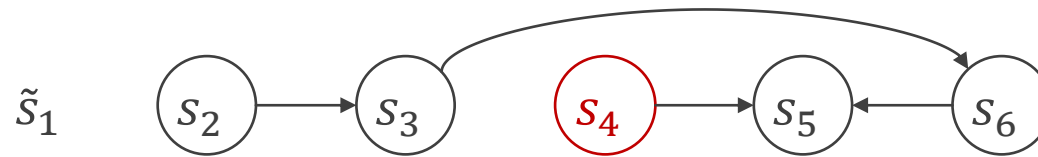
Sample  $\tilde{s}_1 \sim p(s_1)$



# Ancestral Sampling Example

$$\{\cancel{s_1}, s_2, \mathbf{s_4}\} < \{s_3\} < \{s_6\} < \{s_5\}$$

Sample  $\tilde{s}_4 \sim p(\mathbf{s_4})$

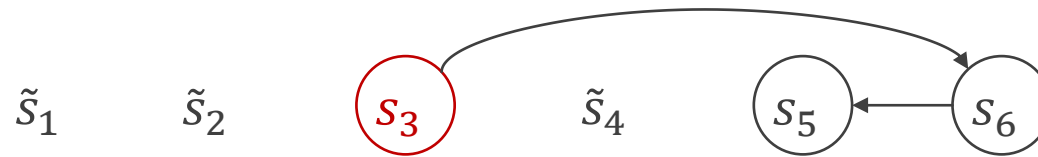




# Ancestral Sampling Example

$$\{\cancel{s_1}, \cancel{s_2}, \cancel{s_4}\} < \{s_3\} < \{s_6\} < \{s_5\}$$

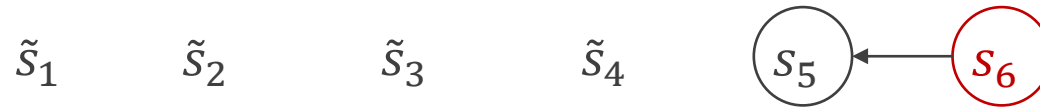
$$\text{Sample } \tilde{s}_3 \sim p(s_3 | \tilde{s}_1, \tilde{s}_2)$$



# Ancestral Sampling Example

$$\{\cancel{s_1}, \cancel{s_2}, \cancel{s_4}\} < \{\cancel{s_3}\} < \{s_6\} < \{s_5\}$$

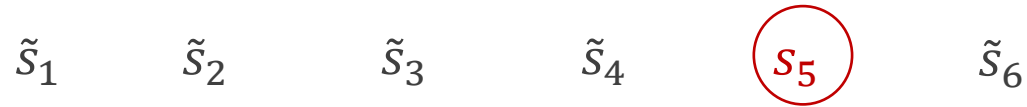
$$\text{Sample } \tilde{s}_6 \sim p(s_6 | \tilde{s}_3)$$



# Ancestral Sampling Example

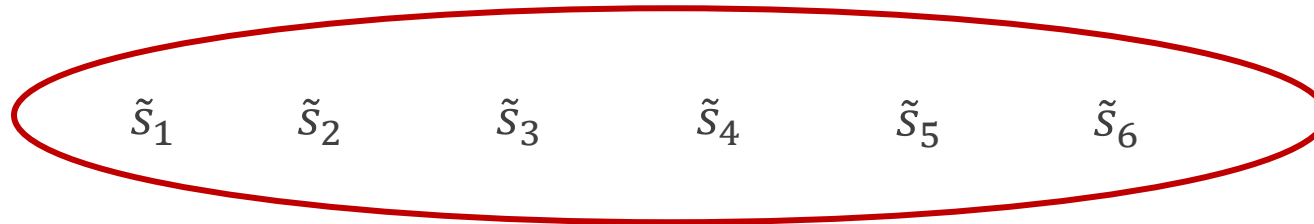
$$\{\cancel{s_1}, \cancel{s_2}, \cancel{s_4}\} < \{\cancel{s_3}\} < \{\cancel{s_6}\} < \{s_5\}$$

$$\text{Sample } \tilde{s}_5 \sim p(s_5 | \tilde{s}_4, \tilde{s}_6)$$



# Ancestral Sampling Example

$$\{\cancel{s}_1, \cancel{s}_2, \cancel{s}_4\} < \{\cancel{s}_3\} < \{\cancel{s}_6\} < \{\cancel{s}_5\}$$



This is a single sample  $x^l$ !

AS performs **exact sampling** since each sample  $x^l$  is drawn from  $p(x)$ .

The samples are also **independent!**  
Thus, AS has **low variance!**

# Sampling with evidence

Suppose that a subset of variables  $s_\epsilon$  are **visible**; writing  $s = s_\epsilon \cup s_{\setminus\epsilon}$ , we would like to sample from:

$$p(s_{\setminus\epsilon} | s_\epsilon) = \frac{p(s_{\setminus\epsilon}, s_\epsilon)}{p(s_\epsilon)}$$

## Can we still use AS?

- ◇ Clamping variables **changes the structure** of the BN.  
In previous example  $s_1 \perp\!\!\!\perp s_2$ , but  $s_1 \not\perp\!\!\!\perp s_2 | s_3$ .

Computing the new structure is complex as running exact inference!

- ◇ We can run AS on the old structure and then **discard** any samples which do not match the **evidence**.

We discard a lot of samples!

# The need of new sampling procedures

Sampling under evidence is **important!**

In probabilistic models, the inference is based on the **posterior**:

$$p(h|v) = \frac{p(h,v)}{p(v)},$$

where:

- ◇ where  $h$  is set of **hidden variables**
- ◇ where  $v$  is set of **visible variables** (i.e. the data)

We need an **efficient** method to sample under evidence!

In the next slide, we introduce the Gibbs sampling procedure.

# Approximate Samplers

# Gibbs Sampling

## Example

The idea is to start from a sample  $x_1 = \{s_1^1, \dots, s_n^1\}$  and to **update only one variable** at a time.

Sample	$s_1$	$s_2$	$s_3$	$s_4$	$s_5$
$x^1$	1	1	2	4	5

# Gibbs Sampling

## Example

The idea is to start from a sample  $x_1 = \{s_1^1, \dots, s_n^1\}$  and to **update only one variable** at a time.

Sample	$s_1$	$s_2$	$s_3$	$s_4$	$s_5$
$x^1$	1	1	2	4	5
$x^2$	3	1	2	4	5

# Gibbs Sampling

## Example

The idea is to start from a sample  $x_1 = \{s_1^1, \dots, s_n^1\}$  and to **update only one variable** at a time.

Sample	$s_1$	$s_2$	$s_3$	$s_4$	$s_5$
$x^1$	1	1	2	4	5
$x^2$	3	1	2	4	5
$x^3$	3	4	2	4	5

# Gibbs Sampling

## Example

The idea is to start from a sample  $x_1 = \{s_1^1, \dots, s_n^1\}$  and to **update only one variable** at a time.

Sample	$s_1$	$s_2$	$s_3$	$s_4$	$s_5$
$x^1$	1	1	2	4	5
$x^2$	3	1	2	4	5
$x^3$	3	4	2	4	5
$x^4$	3	4	2	1	5

# Gibbs Sampling

## Example

The idea is to start from a sample  $x_1 = \{s_1^1, \dots, s_n^1\}$  and to **update only one variable** at a time.

Sample	$s_1$	$s_2$	$s_3$	$s_4$	$s_5$
$x^1$	1	1	2	4	5
$x^2$	3	1	2	4	5
$x^3$	3	4	2	4	5
$x^4$	3	4	2	1	5
$x^5$	3	4	6	1	5

# Gibbs Sampling

## Example

The idea is to start from a sample  $x_1 = \{s_1^1, \dots, s_n^1\}$  and to **update only one variable** at a time.

Sample	$s_1$	$s_2$	$s_3$	$s_4$	$s_5$
$x^1$	1	1	2	4	5
$x^2$	3	1	2	4	5
$x^3$	3	4	2	4	5
$x^4$	3	4	2	1	5
$x^5$	3	4	6	1	5
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

# Gibbs Sampling

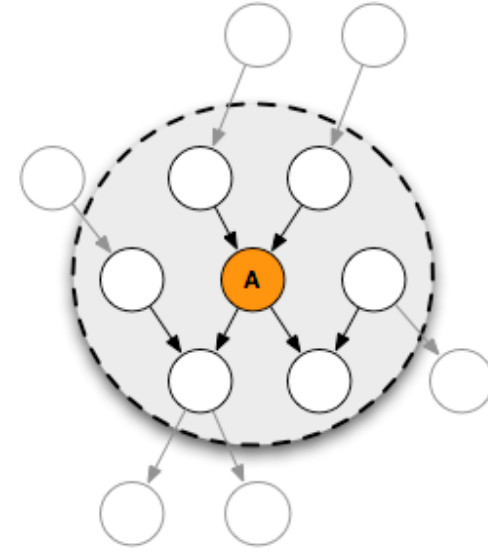
## Definition

During the  $(l + 1)$ -th iteration,

- ◆ we **select** a variable  $s_j$ ;
- ◆ we **sample** its value according to

$$s_j^{l+1} \sim p(s_j | s_{\setminus j}) = \frac{1}{Z} p(s_j | pa(s_j)) \prod_{k \in ch(j)} p(s_k | pa(s_k)),$$

where variables  $s_{\setminus j}$  are clamped to  $\{s_1^l, \dots, s_{j-1}^l, s_{j+1}^l, \dots, s_n^l\}$ .



It depends only on the **Markov blanket** of  $s_j$ ! **Easy to sample!**

Dealing with **evidence** is **easy!**  
We just do not select a variable!

# Gibbs Sampling

## Properties

The Gibbs sampling draws a new sample  $x^l$  from  $q(x^l|x^{l-1})$ .

◆ **Is the Gibbs sampling a valid sampling procedure?**

We are **not** sampling from  $p(x)$ !

We cannot ensure that the sampling distribution has the same marginals of  $p(x)$ .

However, if we compute the limit to  $L \rightarrow \infty$ , the series  $\{x_1, \dots, x_L\}$  **converges** to samples taken from  $p(x)$ !

In the limit of **infinite** samples, the Gibbs sampler is **valid**!

◆ **Has the Gibbs sampler low variance?**

No, samples are **highly** dependent!

# So why does Gibbs work?

- ◇ The Gibbs sampler define a **Markov chain** over the samples state
- ◇ The state-transition distribution is:

$$q(x^l | x^{l-1}) = \sum_j q(x_j) p(x_j^l | x_{\neq j}^{l-1}) \prod_{j \neq i} \delta(x_i^l, x_i^{l-1})$$

- ◇ If the Markov Chain is ergodic it has a **steady-state distribution**

$$q(x_\infty) = \mathbf{v}_\infty = A^\infty \mathbf{v}_\infty$$

where  $A$  is the usual state transition  $q(x_{t+1} | x_t)$  and  $\mathbf{v}_t$  is the marginal of states at time  $t$

- ◇ It can be demonstrated that the **steady state of Gibbs sampling distribution  $q$**  is the **target distribution  $p(x)$** 
  - ◇ In the limit  $t \rightarrow \infty$  we sample from the right distribution

# Gibbs Sampling Tips

- ◇ Due to its simplicity, Gibbs sampling is one of the most popular sampling methods
  - ◇ It is particularly convenient when applied to BN due to the Markov blanket property
- ◇ There are other versions:
  - ◇ Structured Gibbs sampling → We sample more than one variable from  $p(x_j, x_k \mid MB(x_j), MB(x_k))$
  - ◇ Collapsed Gibbs Sampling → We marginalise out some variables before sampling

## BEWARE:

1. Not all the samples are from  $p(x)$ !
  - ◇ **Burn in:** we discard the first  $N$  samples
2. Two consecutive samples are highly dependent!
  - ◇ **Subsampling:** we select a sample every  $K$
3. A Gibbs sampler is less effective when variables are strongly correlated
  - ◇ It is a valid sampler when the distribution is factorised

# MCMC Sampling Framework

Gibbs sampling is a **specialization** of the **Markov Chain Monte Carlo** (MCMC) sampling framework.

Build a **Markov Chain** whose stationary distribution is  $p(x)$ .

Let  $q(x^{l+1}|x^l)$  the MC state-transition distribution, we **must ensure** that the Markov Chain is:

- ◇ **irreducible** → it is possible to reach any state from anywhere;
- ◇ **aperiodic** → at each time-step, we can be anywhere.

Hence, the Markov Chain has a **unique stationary distribution**.

There are different  $q(\cdot)$  which converge to  $p(\cdot)$ .

# MCMC Sampling Procedures

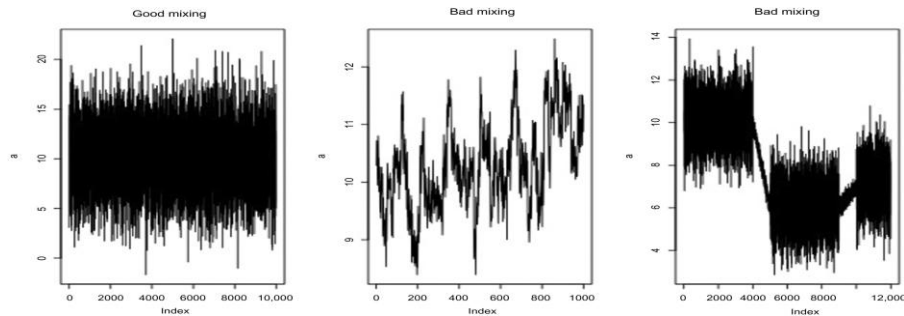
There are **many sampling procedures** in the MCMC framework, defining different **state-transition distribution**  $q(\cdot)$ :

- ◇ Gibbs Sampling
  - ◇  $q(\cdot)$  relies on **marginals**  $p(s_j | s_{\setminus j})$ !
  - ◇ it works well when variables are **not strongly related**!
- ◇ Metropolis-Hastings Sampling
  - ◇ based on a **proposal distribution**  $\tilde{q}(x^{l+1} | x^l)$ !
  - ◇ the choice of  $\tilde{q}(\cdot)$  is **crucial**!
- ◇ Particle Filtering
  - ◇ it is used in **recursive models** such as HMM!
- ◇ Hybrid Monte Carlo
- ◇ Swendsen-Wang
- ◇

Each of them has different characteristics!  
We should choose the most suitable for our purpose!

# The Convergence Issue

- ◇ It is difficult to detect convergence of MCMC
  - ◇ We can observe the trace plot of the sampled values



- ◇ We can compute running means
  - ◇ They should not vary at convergence
- ◇ We can execute multiple runs
  - ◇ Also, the convergence speed depends on the initial guess
- ◇ In practice, we wait as much as possible

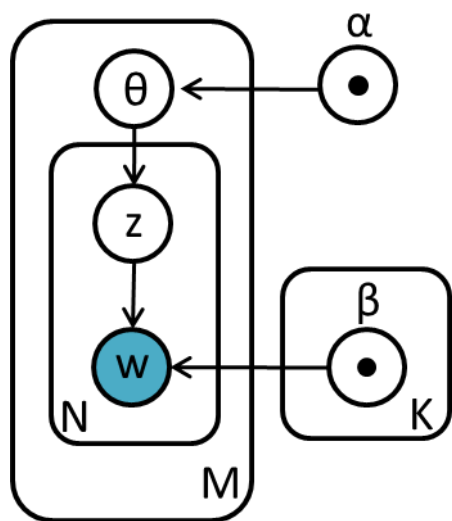
# Sampling for approximated inference

# How to use sampling in a Bayesian Model

- ◇ Build a model  $p(\theta, x)$  with intractable posterior
- ◇ Define your MCMC sampling schema  $q$  for the target distribution  $p(\theta, z | x)$
- ◇ Define the initial guess  $\theta^0$ ,  $l = 0$ ,  $S = \emptyset$
- ◇ Until you have time
  - ◇  $l = l + 1$
  - ◇  $\theta^l \sim q(\theta^l | \theta^{l-1})$
  - ◇ If  $l < N$  and  $(l - N) \% K = 0$ : **(burn-in and down-sampling)**
    - ◇  $S = S \cup \{\theta^l\}$
- ◇ At inference time, let  $x'$  be a new data-point

$$\tilde{p}(z' | x', S) = \mathbb{E}_{\theta \sim S}[p(z | x', \theta)] \approx \frac{1}{|S|} \sum_{\theta^* \in S} p(z | x', \theta^*)$$

# LDA Gibbs Sampling



Start from an initial guess  $\{z_{ij}^0, \theta_i^0, \beta^0\}$ .

Do:

1.  $z_{ij}^{l+1} \sim P(z_{ij} | \mathbf{w}, \mathbf{z}^{l-1}, \theta^l, \beta^l) \approx \text{Cat}(z_{ij}, \beta_{w_{ij}}^l * \theta_i)$
2.  $\theta_i^{l+1} \sim P(\theta_i | \mathbf{w}, \mathbf{z}^{l+1}, \beta^l, \alpha) \approx \text{Dir}(\theta_i | n_i + \alpha_i)$
3.  $\beta^{l+1} \sim P(\beta | \mathbf{w}, \mathbf{z}^{l+1}, \theta^{l+1}) \dots$

Repeat until convergence.

The derivation of the sampling distributions can be quite **mathematically involved**

The convergence criteria is based on  $P(\mathbf{z}, \mathbf{w}, \theta | \beta, \alpha)$ .  
The procedure terminates when the likelihood stops increasing.

# Wrap-up

# Take home messages

- ◇ **Sampling is useful** to deal with intractable  $p(x)$ :
  - ◇ we can **approximate expectations**;
  - ◇ we can perform **inference** in **Bayesian models**;
- ◇  $p(x)$  **univariate** → sampling is **easy**!
- ◇  $p(x)$  **multivariate** → sampling is **difficult**!
  - ◇ **naive** approaches are **not feasible**;
  - ◇ if  $p(x)$  is a **BN**, we can use **AS** (**valid** and with **low variance**);
  - ◇ **AS** does not work with **evidence** (**we always have it!**);
- ◇ **MCMC framework** approximates the sampling procedure:
  - ◇ we can easily deal with evidence;
  - ◇ the sampler defines a MC whose stationary distribution is  $p(x)$ ;
    - ◇ the sampler is valid in the limit  $l \rightarrow \infty$ ;
  - ◇ different state-transition  $q$  leads to different procedures:
    - ◇ Gibbs Sampling, Metropolis-Hastings Sampling, . . .

# Next Lecture

(Conditional) [Markov Random Fields](#)

- ◇ Undirected graphical models
- ◇ Express constraints between RV without needing to use probabilities
- ◇ Topology follows data structure/relations and allow embedding prior information
- ◇ Conditional MRF learning discriminative posteriors

[Boltzmann machines](#)

- ◇ The missing link between probabilistic (MRF) and neural models (RNNs)
- ◇ The originator of the whole deep learning fuzz
- ◇ A good way to see Gibbs sampling at work