Sampling methods

Intelligent Systems for Pattern Recognition (ISPR)

Daniele Castellana

Department of Computer Science
Università di Pisa
daniele.castellana@di.unipi.it

Outline

- **Recap probabilistic concepts**
- **Sampling**
  - What is it?
  - Why do we need it?
  - Properties of samplers
- **Sampling from univariate distributions**
- **Sampling from multivariate distributions**
  - Ancestor sampling
  - Gibbs Sampling
  - Monte Carlo Markov Chain (MCMC)
  - Other methods

Probability recap

- **Discrete Random Variables**
  - x is a **discrete random variable** with C state;
  - $p(x = i),\ i \in [1, C]$ is its **probability distribution**;
  - $p(x_1, \ldots, x_n)$ **joint distribution** of n discrete random variable;

- **Expectation**
  - let $f(\cdot)$ a **function** over a random variable x;
  - $E_{p(x)}[f(x)] = \sum_{i=1}^{C} f(i)p(x = i)$ is its **expected value**;

In this lesson, we will focus only on **discrete variables**, but the same results hold in the continue case.

What is sampling?

Sampling consists in drawing a set of **realisations** $\mathcal{X} = \{x^1, \dots, x^L\}$ of a random variable x with distribution $p(x)$.

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

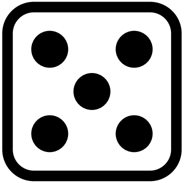
What is sampling?

Sampling consists in drawing a set of **realisations** $\mathcal{X} = \{x^1, \ldots, x^L\}$ of a random variable x with distribution $p(x)$.

Example:

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

l	x^l
1	5

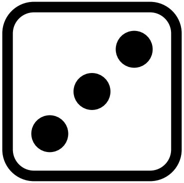


What is sampling?

Sampling consists in drawing a set of **realisations** $\mathcal{X} = \{x^1, \ldots, x^L\}$ of a random variable x with distribution $p(x)$.

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

l	x^l
1	5
2	3

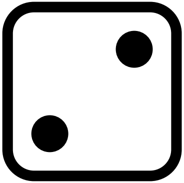


What is sampling?

Sampling consists in drawing a set of **realisations** $\mathcal{X} = \{x^1, \ldots, x^L\}$ of a random variable x with distribution $p(x)$.

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

l	x^l
1	5
2	3
3	2

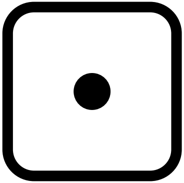


What is sampling?

Sampling consists in drawing a set of **realisations** $\mathcal{X} = \{x^1, \ldots, x^L\}$ of a random variable x with distribution $p(x)$.

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

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

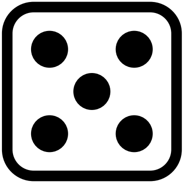


What is sampling?

Sampling consists in drawing a set of **realisations** $\mathcal{X} = \{x^1, \ldots, x^L\}$ of a random variable x with distribution $p(x)$.

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

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

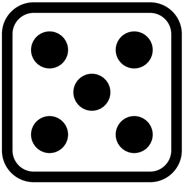


What is sampling?

Sampling consists in drawing a set of **realisations** $\mathcal{X} = \{x^1, \ldots, x^L\}$ of a random variable x with distribution $p(x)$.

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

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



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

Suppose that we want to compute the expectation $E_{p(x)}[f(x)]$.

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

If we have a sample set $\mathcal{X} = \{x^1, \ldots, x^L\}$, then we can approximate the expectation as:

$$E_{p(x)}[f(x)] \approx \frac{1}{L}\sum_{l=1}^{L} f(x^l) \equiv \hat{f}_{\mathcal{X}} \tag{1}$$

There are a lot of cases where $p(x)$ is **intractable**:

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

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:

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

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

In this case, we sample:

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

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

Properties of sampling

The most important properties of sampling are:

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

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

  where $\mathbb{I}[c] = 1$ if and only if c is true;
- the **sampling approximation** $\hat{f}_{\mathcal{X}}$ of the expectation can be an **unbiased estimator**;
- the **sampling approximation** $\hat{f}_{\mathcal{X}}$ of the expectation can have **low variance**;

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

Quick refresh:

Unbiased estimator $\hat{\theta}$ of the unknown θ. $\Longrightarrow$ the approximation is exact on average.

Let $\tilde{p}(\mathcal{X})$ the **distribution over all possible realisations** of the sampling set $\mathcal{X}$, then $\hat{f}_{\mathcal{X}}$ is an **unbiased estimator** if:

$$E_{\tilde{p}(\mathcal{X})}\left[\hat{f}_{\mathcal{X}}\right] = E_{p(x)}\left[f(x)\right]. \tag{2}$$

Unbiased Sampling Approximation

Quick refresh:

Unbiased estimator $\hat{\theta}$ of the unknown θ. $\Longrightarrow$ the approximation is exact on average.

Let $\tilde{p}(\mathcal{X})$ the **distribution over all possible realisations** of the sampling set $\mathcal{X}$, then $\hat{f}_{\mathcal{X}}$ is an **unbiased estimator** if:

$$E_{\tilde{p}(\mathcal{X})}\left[\hat{f}_{\mathcal{X}}\right] = E_{p(x)}\left[f(x)\right]. \tag{2}$$

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

The proof is given in the Appendix.

Unbiased Sampling Approximation

Quick refresh:

Unbiased estimator $\hat{\theta}$ of the unknown θ. $\Longrightarrow$ the approximation is exact on average.

Let $\tilde{p}(\mathcal{X})$ the **distribution over all possible realisations** of the sampling set $\mathcal{X}$, then $\hat{f}_{\mathcal{X}}$ is an **unbiased estimator** if:

$$E_{\tilde{p}(\mathcal{X})}\left[\hat{f}_{\mathcal{X}}\right] = E_{p(x)}\left[f(x)\right]. \tag{2}$$

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

The proof is given in the Appendix.

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

Variance of Sampling Approximation
Definition

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

Let

$$\Delta \hat{f}_{\mathcal{X}} = \hat{f}_{\mathcal{X}} - E_{\tilde{p}(\mathcal{X})}\left[\hat{f}_{\mathcal{X}}\right], \tag{3}$$

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

$$E_{\tilde{p}_{\mathcal{X}}}\left[\left(\Delta \hat{f}_{\mathcal{X}}\right)^2\right].$$

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

Let

$$\Delta\hat{f}_{\mathcal{X}} = \hat{f}_{\mathcal{X}} - E_{\tilde{p}(\mathcal{X})}\left[\hat{f}_{\mathcal{X}}\right], \tag{3}$$

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

$$E_{\tilde{p}_{\mathcal{X}}}\left[\left(\Delta\hat{f}_{\mathcal{X}}\right)^2\right].$$

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

Variance of Sampling 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

$$E_{\tilde{p}_{\mathcal{X}}}\left[\left(\Delta\hat{f}_{\mathcal{X}}\right)^2\right] = \frac{1}{L}\mathrm{Var}_{p(x)}[f(x)]. \tag{4}$$

The proof is given in the Appendix.

Variance of Sampling 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

$$E_{\tilde{p}_{\mathcal{X}}}\left[\left(\Delta\hat{f}_{\mathcal{X}}\right)^2\right] = \frac{1}{L}\mathrm{Var}_{p(x)}[f(x)]. \quad (4)$$

The proof is given in the Appendix.

> We can reduce the variance using a **small** number of samples!

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

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

$$\mathbf{p}(\mathcal{X}) \neq \mathbf{p}(\mathbf{x})$$

$p(x)$ $\longrightarrow$ The distribution **to** sample $\longrightarrow$ **Does not depdend** on the **sampling procedure.**

$\tilde{p}(\mathcal{X})$ $\longrightarrow$ The distribution **of** the samples. $\longrightarrow$ **Depdends** on the **sampling procedure.**

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 $\tilde{p}(\mathcal{X})$:
  - $\tilde{p}(x^l) = p(x^l) \implies$ **valid sampler**;
  - $\tilde{p}(x^l, x^{l'}) = \tilde{p}(x^l)\tilde{p}(x^{l'}) \implies$ **low approximation variance**.

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 $\tilde{p}(\mathcal{X})$:
  - $\tilde{p}(x^l) = p(x^l) \implies$ **valid sampler**;
  - $\tilde{p}(x^l, x^{l'}) = \tilde{p}(x^l)\tilde{p}(x^{l'}) \implies$ **low approximation variance**.

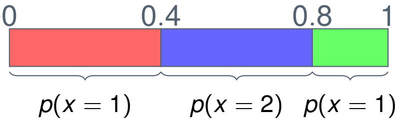
In the **next slides**, we introduce examples of **sampling procedures**:

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

Drawing samples from an **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 $\{s_1, \ldots, s_n\}$, where each s_i is a **discrete variable** with C states.

Hence, each sample x^l contains n values.

$\mathcal{X}$	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 an **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**!

Naive Multivariate Sampling - 1

We build an **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:

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

Them, we sample the variables in the **following order**:

1. sample $\tilde{s}_1 \sim p(s_1)$;
2. sample $\tilde{s}_2 \sim p(s_2 \mid \tilde{s}_1)$;
3. sample $\tilde{s}_3 \sim p(s_3 \mid \tilde{s}_1, \tilde{s}_2)$;

$\vdots$

n. sample $\tilde{s}_n \sim p(s_n \mid \tilde{s}_1, \dots, \tilde{s_{n-1}})$.

Naive Multivariate Sampling - 2

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

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

Them, we sample the variables in the **following order**:

1. sample $\tilde{s}_1 \sim p(s_1)$;
2. sample $\tilde{s}_2 \sim p(s_2 \mid \tilde{s}_1)$;
3. sample $\tilde{s}_3 \sim p(s_3 \mid \tilde{s}_1, \tilde{s}_2)$;

   ⋮

n. sample $\tilde{s}_n \sim p(s_n \mid \tilde{s}_1, \ldots, \tilde{s}_{n-1})$.

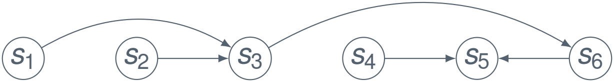
Easy because **univariate**!

Unfortunately, computing the distribution $p(s_i \mid s_{j<i})$ can require summation over an **exponential number of states**!

Ancestral Sampling

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

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

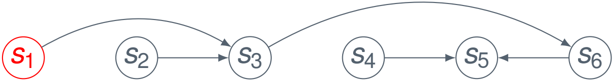


The **BN ancestral order** tell us the sampling order.

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

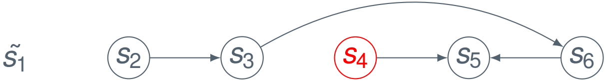
$$\{s_1, s_2, s_4, \} \prec \{s_3\} \prec \{s_6\} \prec \{s_5\}$$

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



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

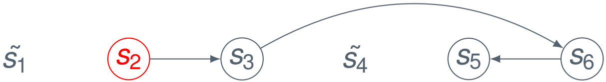
Sample $\tilde{s}_4 \sim p(s_4)$



Ancestral Sampling
Example

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

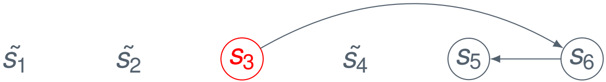
Sample $\tilde{s}_2 \sim p(s_2)$



Ancestral Sampling
Example

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

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



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

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

$$\tilde{s}_1 \qquad \tilde{s}_2 \qquad \tilde{s}_3 \qquad \tilde{s}_4 \qquad s_5 \leftarrow s_6$$

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

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

$\tilde{s}_1$ $\tilde{s}_2$ $\tilde{s}_3$ $\tilde{s}_4$ s_5 $\tilde{s}_6$

$$\{\cancel{S_1}, \cancel{S_2}, \cancel{S_4},\} \prec \{\cancel{S_3}\} \prec \{\cancel{S_6}\} \prec \{\cancel{S_5}\}$$

$$\tilde{s}_1 \quad \tilde{s}_2 \quad \tilde{s}_3 \quad \tilde{s}_4 \quad \tilde{s}_5 \quad \tilde{s}_6$$

This is a single sample x^l!

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

$$\tilde{s}_1 \quad \tilde{s}_2 \quad \tilde{s}_3 \quad \tilde{s}_4 \quad \tilde{s}_5 \quad \tilde{s}_6$$

This is a single sample x^l!

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

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

$$\tilde{s}_1 \quad \tilde{s}_2 \quad \tilde{s}_3 \quad \tilde{s}_4 \quad \tilde{s}_5 \quad \tilde{s}_6$$

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_ϵ are **visible**; writing $s = s_\epsilon \cup s_{\setminus \epsilon}$, we would like to sample from:

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

Can we still use AS?

Sampling with evidence

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

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

Can we still use AS?

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

Computing the new structure is complex as running exact inference!

Sampling with evidence

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

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

Can we still use AS?

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

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 needing of new sampling procedures

Sampling under evidence is **important**!

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

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

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.

The idea is to start from a sample $x^1 = \{s_1^1, \ldots, s_n^1\}$ 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

The idea is to start from a sample $x^1 = \{s_1^1, \ldots, s_n^1\}$ 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

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

The idea is to start from a sample $x^1 = \{s_1^1, \ldots, s_n^1\}$ 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

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

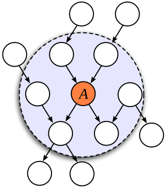
Sample	s_1	s_2	s_3	s_4	s_5
x^1	1	1	2	4	5
x^2	3	1	2	4	5
x^3	3	4	2	4	5
x^4	3	4	2	1	5
x^5	3	4	6	1	5

The idea is to start from a sample $x^1 = \{s_1^1, \ldots, s_n^1\}$ 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
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

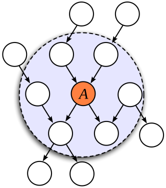
During the $(l + 1)$-th iteration,

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



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

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



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

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

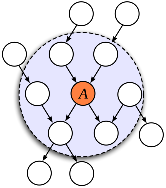
$$s_j^{l+1} \sim p(s_j \mid s_{\setminus j}) = \frac{1}{Z} p(s_j \mid pa(s_j)) \prod_{k \in ch(j)} p(s_k \mid pa(s_k)),$$

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

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

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

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



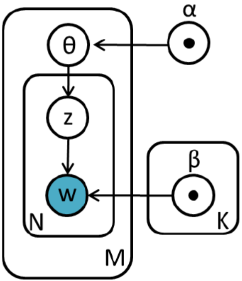
$$s_j^{l+1} \sim p(s_j \mid s_{\setminus j}) = \frac{1}{Z} p(s_j \mid pa(s_j)) \prod_{k \in ch(j)} p(s_k \mid pa(s_k)),$$

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!

LDA Gibbs Sampling



Start form initial guess $\{z_{ij}^0, \theta_i^0, \beta^0\}$.
Do:

1. $z_{ij}^{l+1} \sim P(z_{ij} \mid \mathbf{w}, z_{-ij}^l, \theta^l, \beta^l, \alpha)$
2. $\theta_i^{l+1} \sim P(\theta_i \mid \mathbf{w}, \mathbf{z}^{l+1}, \beta^l, \alpha)$
3. $\beta^{l+1} \sim P(\beta \mid \mathbf{w}, \mathbf{z}^{l+1}, \theta^{l+1}, \alpha)$

Repeat until convergence.

The **derivation** of the sampling formulas can require strong **mathematical skills**!

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

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

- **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 $p(x)$.

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

- **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 $L \to \infty$, the series $\{x^1, \ldots, x^L\}$ **converges** to samples taken from $p(x)$!

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

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

- **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 $L \to \infty$, 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**!

MCMC Sampling Framework

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

The idea is to build a **Markov Chain** whose stationary distribution is $p(x)$.

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

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

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

MCMC Sampling Framework

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

The idea is to build a **Markov Chain** whose stationary distribution is $p(x)$.

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

- **irreducible** $\rightarrow$ it is possible to reach any state from anywhere;
- **aperiodic** $\rightarrow$ 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
- Metropolis-Hastings Sampling
- Particle Filtering
- Hybrid Monte Carlo
- Swendson-Wang

  $\vdots$

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

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** $p(s_j \mid s_{\setminus j})$!
  - it works well when variables are **not strogly related**!
- Metropolis-Hastings Sampling
- Particle Filtering
- Hybrid Monte Carlo
- Swendson-Wang

  ⋮

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

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** $p(s_j \mid s_{\setminus j})$!
  - it works well when variables are **not strogly related**!
- Metropolis-Hastings Sampling
  - based on a **proposal distribution** $\tilde{q}(x^{l+1} \mid x^l)$!
  - the choice of $\tilde{q}(\cdot)$ **is crucial**!
- Particle Filtering
- Hybrid Monte Carlo
- Swendson-Wang

  ⋮

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

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

$\vdots$

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

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** $\rightarrow$ sampling is **easy**!
- $p(x)$ **multivariate** $\rightarrow$ 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** sampling procedure:
  - we can **easily** deal with evidence;
  - the sampler defines a **MC** whose **stationary** distr. is $p(x)$;
    - the sampler is valid in the limit $l \rightarrow \infty$;
  - **different** state-trans. q leads to **different** procedures:
    - **Gibbs Sampling**, Metropolis-Hastings Sampling, ...

See Chapter 27 of **BRML** book!

Appendix

In the following slides, we provide:

- some properties of the **expectation** that are useful in the proofs;
- the proof of the **average approximation**;
- the proof of the **variance approximation**;

Expectation properties

The following property are used during the proofs:

- **Linearity**:

$$\begin{aligned} E_{p(x)}\left[f(x)+g(x)\right] = E_{p(x)}\left[f(x)\right] + E_{p(x)}\left[g(x)\right] \\ E_{p(x)}\left[c\, f(x)\right] = c\, E_{p(x)}\left[f(x)\right]. \end{aligned} \tag{5}$$

- **Expected value of a constant**:

$$E_{p(x)}\left[c\right] = c. \tag{6}$$

Also, we use the symbol $\overset{(n)}{=}$ to indicate that the statement in the equation (n) is used to make a step in the proof.

Proof Average Approximation

We want to prove that

$$E_{\tilde{p}(\mathcal{X})}\left[\hat{f}_{\mathcal{X}}\right] = E_{p(x)}\left[f(x)\right]. \tag{2}$$

assuming

$$\tilde{p}(x^l) = p(x) \tag{7}$$

Proof.

$$E_{\tilde{p}(\mathcal{X})}\left[\hat{f}_{\mathcal{X}}\right] \stackrel{(1)}{=} E_{\tilde{p}(\mathcal{X})}\left[\frac{1}{L}\sum_{l=1}^{L} f(x^l)\right] \stackrel{(5)}{=} \frac{1}{L}\sum_{l=1}^{L} E_{\tilde{p}(x^l)}\left[f(x^l)\right] =$$

$$\stackrel{(7)}{=} \frac{1}{L}\sum_{l=1}^{L} E_{p(x)}\left[f(x)\right] = \frac{1}{L} \times L \times E_{p(x)}\left[f(x)\right] = E_{p(x)}\left[f(x)\right].$$

Proof Variance Approximation I

We want to prove that

$$E_{\tilde{p}_{\mathcal{X}}}\left[\left(\Delta\hat{f}_{\mathcal{X}}\right)^2\right] = \frac{1}{L}\mathrm{Var}_{p(x)}[f(x)]. \tag{4}$$

assuming

$$\tilde{p}(x^l) = p(x) \tag{7}$$

$$\tilde{p}(x^i, x^j) = \tilde{p}(x^i)\tilde{p}(x^j) \tag{8}$$

This holds due to assumption (7)!

Proof.

$$\Delta\hat{f}_{\mathcal{X}} \overset{(3)}{=} \hat{f}_{\mathcal{X}} - E_{\tilde{p}(\mathcal{X})}\left[\hat{f}_{\mathcal{X}}\right] \overset{(1)+(2)}{=} \frac{1}{L}\sum_{l=1}^{L} f(x^l) - E_{p(x)}\left[f(x)\right] =$$

$$= \frac{1}{L}\sum_{l=1}^{L} f(x^l) - \frac{1}{L}\sum_{l=1}^{L} E_{p(x)}\left[f(x)\right] = \frac{1}{L}\sum_{l=1}^{L}\left(f(x^l) - E_{p(x)}\left[f(x)\right]\right) \tag{9}$$

Proof Variance Approximation II

Then, naming

$$\Delta f(x^l) = f(x^l) - E_{p(x)}\left[f(x)\right], \tag{10}$$

we obtain

$$\begin{aligned}
E_{\tilde{p}_{\mathcal{X}}}\left[\left(\Delta\hat{f}_{\mathcal{X}}\right)^2\right] &\overset{(9)}{=} E_{\tilde{p}_{\mathcal{X}}}\left[\left(\frac{1}{L}\sum_{l=1}^{L}\left(f(x^l) - E_{p(x)}\left[f(x)\right]\right)\right)^2\right] \overset{(10)}{=} \\
&= E_{\tilde{p}_{\mathcal{X}}}\left[\left(\frac{1}{L}\sum_{l=1}^{L}\Delta f(x^l)\right)^2\right] = E_{\tilde{p}_{\mathcal{X}}}\left[\frac{1}{L^2}\sum_{l,l'}\Delta f(x^l)\Delta f(x^{l'})\right] \overset{(5)}{=} \\
&= \frac{1}{L}E_{\tilde{p}(x)}\left[(\Delta f(x))^2\right] + \frac{1}{L^2}\sum_{l\neq l'}E_{\tilde{p}(x^l,x^{l'})}\left[\Delta f(x^l)\Delta f(x^{l'})\right].
\end{aligned} \tag{11}$$

Proof Variance Approximation III

The term

$$
\begin{aligned}
&\frac{1}{L^2}\sum_{l\neq l'} E_{\tilde{p}(x^l,x^{l'})}\left[\Delta f(x^l)\Delta f(x^{l'})\right] \overset{(8)}{=} \\
&= \frac{1}{L^2}\sum_{l\neq l'} E_{p(x^l)}\left[\Delta f(x^l)\right] E_{p(x^{l'})}\left[\Delta f(x^{l'})\right] = 0,
\end{aligned}
\tag{12}
$$

since

$$
\begin{aligned}
E_{p(x^l)}\left[\Delta f(x^l)\right] &\overset{(10)}{=} E_{p(x^l)}\left[f(x^l) - E_{p(x)}\left[f(x)\right]\right] \overset{(5)+(6)}{=} \\
&= E_{p(x)}\left[f(x)\right] - E_{p(x)}\left[f(x)\right] = 0.
\end{aligned}
$$

Finally, combining (11) and (12), we obtain

$$
E_{\tilde{p}_{\mathcal{X}}}\left[\left(\Delta \hat{f}_{\mathcal{X}}\right)^2\right] = \frac{1}{L} E_{\tilde{p}(x)}\left[(\Delta f(x))^2\right] = \frac{1}{L}\mathrm{Var}_{p(x)}[f(x)].
$$