Montecarlo methods recall

Course of Machine Learning
Master Degree in Computer Science
University of Rome "Tor Vergata"
a.a. 2021-2022

Giorgio Gambosi

The basic problem

Integrate a hard (for example high dimensional) function

$$\int_a^b g(x)dx$$

Idea
See the integral as an expectation

Approach
Assume we have a function $f(x)$ and a density $p(x)$ in $[a,b]$ such that $g(x) = f(x)p(x)$, we may write

$$\int_a^b g(x)dx = \int_a^b f(x)p(x)dx = E_{p(x)}[f(x)]$$

and approximate this value through the mean of n values $f(x_1), \dots, f(x_n))$ sampled from $p(x)$:

$$E[f(x)] \approx \frac{1}{n}\sum_{i=1}^n f(x_i)$$

Approach

1. Sample a sequence $x^{(1)}, x^{(2)}, \dots, x^{(n)}$ of values from distribution $p(x)$, that is such that $Pr(X = x^{(i)}) = p(x^{(i)})$
2. Apply function $f(x)$ to such values
3. Average the set of values obtained

Sampling for expectations

Problem
But how to sample values from $p(x)$?

$$E_p[f(x)] = \int_x f(x)p(x)dx$$

where $p(x)$ is hard to derive analitically.

Hypothesis
Assume a (pseudo) random generator $\mathcal{R}$ is available which returns a sequence of values (approximately) uniformly distributed in the interval $[0, 1]$.

Given a distribution $p(x)$, the sequence of values provided by $\mathcal{R}$ can be exploited to derive a different (possibly shorter) sequence of values with distribution $p(x)$

General issue

Problem underlying this method:

How can we sample from any distribution, especially if we do not have an analytical representation of it?

Sampling: easy case

Assume we know $p(x)$: we wish to find a function $f(z)$ such that if $z \sim U(0, 1)$, then $f(z) \sim p(z)$.

- This is equivalent to saying that for each $z \in [0, 1]$ the cumulative probability of z wrt to the uniform distribution (which is z itself), must be equal to the cumulative probability of $f(z)$ wrt to $p(z)$ (which, by definition, is $Pr(\zeta \leq f(z))$)
- that is,

  $$z = \int_0^z d\zeta = \int_{-\infty}^{f(z)} p(\zeta)d\zeta = P(f(z))$$

  where $P(t)$ is the cumulative distribution of t if $t \sim p(x)$
- as a corollary, since it results $z = P(f(z))$, we have that $f(z) = P^{-1}(z)$

Sampling in the easy case: an example

Example
Given $\mathcal{R}$, we use it to produce exponentially distributed values, that is values distributed according to

$$p(x|\lambda) = \lambda e^{-\lambda x}, 0 \leq x < \infty$$

The exponential cumulative distribution is

$$P(x) = \int_0^x \lambda e^{-\lambda\xi} d\xi = 1 - e^{-\lambda x}$$

by setting $z = P(f(z)) = 1 - e^{-\lambda f(z)}$ we get

$$\begin{aligned} e^{-\lambda f(z)} &= 1 - z \\ -\lambda f(z) &= \ln(1 - z) \\ f(z) &= -\frac{1}{\lambda}\ln(1 - z) \end{aligned}$$

Sampling on general distributions

Approaches
Applying the above method is possible only for simple distributions. In most cases, you cannot immediately derive values distributed according to $p(x)$

Many sampling methods have been introduced

- rejection sampling
- importance sampling

- adaptive rejection sampling
- sampling-importance-sampling
- ...

Rejection sampling

Context

- $p(x)$ is difficult to sample, for example we have no analytical definition
- there exists $q(x)$
  - easier to sample
  - there exists K such that $Kq(x) \geq p(x)$, $\forall x$

Method

1. Sample $\overline{x} \sim q(x)$
2. Sample $u^* \sim U(0, Kq(\overline{x}))$
3. If $u^* \leq p(\overline{x})$ accept and return the sample, otherwise discard it

Importance sampling

Context
Assume we have to approximate the expectation of $f(x)$ wrt $p(x)$

$$E_{p(x)}[f(x)] = \int_x f(x)p(x)dx$$

If we are able to sample n values $x^{(i)}$ from $p(x)$ then we may apply the approximation

$$\int_x f(x)p(x)dx \approx \frac{1}{n}\sum_{i=1}^{n} f(x^{(i)})$$

Assume $p(x)$ hard to sample, but easy to evaluate for all x. Let $q(x)$ be some easy distribution to sample. Then,

$$\int_x f(x)p(x)dx = \int_x \frac{f(x)}{q(x)}p(x)q(x)dx \approx \frac{1}{n}\sum_{i=1}^{n} \frac{p(x^{(i)})}{q(x^{(i)})} f(x^{(i)})$$

The ratios $p(x^{(i)})/q(x^{(i)})$ are sampling weights, associated to samples from $q(x)$ (hence avoiding sampling from $p(x)$)

Markov chains

Definition
Given a (possibly infinite) sequence of random variables $\mathbf{X} = (X_0, X_1, \ldots)$ and a *state space* $\mathcal{X}$ of possible values for all $X_i \in \mathbf{X}$, a *Markov chain* on $\mathbf{X}$ is a stochastic process which defines for each ordered pair $< x_i, x_j > \in \mathcal{X}^2$, a probability p_{ij} of transition from x_i to x_j such that $p(X_t = x_i | X_{t-1} = x_j) = p_{ij}$, for all $t > 0$.

State probability
Given an initial state, that is a vaue assigned to X_0, the distribution $p(X_t = x_j|X_0 = x_k)$ of each random variable X_t on the set of state can be easily obtained (by matrix multiplication).

Markov chains

Stationary distribution
Under suitable conditions on its structure, a Markov chain is *ergodic*, that is the probability $p(X_t = x_j|X_0 = x_k)$, as $n \to \infty$,

- is independent from the initial state

$$p(X_t = x_j|X_0 = x_k) = p(X_t = x_j)$$

- is stationary

$$p(X_t = x_j) = p(X_{t+1} = x_j)$$

Markov chain Montecarlo (MCMC)

Idea
Given a hard to sample distribution $p(x)$, derive an ergodic Markov chain such that:

- its transition probability $q(x_i|x_{i-1})$ is easy to sample
- stationary distribution is $p(x)$

Markov chain Montecarlo (MCMC)

Ho to use it?
Given the Markov chain,

- a sequence of random transitions is performed, starting from any initial state (value of x).
- apply a certain number of initial transitions (*burn-in time*)
- after that, at each step the value $\overline{x}$ reached by the MC is tested wrt a predefined criterion: if the test is positive, the value is returned

The returned values are (approximately) distributed as $p(x)$: hence their sequence can be used as a sequence of samplings from $p(x)$

MCMC methods
Several MCMC methods have been defined, differing each other by the structure of the chain and the acceptance criterion applied.

Metropolis algorithm

Idea
After the burning time, let $x^{(i-1)}$ be the current state and let $\overline{x}$ be the value produced by a random transition from $x^{(i-1)}$, obtained by sampling $q(x|x^{(i-1)})$

$\overline{x}$ is accepted, and returned as a sample, with probability

$$A(\overline{x}, x^{(i-1)}) = \text{min} \left(1, \frac{p(\overline{x})}{p(x^{(i-1)})}\right)$$

Notice that if $\overline{x}$ has higher probability than $x^{(i-1)}$ with respect to the target distribution $p(x)$, it is accepted, while if its probability is smaller, it is accepted with probability equal to the ratio between them.

If $\overline{x}$ is accepted, then $x^i = \overline{x}$ becomes the current state, otherwise the current state is not modified, that is $x^{(i)} = x^{(i-1)}$)

Note
Observe that the same holds if $\pi(x) = Kp(x)$ is applied in the definition of $A(\overline{x}, x^{(i-1)})$; observe also that the value of K needs not being known

Metropolis algorithm

Structure
At the i-th iteration:

1. Sample a value $\overline{x}$ from $q(x|x^{(i-1)})$
2. With probability $A(\overline{x}, x^{(i-1)})$
   - let $x^{(i)} = \overline{x}$, return $\overline{x}$
   - else let $x^{(i)} = x^{(i-1)}$

Note
For any pair x, x', the real probability of transition from x to x' is given by the *transition kernel*

$$T(x|x') = q(x|x')A(x, x')$$

Metropolis algorithm

Detailed balance
Given the target distribution $p(x)$, a Markov chain has the *detailed balance* property with respect to $p(x)$, if for each x, x',

$$p(x)q(x'|x) = p(x')q(x|x')$$

that is the probability that at a certain step the current state is x and the following state is x' is equal to the one that the current state is x' and the next state is x.

In this case, $p(x)$ is the stationary distribution of the Markov chain. In fact, let us remind that if $p^*(x)$ is the stationary distribution then by definition

$$p^*(x) = \sum_{x'} q(x|x')p^*(x')$$

and for $p(x)$ we have

$$p(x) = \sum_{x'} q(x|x')p(x') = \sum_{x'} q(x'|x)p(x) = p(x)\sum_{x'} q(x'|x) = p(x)$$

Metropolis algorithm

Uniqueness of the stationary distribution
Even in the case that $p(x)$ is a stationary distribution, we must be sure that the Markov chain tends to $p(x)$ for any initial state, that is that it is ergodic.

A sufficient condition for ergodicity is that for all pairs x, x' the transition probability is positive, that is $q(x|x') > 0$

Metropolis algorithm

Why it does work

- Assume the transition probability distribution is
  - symmetric : $q(x|x') = q(x'|x)$, $\forall x, x'$
  - positive: $q(x|x') > 0$, $\forall x, x'$
- then
  - for the probability $T(x|x') = q(x|x')A(x, x')$ the detailed balance property holds wrt $p(x)$

$$\begin{aligned}
p(x)T(x'|x) = p(x)q(x'|x)A(x', x) &= \text{min}\left(p(x)q(x'|x), \frac{p(x)q(x'|x)p(x')}{p(x)}\right) \\
&= \text{min}\,(p(x)q(x'|x), p(x')q(x'|x)) = \text{min}\,(p(x)q(x|x'), p(x')q(x|x')) \\
&= \text{min}\left(\frac{p(x')q(x|x')p(x)}{p(x')}, p(x')q(x|x')\right) = p(x')q(x|x')A(x, x') \\
&= p(x')T(x|x')
\end{aligned}$$

    hence $p(x)$ is a stationary distribution
  - all transition probabilities are positive, hence the chain is ergodic and always tends to $p(x)$

Metropolis-Hastings algorithm

Idea

- Applied for non symmetric $q(x|x')$
- In this case, the transition kernel $T'(x|x') = q(x|x')A'(x, x')$ refers to

$$A'(x, x') = \text{min}\left(1, \frac{p(x)q(x'|x)}{p(x')q(x|x')}\right)$$

Why does it works
The detailed balance property still holds for the transition kernel

$$\begin{aligned}
p(x)T'(x'|x) = p(x)q(x'|x)A'(x', x) &= \text{min}\left(p(x)q(x'|x), \frac{p(x)q(x'|x)p(x')q(x|x')}{p(x)q(x'|x)}\right) \\
&= \text{min}\,(p(x)q(x'|x), p(x')q(x|x')) = p(x')q(x|x')A'(x, x') = p(x')T'(x|x')
\end{aligned}$$

Transition probability

Structure
Anything goes, as soon as the detailed balance property is verified. Usually, built on a set of base transitions $B_k(x|x')$

- Mixture of base transitions

$$T(x|x') = \sum_{i=1}^{K} \alpha_i B_i(x|x')$$

with $\alpha_i \geq 0, \forall i$ and $\sum_{i=1}^{K} \alpha_i = 1$

- Sequence of applications

$$T(x|x') = \sum_{x_1} \sum_{x_2} \ldots \sum_{x_{K-1}} B_1(x|x_1) B_2(x_1|x_2) \ldots B_K(x_{K-1}|x')$$

Gibbs sampling

Use
Gibbs sampling is a MCMC applied in cases when:

- x has dimensionality at least 2, $\mathbf{x} = (x_1, \ldots, x_m)$, with $m > 1$
- for all $i = 1, \ldots, m$, the conditional distribution $p(x_i|\mathbf{x}_{-i})$ is easy to sample, where $\mathbf{x}_{-i} = \{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_m\}$

Idea
Instead of sampling the next state in a single step from $q(\mathbf{x}|\mathbf{x}')$, a sequence of m transitions is sampled, each wrt a component x_i of $\mathbf{x}$ and to distribution $p(x_i|\mathbf{x}_{-i})$.

The basic idea in Gibbs sampling is that rather than probabilistically picking the next state of all at once, a separate probabilistic choice is performed for each of the m dimensions, with each choice depending on the other $k-1$ dimensions.

Gibbs sampling

Algorithm structure

- Sample m values for the initial state $\mathbf{x}^{(0)} = (x_1^{(0)}, \ldots, x_m^{(0)})$
- For $i = 1, \ldots, T$
  - For $k = 1, \ldots, m$ sample $x_k^{(i)}$ from

$$\begin{aligned} p(x_k|\mathbf{x}_{-k}^{(i)}) &= p(x_k|x_{-1}^{(i)}, \ldots, x_{k-1}^{(i)}, x_{k+1}^{(i-1)}, \ldots, x_m^{(i-1)}) \\ &= \frac{p(x_1^{(i)}, \ldots, x_{k-1}^{(i)}, x_k^{(i-1)}, x_{k+1}^{(i-1)}, \ldots, x_m^{(i-1)})}{p(x_1^{(i)}, \ldots, x_{k-1}^{(i)}, x_{k+1}^{(i-1)}, \ldots, x_m^{(i-1)})} \end{aligned}$$

  - set $\mathbf{x}^{(i)} = (x_1^{(i)}, \ldots, x_m^{(i)})$

Gibbs sampling

Why it does work

- it is possible to prove that $p(\mathbf{x})$ is a stationary distribution of the Markov chain
- also, if distributions $p(x_i|\mathbf{x}_{-i})$ are never equal to zero, the chain is ergodic, and tends to $p(\mathbf{x})$

MCMC and bayesian models

- MCMC can be applied (as it frequently happens) in bayesian inference by observing that the posterior distribution is defined as
$$p(\theta|\mathbf{X}) = \frac{p(\mathbf{X}|\theta)p(\theta)}{p(\mathbf{X})} = \frac{p(\mathbf{X}|\theta)p(\theta)}{Z}$$
where Z is usually hard to compute
- Let us remind that MCMC is able to sample a distribution $p(\mathbf{x})$ assuming that a proportional function $\pi(\mathbf{x}) = Kp(\mathbf{x})$ can be evaluated, for any unknown K
- Thus, samples of the posterior distribution of parameters can be obtained if both the prior $p(\theta)$ and the likelihood $p(\mathbf{X}|\theta) = \prod_i p(\mathbf{x}_i|\theta)$ can be evaluated for any value θ

Sampling the evidence

- Actually, the evidence
$$p(\mathbf{X}) = \int p(\mathbf{X}|\theta)p(\theta)d\theta$$
could be explicitly evaluated, if necessary, as the average of a set of m values
$$p(\mathbf{X}|\theta_i) \qquad i = 1, \ldots, m$$
computed from the set of samples $\theta_1, \ldots, \theta_m$ of $p(\theta)$

Sampling the predictive distribution

- For what regards the predictive distribution
$$p(\mathbf{x}|\mathbf{X}) = \int_\theta p(\mathbf{x}|\theta)p(\theta|\mathbf{X})d\theta$$
the same considerations apply, averaging the set of values
$$p(\mathbf{x}|\theta_i) \qquad i = 1, \ldots, m$$
computed from the set of samples $\theta_1, \ldots, \theta_m$ of the posterior distribution $p(\theta|\mathbf{X})$

Gibbs sampling and Metropolis Hastings

Gibbs sampling as a particular case of MH
Gibbs sampling can be seen as Metropolis Hastings with sequential sampling from base distributions $B_k(x|x') = p(x_k|\mathbf{x}_{(-k)})$. Since $p(\mathbf{x}) = p(\mathbf{x}_{(-k)})p(x_k|\mathbf{x}_{(-k)})$, then

$$\begin{aligned}
A(\mathbf{x}, \mathbf{x}') &= \text{min}\left(1, \frac{p(\mathbf{x})B_k(\mathbf{x}'|\mathbf{x})}{p(\mathbf{x}')B_k(\mathbf{x}|\mathbf{x}')}\right) \\
&= \text{min}\left(1, \frac{p(\mathbf{x}_{(-k)})p(x_k|\mathbf{x}_{(-k)})p(x_k|\mathbf{x}'_{(-k)})}{p(\mathbf{x}'_{(-k)})p(x_k|\mathbf{x}'_{(-k)})p(x_k|\mathbf{x}_{(-k)})}\right) \\
&= \text{min}\left(1, \frac{p(\mathbf{x}_{(-k)})}{p(\mathbf{x}'_{(-k)})}\right) = \text{min}\ (1, 1) = 1
\end{aligned}$$

Transitions are never discarded.