INF367A: Probabilistic machine learning

Lecture 10: Sampling

Pekka Parviainen

University of Bergen

10.3.2020

Outline

Monte Carlo integration

Sampling univariate distributions

Markov chains

Markov chain Monte Carlo (MCMC)
- Gibbs sampling
- Metropolis-Hastings algorithm

Approximating distributions

- What if we do not have a closed-form representation for the posterior?
- Approximate the posterior with a simpler family of distributions
  - Laplace approximation
  - Variational inference (next week)
- Represent the posterior using samples from it
  - MCMC (this week)

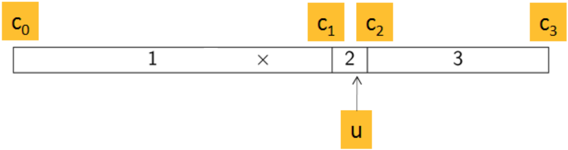
Ways to sample

- Inverse transform sampling
- Rejection sampling
- Importance sampling
- Gibbs sampling (MCMC)
- Metropolis algorithm (MCMC)
- Metropolis-Hastings algorithm (MCMC)
- Hamiltonian Monte Carlo
- Particle filtering
- ...

Sampling from a discrete distribution

- Consider a one-dimensional discrete distribution $P(x)$ where $dom(x) = \{1, 2, 3\}$, with

$$P(x) = \begin{cases} 0.6, & x = 1 \\ 0.1, & x = 2 \\ 0.3, & x = 3 \end{cases}$$



- We can sample from this distribution by sampling uniformly from $[0, 1]$, say $u = 0.66$. Then the sampled state would be state 2, since u is in the interval $(c_1, c_2]$

$$(c_0, c_1, c_2, c_3) = (0, 0.6, 0.7, 1.0)$$

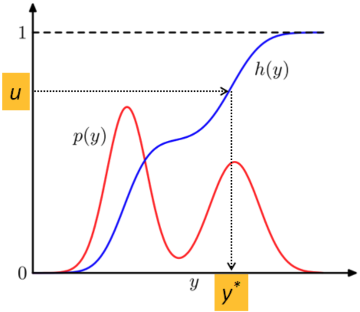
Inverse transform sampling

- The cumulative distribution function (CDF)

$$h(x) = \int_{-\infty}^{x} P(t)\mathrm{d}t$$

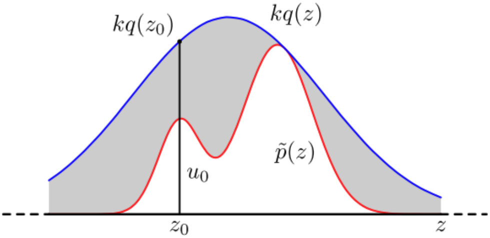
- Samples x^* from distribution P can be drawn as follows:

1. Sample $u \sim Unif(0,1)$
2. Compute x^* using $x^* = h^{-1}(u)$



Rejection sampling

- Use a simple proposal distribution $q(x)$ such that $k \cdot q(x) \geq P(x)$
- Idea: draw uniformly from under the curve $k \cdot q(z)$. Reject the sample if it falls in the gray area. The retained samples are from $P(x)$



- Downside: Might be very ineffective

Monte Carlo integration: Idea

- The goal in Monte Carlo integration is to compute integrals of kind
$$\int h(x)f(x)\mathrm{d}x = E_{f(x)}[h(x)],$$
where $f(x)$ is a probability density function from which we can generate samples
- An approximation is obtained by
$$E_{f(x)}[h(x)] \approx \frac{1}{n}\sum_{i=1}^{n} h(x_i),$$
where $x_i, i = 1, \ldots, n$ are sampled from distribution f
- Justified by the law of large numbers

Example, area of a unit circle

$$
\begin{aligned}
A &= \int_{x^2+y^2<1} \mathrm{d}x\mathrm{d}y = \int_{-1}^{1}\int_{-1}^{1} I(x^2+y^2<1)\mathrm{d}x\mathrm{d}y \\
&= 4 \times \int_{-1}^{1}\int_{-1}^{1} I(x^2+y^2<1) \times \frac{1}{4}\mathrm{d}x\mathrm{d}y \\
&= 4 \times \int_{-1}^{1}\int_{-1}^{1} I(x^2+y^2<1) \times f(x,y)\mathrm{d}x\mathrm{d}y,
\end{aligned}
$$

where

$$
f(x,y) = \begin{cases} \frac{1}{4}, & \text{if } (x,y) \in [-1,1]\times[-1,1] \\ 0, & \text{otherwise} \end{cases}
$$

Thus, we can estimate

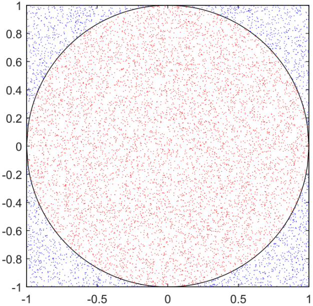
$$
\bar{A} = 4 \times \frac{1}{n}\sum_{i=1}^{n} I(x_i^2 + y_i^2 < 1),
$$

where (x_i, y_i), $i = 1, \ldots, n$ have been sampled uniformly over the square $[-1,1]\times[-1,1]$

Example, area of a unit circle

- Compute the area of the unit circle using Monte Carlo integration
  1. Simulate n points uniformly from the square $[-1,1] \times [-1,1]$
  2. Compute the fraction of points inside the circle
  3. The area of the circle is th fraction times the area of the square

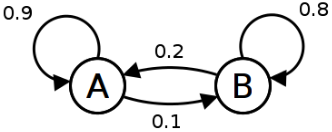
- With $n = 10000$,
  $\text{Area} \approx 3.1672$
  $95\%\ \text{CI} = [3.1347, 3.1997]$



Markov chains

- A Markov chain is a sequence of random variables $X_0, X_1, X_2, \ldots$ that satisfies the Markov property
- **Markov property**: The current state (value of the variable) depends only on a finity history of the previous states
- For example, first order Markov chain

$$P(X_t \mid X_{0:t-1}) = P(X_t \mid X_{t-1})$$



Markov chains

- A Markov chain is specified by the transition probabilities

$$T(x\,|\,y)$$

- On the previous slide, $T(x^{(t+1)} = A|x^{(t)} = A) = 0.9$,
  $T(x^{(t+1)} = B\,|\,x^{(t)} = A) = 0.1$,
  $T(x^{(t+1)} = A|\,x^{(t)} = B) = 0.2$, and
  $T(x^{(t+1)} = B\,|\,x^{(t)} = B) = 0.8$
- A Markov chain is *irreducible* if it possible to move from any state to any other state (directly or indirectly)
- A state x has period k if any return to the state x has to occur in multiples of k time steps. If $k = 1$ then the state is aperiodic. A Markov chain is *aperiodic* if it has at least one aperiodic state

Markov chains - Theory

- Markov chain is called *ergodic* if the distribution of the chain eventually converges to a stationary distribution
- Ergodicity follows from weak conditions (*irreducibility*, *aperiodicity*)
- $P^*(x)$ is the **stationary distribution** of the chain, if it satisfies the *detailed balance*

  $$P^*(x)T(y\,|\,x) = P^*(y)T(x\,|\,y),$$

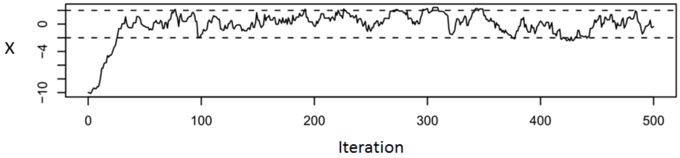
  for all x and y
- In other words, if we simulate the chain long enough, the distribution of the samples converges to the stationary distribution

MCMC - the idea

- Simulate a sequence of samples $x^{(1)}, x^{(2)}, \ldots$ (Monte Carlo) such that the next sample depends only on the previous sample (Markov chain)

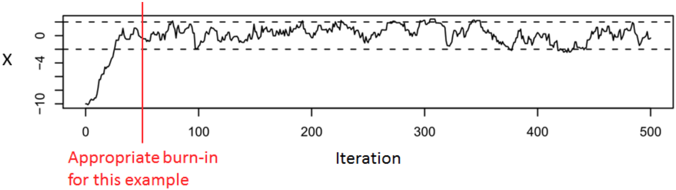
$$P(x^{(t+1)} | x^{(1)}, \ldots, x^{(t)}) = T(x^{(t+1)} | x^{(t)})$$

- By selecting the transition probabilities $T(x^{(t+1)} | x^{(t)})$ appropriately, the chain can be made to converge to a desired distribution P
- Example: Markov chain to sample from N(0, 1)



MCMC - the idea

- The early samples before the chain is converged must be discarded (**burn-in**)



MCMC - the idea

- If the stationary distribution of the Markov chain has density f, then we can approximate as before using

$$E_{f(x)}[h(x)] = \int h(x)f(x)\mathrm{d}x \approx \frac{1}{n-t}\sum_{i=t+1}^{n} h(x^{(i)}),$$

  where $X^{(i)}$, $i = t, \ldots, n$ are samples from the chain after the burn-in t
- This holds even if the samples $x^{(t+1)}, x^{(t+2)}, \ldots$ are dependent. However, the variance of the estimator increases compared to independent samples
- **Thinning** (saving, e.g., every 10th sample only) reduces dependency

MCMC - Convergence and mixing

- **Convergence**: How long does it take for the chain to start producing samples from the stationary distribution
  - The slower the chain converges, the longer burn-in you need
  - Diagnosing convergence is usually difficult
- **Mixing**: How correlated the samples from the target distribution are. The higher the correlation is, the smaller the *effective sample size* is

Markov chain Monte Carlo (MCMC)

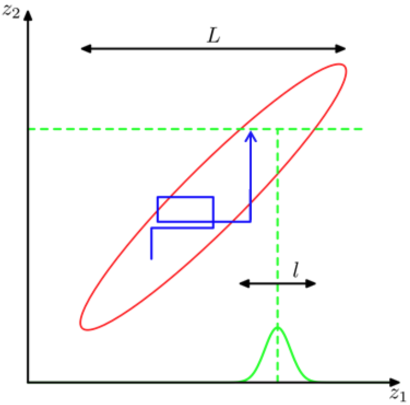
- Constructing the Markov chain corresponds to selecting the transition probabilities $T(\theta_{t+1} \mid \theta_t)$ in such a way that the stationary distribution corresponds to the distribution of interest $P(\theta \mid D)$
- Common ways to do this:
  - Gibbs sampling
  - Metropolis sampling
  - Metropolis-Hastings sampling

Gibbs sampling

- Gibbs sampling consists of simulating a multivariate distribution by iteratively sampling from the full conditional distribution of one variable given all the other variables
- For example, sample from $P(\theta_1, \theta_2, \theta_3)$

1. Initialize $(\theta_1^{(1)}, \theta_2^{(1)}, \theta_3^{(1)})$
2. For $t = 1, \ldots, T$
   - sample $\theta_1^{(t+1)} \sim P(\theta_1 \,|\, \theta_2^{(t)}, \theta_3^{(t)})$
   - sample $\theta_2^{(t+1)} \sim P(\theta_2 \,|\, \theta_1^{(t+1)}, \theta_3^{(t)})$
   - sample $\theta_3^{(t+1)} \sim P(\theta_3 \,|\, \theta_1^{(t+1)}, \theta_2^{(t+1)})$

Gibbs sampling (illustration)



Source: Bishop

Deriving a Gibbs in practice

- When computing conditional distribution for variable θ, we can ignore all terms do not depend on θ (Markov Blanket!)
- Using (semi-)conjugate priors make computing the conditional distributions easy

Gibbs sampling - example (1/7)

- Model: assume that we have $\mathbf{x} = (x_1, \ldots, x_n)$ observations from a mixture of two univariate Gaussians:

$$P(x_i \,|\, \theta, \pi) = (1-\pi)N(x_i \,|\, 0, 1) + \pi N(x_i \,|\, \theta, 1)$$

- Priors:

$$\pi \sim Beta(a_0, b_0)$$

$$\theta \sim N(0, \beta_0^{-1})$$

- Formulation using latent variables $\mathbf{z} = (z_1, \ldots, z_n)$:

$$\begin{aligned} P(\mathbf{z} \,|\, \pi) &= \prod_{i=1}^{n} \pi^{z_{i2}}(1-\pi)^{z_{i1}} \\ P(\mathbf{x} \,|\, \mathbf{z}, \theta) &= \prod_{i=1}^{n} N(x_i \,|\, 0, 1)^{z_{i1}} N(x_i \,|\, \theta, 1)^{z_{i2}} \end{aligned}$$

Gibbs sampling - example (2/7)

- We have a joint distribution

$$P(\mathbf{x}, \mathbf{z}, \theta, \pi) = P(\pi)P(\theta)P(\mathbf{z} \,|\, \pi)P(\mathbf{x} \,|\, \mathbf{z}, \theta)$$

- **Goal**: Posterior distribution

$$P(\theta, \pi \,|\, \mathbf{x})$$

- To derive a Gibbs sampler, we need to compute the conditional distributions for all unknowns

$$P(z_i \,|\, \mathbf{x}, \mathbf{z}_{-i}, \pi, \theta)$$

$$P(\theta \,|\, \mathbf{x}, \mathbf{z}, \pi)$$

$$P(\pi \,|\, \mathbf{x}, \mathbf{z}, \theta)$$

Here $\mathbf{z}_{-i}$ means all other variables in $\mathbf{z}$ except z_i

Gibbs sampling - example (3/7)

- We can write

$$\begin{aligned} P(z_i \mid \mathbf{x}, \mathbf{z}_{-i}, \pi, \theta) &\propto P(x_i \mid z_i, \theta) P(z_i \mid \pi) \\ &= N(x_i \mid 0, 1)^{z_{i1}} N(x_i \mid \theta, 1)^{z_{i2}} \pi^{z_{i2}} (1-\pi)^{z_{i1}} \end{aligned}$$

or, equivalently,

$$P(z_i \mid \mathbf{x}, \mathbf{z}_{-i}, \pi, \theta) \propto \begin{cases} (1-\pi) N(\mathbf{x}_i \mid 0, 1), & \text{if } z_{i1} = 1 \\ \pi N(\mathbf{x}_i \mid \theta, 1), & \text{if } z_{i2} = 1 \end{cases}$$

Gibbs sampling - example (4/7)

- After normalizing, we get

$$P(z_i \,|\, \mathbf{x}, \mathbf{z}_{-i}, \pi, \theta) \;=\; \begin{cases} r_{i1}, & \text{if } z_{i1} = 1 \\ 1 - r_{i1}, & \text{if } z_{i2} = 1 \end{cases}$$

where

$$r_{i1} = \frac{(1-\pi)N(\mathbf{x}_i \,|\, 0, 1)}{(1-\pi)N(\mathbf{x}_i \,|\, 0, 1) + \pi N(\mathbf{x}_i \,|\, \theta, 1)}$$

Gibbs sampling - example (5/7)

- Next, let us derive the conditional distribution for θ:

$$
\begin{aligned}
P(\theta \,|\, \mathbf{x}, \mathbf{z}, \pi) &\propto P(\theta) \prod_{i=1}^{n} P(x_i \,|\, z_i, \theta) \\
&\propto N(\theta \,|\, 0, \beta^{-1}) \prod_{i=1}^{n} N(x_i \,|\, \theta, 1)^{z_{i2}} \\
&\propto e^{-\frac{1}{2}\beta\theta^2} \cdot e^{-\frac{1}{2}\sum_{i=1}^{n} z_i (x_i - \theta)^2}
\end{aligned}
$$

- By completing the square, we get

$$
\theta \sim N\left(\frac{\sum_{i=1}^{n} x_i z_{i2}}{\beta + \sum_{i=1}^{n} z_{i2}}, \frac{1}{\beta + \sum_{i=1}^{n} z_{i2}} \right)
$$

Gibbs sampling - example (6/7)

- We can write the conditional probability as follows

$$
\begin{aligned}
P(\pi \,|\, \mathbf{x}, \mathbf{z}, \theta) &\propto P(\pi) \prod_{i=1}^{n} P(z_i \,|\, \pi) \\
&\propto \pi^{a_0 - 1} (1 - \pi)^{b_0 - 1} \prod_{i=1}^{n} \pi^{z_{i2}} (1 - \pi)^{z_i 1} \\
&= \pi^{a_0 - 1 + \sum_{i=1}^{n} z_{i2}} (1 - \pi)^{b_0 - 1 + \sum_{i=1}^{n} z_{i1}}
\end{aligned}
$$

- We notice that

$$
\pi \sim Beta(a_0 + \sum_{i=1}^{n} z_{i2}, b_0 + \sum_{i=1}^{n} z_{i1})
$$

Gibbs sampling - example (7/7)

- Gibbs sampler

1. Initialize $\mathbf{z}^{(1)}$, $\theta^{(1)}$, and $\pi^{(1)}$
2. For $t = 1, \ldots, T$:
   - For $i = 1, \ldots, n$:
     - Sample $z_{i1}^{(t+1)} \sim Bernoulli(r_{i1}^{(t)})$ where $r_{i1}^{(t)} = \frac{(1-\pi^{(t)})N(\mathbf{x}_i|0,1)}{(1-\pi^{(t)})N(\mathbf{x}_i|0,1)+\pi^{(t)}N(\mathbf{x}_i|\theta^{(t)},1)}$
   - Sample $\theta^{(t+1)} \sim N(\frac{\sum_{i=1}^{n} x_i z_{i2}^{(t+1)}}{\beta+\sum_{i=1}^{n} z_{i2}^{(t+1)}}, \frac{1}{\beta+\sum_{i=1}^{n} z_{i2}^{(t+1)}})$
   - Sample $\pi^{(t+1)} \sim Beta(a_0 + \sum_{i=1}^{n} z_{i2}^{(t+1)}, b_0 + \sum_{i=1}^{n} z_{i1}^{(t+1)})$

Metropolis algorithm

- To derive a Gibbs sampler for $P(\theta_1, \theta_2, \theta_3)$ one need to be able to sample from the full conditionals $P(\theta_1 \mid \theta_2^{(t)}, \theta_3^{(t)})$
- In the Metropolis algorithm, it is only required that the ratio

$$\frac{P(\theta^*)}{P(\theta^{(t)})}$$

can be evaluated where $\theta^{(t)} = (\theta_1^{(t)}, \theta_2^{(t)}, \theta_3^{(t)})$ is the current value and $\theta^* = (\theta_1^*, \theta_2^*, \theta_3^*)$ is a value drawn from some proposal distribution

$$q(\theta^* \mid \theta^{(t)})$$

Proposal distribution

- In the Metropolis algorithm, it is assumed that the proposal distribution is symmetric

$$q(\theta^* \,|\, \theta^{(t)}) = q(\theta^{(t)} \,|\, \theta^*)$$

- Furthermore, the chain should be ergodic (irreducibility, aperiodicity)
- Otherwise, you are free to choose the proposal distribution
  - Though, the choice will affect the quality of the results

Metropolis algorithm

Metropolis algorithm to sample from $P(\theta \,|\, D)$

1. Select initial value $\theta^{(1)}$
2. For $t = 2, \ldots, T$:
   - Draw a candidate θ^* from the proposal $q(\theta^* \,|\, \theta^{(t-1)})$
   - Compute acceptance probability

$$
\begin{aligned}
A(\theta^* \,|\, \theta^{(t-1)}) &= \min\left\{1, \frac{P(\theta^* \,|\, D)}{P(\theta^{(t-1)} \,|\, D)}\right\} \\
&= \min\left\{1, \frac{P(D \,|\, \theta^*)P(\theta^*)}{P(D \,|\, \theta^{(t-1)})P(\theta^{(t-1)})}\right\}
\end{aligned}
$$

   - Draw $u \sim Unif(0, 1)$
     - If $u < A$, set $\theta^{(t)} = \theta^*$ (proposal accepted)
     - Otherwise, set $\theta^{(t)} = \theta^{(t-1)}$ (proposal rejected)

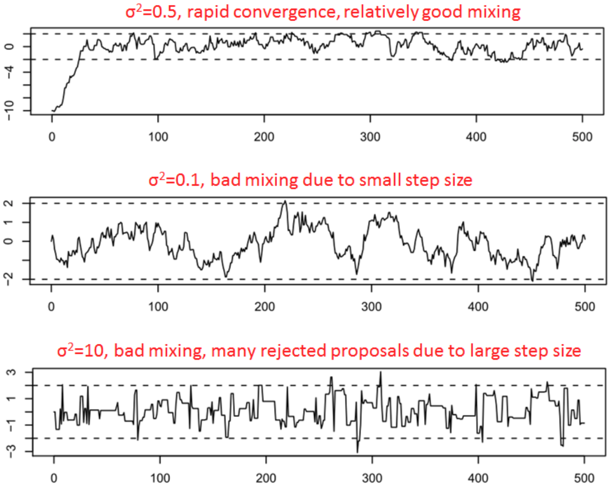
Metropolis sampling - Example (1/2)

- Simulate from $N(0, 1)$
- Use a proposal distribution

$$q(x^* \,|\, x^{(t)}) = N(x^* \,|\, x^{(t)}, \sigma^2)$$

- Different values of σ^2 lead to different mixing properties

Metropolis sampling - Example (2/2)



Metropolis-Hastings algorithm

- Very similar to the Metropolis algorithm
- The difference is that the proposal distribution $q(\theta^* \mid \theta^{(t)})$ does not need to be symmetric, that is, $q(\theta^* \mid \theta^{(t)})$ can differ from $q(\theta^{(t)} \mid \theta^*)$
- The algorithm is otherwise exactly the same as in Metropolis algorithm except the acceptance probability needs to be adjusted due to the asymmetry

$$A(\theta^* \mid \theta^{(t-1)}) = \min\left\{1, \frac{P(\theta^* \mid D)}{P(\theta^{(t-1)} \mid D)} \frac{q(\theta^{(t-1)} \mid \theta^*)}{q(\theta^* \mid \theta^{(t-1)})}\right\}$$

Representing distributions with samples

- Once you have samples from posterior, you can approximate any property of the posterior
  - Posterior mean $\approx$ Sample mean
  - Posterior variance $\approx$ Sample variance
  - ...
- The more (independent) samples you have, the more accurate the approximation is
  - If your chain does not mix well, you need more samples

Predictive distributions

- You can sample from the posterior predictive distribution using the following procedure:
  - For $t = 1, \ldots T$:
    1. Sample $\theta^{(t)} \sim P(\theta \,|\, D)$
    2. Sample $x^{(t)} \sim P(x \,|\, \theta^{(t)})$
- Now $x^{(t)}$, $t = 1, \ldots, T$ are a sample from the posterior predictive distribution

Remarks

- Gibbs vs. Metropolis-Hastings
  - You can use Metropolis-Hastings whenever you can compute $P(\mathbf{x}|\theta)P(\theta)$
  - In Gibbs, you need to be able to compute the full conditional distributions which may be difficult if there are non-conjugate priors involved
  - Rule of a thumb (based on my personal experience): Use Gibbs if you can derive the updates, otherwise MH
- Challences with MCMC
  - Storing the sample requires lots of memory (especially for complex models)
  - The stationary distribution is achieved only asymptotically
    - With finite sample size, it is difficult to estimate whether the chain has converged

Further readings

- Bishop 11

Sources

- These slides are mostly based on the slides from the course "Machine Learning: Advanced Probabilistic Methods" by P. Marttinen (Aalto University)