Sampling[1]

Dmitry Adamskiy, David Barber

University College London

[1] These slides accompany the book *Bayesian Reasoning and Machine Learning*. The book and demos can be downloaded from www.cs.ucl.ac.uk/staff/D.Barber/brml. Feedback and corrections are also available on the site. Feel free to adapt these slides for your own purposes, but please include a link the above website.

Overview

Problem Setting

Importance Sampling

Rejection Sampling

MCMC methods: Gibbs and Metropolis-Hastings

Problem Setting

Problem Setting

- Sometimes we have to deal with (reasonably) complex probability distributions: for example, we may know them up to normalising constant.
  - EM-like inference: posterior $p(h|v, \theta)$ could be nasty, but joint $p(h, v|\theta)$ is fine.
  - In undirected models, $p = \frac{1}{Z}e^{-E(x)}$, we know energy $E(x)$, but normalization constant is a problem.
  - ...
- Questions we might be interested in:
  1. Draw samples from $p(x)$.
  2. Compute expectations with respect to $p(x)$.
- Two possible approaches: deterministic approximations (last week) and sampling (today).

Basic Monte Carlo

Using samples to approximate averages

Given a finite set of samples $\mathcal{X}$, one can approximate expectation as

$$\langle f(x) \rangle_{p(x)} \approx \frac{1}{L} \sum_{l=1}^{L} f(x^l) \equiv \hat{f}_{\mathcal{X}}$$

The subscript indicated that the estimate depends on the set of samples.

- If we have a procedure that 'faithfully' draws samples from $p(x)$, then we can use this to approximate averages.

Sampling as distribition

A sampling procedure produces realisations of set $\mathcal{X}$ and can be considered a distribution $\tilde{p}(\mathcal{X})$. Provided that the marginals are equal to the marginals of target distribution, $\tilde{p}(x^l) = p(x^l)$, the average of the sampling approximation is

$$\left\langle \hat{f}_{\mathcal{X}} \right\rangle_{\tilde{p}(\mathcal{X})} = \frac{1}{L} \sum_{l=1}^{L} \left\langle f(x^l) \right\rangle_{\tilde{p}(x^l)} = \left\langle f(x) \right\rangle_{p(x)}$$

So the mean of the sample approximation is the exact mean of f provided only thet the marginals of $\tilde{p}(\mathcal{X})$ correspond to $p(x)$.

Dependent samples?

Even if the individual samples are dependent, that is $\tilde{p}(\mathcal{X})$ does not factorise into $\prod_l \tilde{p}(x^l)$, the sample average is unbiased.

Variance

What about the variance? Defining $\Delta \hat{f}_{\mathcal{X}} = \hat{f}_{\mathcal{X}} - \left\langle \hat{f}_{\mathcal{X}} \right\rangle_{\tilde{p}(\mathcal{X})}$ and $\Delta f(x) = \hat{f}_{\mathcal{X}} - \langle f(x) \rangle_{p(x)}$ the variance of the approximation becomes (assuming $\tilde{p}(x^l) = p(x)$)

$$\begin{aligned}
\left\langle \Delta^2 \hat{f}_{\mathcal{X}} \right\rangle_{\tilde{p}(\mathcal{X})} &= \frac{1}{L^2} \sum_{l,l'} \left\langle \Delta f(x^l) \Delta f(x^{l'}) \right\rangle_{\tilde{p}(x^l, x^{l'})} \\
&= \frac{1}{L^2} \left(L \left\langle \Delta^2 f(x) \right\rangle_{\tilde{p}(x)} + \sum_{l \neq l'} \left\langle \Delta f(x^l) \Delta f(x^{l'}) \right\rangle_{\tilde{p}(x^l, x^{l'})} \right)
\end{aligned}$$

If the samples are independent, $\tilde{p}(x^l, x^{l'}) = \tilde{p}(x^l)\tilde{p}(x^{l'})$, the last term vanishes and the variance scales inversely with the number of samples.

Drawing independent samples

- The critical difficulty is in actually generating independent samples from $p(x)$.
- A dependent scheme may be unbiased, but if variance is very high we need a large number of samples to get an accurate approximation.

Importance Sampling

Importance Sampling

Consider $p(x) = \frac{p^*(x)}{Z}$, where $p^*(x)$ can be evaluated but $Z = \int_x p^*(x)$ is intractable. The average with respect to p is given by

$$\int_x f(x)p(x) = \frac{\int_x f(x)p^*(x)}{\int_x p^*(x)} = \frac{\int_x f(x)\frac{p^*(x)}{q(x)}q(x)}{\int_x \frac{p^*(x)}{q(x)}q(x)}$$

and we can approximate this sampling from $q(x)$.

$$\int_x f(x)p(x) \approx \frac{\sum_{l=1}^L f(x^l)\frac{p^*(x^l)}{q(x^l)}}{\sum_{l=1}^L \frac{p^*(x^l)}{q(x^l)}} = \sum_{l=1}^L f(x^l)w^l,$$

where we define the normalised importance weights

$$w^l = \frac{p^*(x^l)/q(x^l)}{\sum_{l=1}^L p^*(x^l)/q(x^l)}$$

Problems

- Finding the right distribution Q is not easy
- Diagnosing whether Q is good is also not easy

Let's see how this works...

Rejection Sampling

Rejection Sampling

- How to draw samples from $p(x)$ when we have an efficient sampling procedure for $q(x)$?
- Suppose that we know that $p(x) < cq(x)$ for all x. Then we can sample from $q(x)$ and accept the sample with probability $\frac{p(x)}{cq(x)}$.

Auxiliary variable view

- Let $y \in \{0, 1\}$ be an auxiliary binary variable and define $q(x, y) = q(x)q(y|x)$. If we set $q(y = 1|x) \propto p(x)/q(x)$, then $q(x, y = 1) \propto p(x)$
- So sampling from $q(x, y)$ gives us a procedure for sampling from $p(x)$
- Expected acceptance rate is $1/c$.
- Works with unnormalised $p^*(x)$, if we find c such that $p^*(x) < cq(x)$. The acceptance rate becomes Z/c.

It is not easy to find distribution $q(x)$ such that c is small.

MCMC methods: Gibbs and Metropolis-Hastings

MCMC idea

- Suppose we want to sample from distribution $p(x)$. The idea is to build a Markov chain such that $p(x)$ is the stationary distribution $p_\infty(x)$ of the chain:

$$p_\infty(x') = \sum_x p_\infty(x) T(x \to x')$$

- Then in the long run samples from the chain will be samples from $p(x)$.
- We draw the first sample from some distribution $p_0(x)$ and then use proposal distribution $T(x \to x') = p(x_t = x' | x_{t-1} = x)$.
- How to select the proposal distribution to arrive where we want to?
- There are some sufficient conditions guaranteeing convergence...

Ergodicity and Detailed Balance

- First, we need to converge to unique stationary distribution regardless of the initial state x_0: a form or ergodicity. A sufficient condition for that is that there is some k that it is possible to reach any stat from any other state in exactly k steps:

  $$T^k(x \to x') > 0$$

  for all x, x' and some k.
- Exercise: think of the non-trivial example where this breaks.
- One sufficient condition for $p(x)$ being stationary distribution is this

  $$p(x')T(x' \to x) = p(x)T(x \to x')$$

  This is called detailed balance.

Gibbs Sampling

Sometimes (as you will see in Assignment 3), the whole distribution $p(x)$ is nasty, but the conditionals $p(x_i|x_{\setminus i})$ are easy to sample from.

- Gibbs sampling: loop over the variables (at random or in turn) and sample from the conditional $p(x_i|x_{\setminus i})$.
- Detailed balance. The transition probability is

  $$T(x \to x') = \pi_i p(x'_i|x_{\setminus i}),$$

  where π_i is a probability of picking i-th variable.
- Then

  $$T(x \to x')p(x) = \pi_i p(x'_i|x_{\setminus i})p(x_i|x_{\setminus i})p(x_{\setminus i})$$

  On the other hand:

  $$T(x' \to x)p(x) = \pi_i p(x_i|x'_{\setminus i})p(x'_i|x'_{\setminus i})p(x'_{\setminus i})$$

  But $x'_{\setminus i} = x_{\setminus i}$ so these are the same.

Metropolis-Hastings algorithm

Problems with Gibbs sampling:

- Could be slow
- What is we can't sample from the conditionals?

Idea of Metropolis algorithm: let's use some proposal distribution $Q(x \to x')$ and we'll accept of reject based on density of $p(x)$ (a bit like rejection sampling).
Repeat the following steps starting from some x:

1. Propose a new state x' by sampling from $Q(x'|x) = Q(x \to x')$.
2. Accept the new state with probability

$$\min(1, p(x')Q(x' \to x)/p(x)Q(x \to x'))$$

3. If the sample is not accepted the old state becomes the new sample.

Detailed balance

- The probability $T(x \to x')$ is

$$T(x \to x') = S(x \to x') \min\left(1, \frac{p(x')Q(x' \to x)}{p(x)Q(x \to x')}\right)$$

- Let $p(x')Q(x' \to x) \leq p(x)Q(x \to x')$. Then

$$T(x \to x')p(x) = p(x)Q(x \to x')\frac{p(x')Q(x' \to x)}{p(x)Q(x \to x')}$$

  and

$$p(x')T(x' \to x) = p(x')Q(x' \to x).$$

  So they are equal and detailed balance holds.

Issues

- How do we know when we converged?
- How to select parameters to get a good proposal distribution?
- Burn-in: the target distribution is reached in the limit, so discarding initial samples is a good policy.
- Samples are correlated, so should we skip some?
- One long chain vs. lots of short?
- Lots of algorithms and practical tricks