Approximate Inference - Key Ideas

Probabilistic Artificial Intelligence - HS 2018

Mojmír Mutný
ETH Zürich

November 21, 2018

1 Bayesian Inference

To expose ideas of approximate inference, let us work with the following definition of inference.

Definition 1 (Inference)**.** *Inference is a quantitative statistical query about a quantity Q with the knowledge of some data D and model $\mathcal{M}$ than relates Q and D.*

An example of a query can be the most likely value of a quantity Q given the data D where model $\mathcal{M}$ dictates the conditional probability relation (often $P(D|Q)$). In this case, a query can be for example as,

$$q_{MAP} = \arg\max_q P(Q = q|D) \tag{1}$$

The subscript MAP stands for *maximum a-posterior*, since often, when interpreting Q as random variable and having a prior probability on the quantity Q, $P(Q)$, we can derive the posterior (given the realization of another random variable D) $P(Q|D)$. Utilizing the knowledge of the model that gives us $P(D|Q)$, we can calculate the desired quantity using Bayes' rule,

$$P(Q|D) = \frac{P(D|Q)P(Q)}{P(D)} = \frac{R(Q)}{Z} \tag{2}$$

where

$$Z = P(D) = \sum_{q\in\Omega} P(D|Q = q)P(Q = q)dq, \tag{3}$$

and $R(Q) = P(D|Q)P(Q)$.

As the model $\mathcal{M}$ dictates the conditional dependence $P(D|Q)$ calculation of $R(Q)$ is often not challenging, however, integrating Z can be. Calculating this is not necessary for the query in (1), however different queries such as the following require the calculation of Z.

$$\mathbb{E}[Q|D] = \sum_{q\in\Omega} qP(Q=q|D)dq = \frac{1}{Z}\sum_{q\in\Omega} qP(D|Q=q)P(Q=q)dq \tag{4}$$

Note that this way of inference can be contrasted with the so called *frequentist inference*, where, for example, a parameter that maximizes the likelihood of the event D is chosen as an estimator,

$$\theta_{ML} = \arg\max P(D|Q).$$

1.1 Remark

When $|\Omega| < \infty$, then a probability distribution can be represented as a histogram, or in other words, as a vector.

Large portions of these notes are based on the excellent review of [Andrieu et al., 2003].

2 Monte Carlo (MC)

Monte Carlo is a colloquial name for an approximation of integrals using sampling from probability distribution. It has its roots in Manhattan project [Andrieu et al., 2003]. It can be used to answer queries where an expectation is required such as the above query $\mathbb{E}[Q|D]$ or calculating the whole probability distribution $P(Q|D)$ (not just $\arg\max$).

First, consider definition of expectation,

$$\mathbb{E}_{q\sim p(q)}[f(q)] := \sum_{q\in\Omega} p(q)f(q)dq. \tag{5}$$

Due to *Law of Large Numbers*, we know that expectation can be approximated using sample mean μ_n,

$$\mu_n = \sum_{i=1}^{n} f(q_i) \text{ where } q_i \sim p(q). \tag{6}$$

In other words, we know that

$$\mu_n \to \mathbb{E}_{q\sim p(q)}[f(q)] \text{ as } n \to \infty$$

Hence when faced with an inference problem where Z needs to be approximated we resort to approximation of Z as,

$$Z = \sum_{q \in \Omega} P(D|Q = q)P(Q = q) = \mathbb{E}_{q \sim P(Q)}[P(D|Q = q)] \approx \sum_{i=1}^{n} P(D|Q = q_i),$$

where $q_i \sim P(Q)$. The estimate $\mu_n = \sum_i^n P(D|Q = q_i)$ might converge very slowly to $\mathbb{E}_{q \sim P(Q)}[P(D|Q)]$, since in some cases, as demonstrated in Figure 1, most of the probability mass where the samples from $P(Q)$ lie is elsewhere to where most of the mass of the integrand $P(D|Q)$ lies.

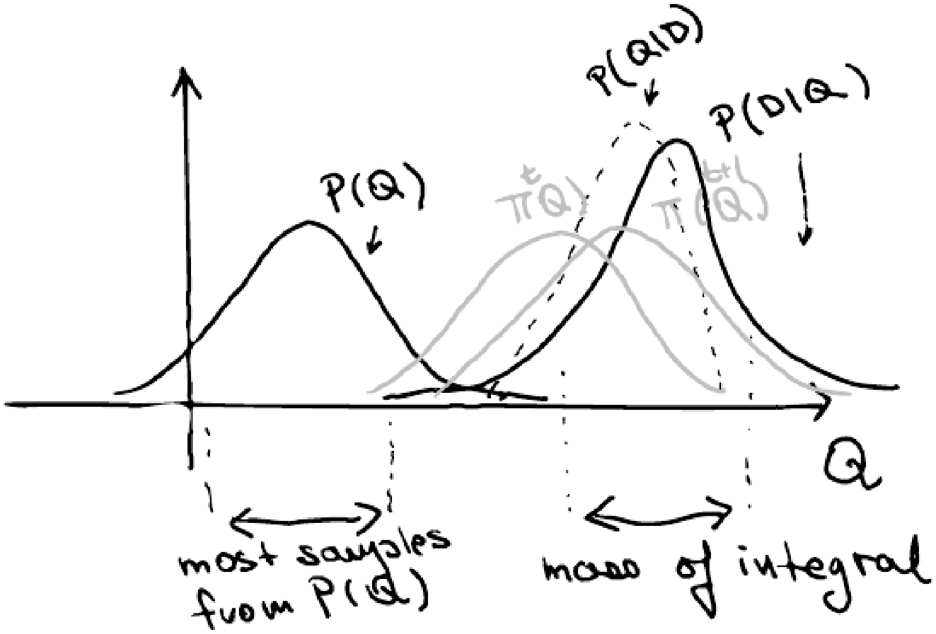

Figure 1: Importance sampling with $\pi(Q)$. In Markov chain Monte Carlo we want to improve our importance sampling distribution over the time in order to sample from the true posterior.

3 Importance Sampling[1]

Importance sampling is a step towards MCMC that we delineate here. It tackles again the problem to approximate (3).

[1] Not examinable

Importance sampling is a way to steer the sampling process of MC with a user defined sampling distribution $\pi(Q)$ by considering the following equivalent statements

$$
\begin{aligned}
Z &= \sum_{q\in\Omega} P(D|Q=q)P(Q=q)dq & (7)\\
&= \sum_{q\in\Omega} \frac{P(D|Q=q)P(Q=q)}{\pi(Q=q)}\pi(Q=q)dq & (8)\\
&= \sum_{q\in\Omega} w(Q=q)\pi(Q=q)dq & (9)\\
&= \mathbb{E}_{q\sim\pi(Q)}[w(Q)], & (10)
\end{aligned}
$$

where $w(Q) = \frac{P(D|Q)P(Q)}{\pi(Q)}$. As the above can be expressed as expectation, we can use Monte Carlo estimate again, but now with sampling according to a different distribution $\pi(Q)$.

A natural questions arises, what is the optimal sampling distribution $\pi(Q)$ such that our estimate converges quickly to the true value. The answer, not surprisingly but unfortunately, turns out to be the actual posterior distribution $P(Q|D)$ that we try to estimate along this procedure.

A way to bypass this obstacle is to use MCMC, where we successively try to improve the quality of our distribution $\pi(Q)$ such that it approaches the optimal sampling distribution $P(Q|D)$.

Remark 1. *A probability distribution over a finite space Ω can be represented a vector or length $|\Omega|$, namely $\pi \in \mathbb{R}^{|\Omega|}$, where sum of its entries is equal to 1. The probability of an event $(i \in \Omega)$can be associated to a component $P(Q=i) = \pi_i$.*

4 Markov Chain Monte Carlo (MCMC)

The idea of MCMC is to start with a importance sampling distribution π^0, and create an iterative procedure, whereby we improve our importance sampling distribution $\pi^t(Q)$ in each iteration $t \to t+1$.

With MCMC, we will be able to generate samples from a distribution π^t which will eventually approach the distribution $\pi^t \to P(Q|D)$ as $t \to \infty$. In order to define the transition from t to $t+1$, we need a language and techniques from Markov chains.

4.1 Markov Chains

Before we dwell into intricacies of Markov chain Monte Carlo, we need to first review basics of Markov chains. For a comprehensive review, please have a look in the standard probability textbooks (e.g. [Ross, 2009]). We cannot go into all technicalities here, however you are recommended to review concepts of preriodicity, reducibility, ergodicity and ergodic theorem.

Definition 2 (Markov stochastic process)**.** *A stochastic process is a collection of random variables $\{Q_t\}_{t\in I}$, where I is the ordered-index set.*

A Markov stochastic process satisfies Markov property if

$$\mathrm{P}(Q_{t'}|\{Q_t\}_{t<t'}) = \mathrm{P}(Q'_t|Q_{t'-1}). \tag{11}$$

An associated object with a Markov stochastic process is so called transition kernel, or in our case transition matrix. We denote it T. It tells us the probability that having observed $Q_t = i$, what is the probability that we observe a specific value of $Q_{t+1} = j$ in the next step of the chain. Specifically,

$$T(i,j) = P(Q_{t+1} = i|Q_t = j) \tag{12}$$

In finite probability space ($|\Omega| < \infty$) we can view this as matrix $T \in \mathbb{R}^{|\Omega|\times|\Omega|}$.

Lastly, we need to introduce a concept of a stationary distribution π^∞.

Definition 3. *Let $\{Q_t\}$ be a Markov process with T transition kernel. A stationary distribution π^∞ is a distribution which fulfills the following relation,*

$$\sum_{q\in\Omega} \pi^\infty(Q_t = q)T(Q_{t+1}, Q_t = q)dq = \pi^\infty(Q_{t+1}). \tag{13}$$

This condition can be interpreted in matrix notation as

$$\pi^\infty T = \pi^\infty, \tag{14}$$

which signalizes that π^∞ represented as a vector is left eigenvector of the matrix T with associated eigenvalue 1. If your Markov chain is irreducible and aperiodic the existence and uniqueness of such eigenvector is guaranteed. If it is not aperiodic, the uniqueness needs to be dropped.

A property of Markov chains that we will not review here is that given suitable conditions they converge to stationary distributions. Please refer to [Ross, 2009]. Namely with matrix notation, this can be interpreted as

$$\vartheta T^t \to \pi^\infty$$

as $t \to \infty$, where $\vartheta \in \mathbb{R}^{|\Omega|}$ is a probability distribution.

4.2 Markov chain of probability distributions

The insight of MCMC is to stipulate that we know the limiting stationary distribution which is $\pi^\infty = P(Q|D)$, and to find a transition kernel such that this is true. Then, we evolve the kernel and use it to *approximately* sample from $P(Q|D)$. In other words, the probability distributions $\pi^{t+1} = T\pi^t$ follow the Markov chain specified by the kernel T.

The difficulty here lies in ensuring that transition kernel T (or proposal distribution) satisfies the eigenvector relation in Equation (14). It turns out there is a stronger condition that can be imposed that is easier to handle.

4.3 Detailed Balance

We have seen that the requirement on the transition kernel $T(Q_{t+1}, Q_t)$ is that it has a stationary distribution equal to the true posterior $P(Q|D)$. This condition is in some circumstances difficult assure directly. Instead, we can formulate a condition that necessary implies that the stationary distribution is $P(Q|D)$.

Definition 4 (Detailed Balance)**.** *Let $\{Q_t\}_t$ be a Markov process with a transition kernel T, and π^∞ distribution, then if the following relation holds*

$$\pi^\infty(Q_t)T(Q_{t+1}, Q_t) = \pi^\infty(Q_{t+1})T(Q_t, Q_{t+1}) \tag{15}$$

we say that the distribution π^∞ under the transition kernel T satisfies the detailed balance condition.

We see that we can make sure that $P(Q|D)$ is stationary distribution for the Markov chain with transition kernel T without knowing $P(Q|D)$ exactly. The only required quantity is $R(Q) = P(D|Q)P(Q)$. Since,

$$\frac{R(Q_t)}{Z}T(Q_{t+1}, Q_t) = \frac{R(Q_{t+1})}{Z}T(Q_t, Q_{t+1}).$$

Claim 1. *Detailed balance of π^∞ and T implies that π^∞ is stationary distribution for the kernel T.*

Proof. We sum both sides of the equation (15) with respect to q.

$$\sum_{q\in\Omega}\pi^{\infty}(Q_t=q)T(Q_{t+1},Q_t=q)dq=\sum_{q\in\Omega}\pi^{\infty}(Q_{t+1})T(Q_t=q,Q_{t+1})dq$$

As $T(Q_{t+1},Q_t)$ is valid probability distribution, it sums to one.

$$\sum_{q\in\Omega}\pi^{\infty}(Q_t=q)T(Q_{t+1},Q_t=q)dq=\pi^{\infty}(Q_{t+1}).$$

□

4.4 Algorithm

We have nearly all ingredients to formulate a practical MCMC algorithm. First consider the toy example, where we know that transition kernel T satisfies detailed balance (or the stationarity condition) with the true posterior. Then the toy Algorithm 1 approximates a sample from the true posterior $P(Q|D)$. This algorithm is a way to produce a sample $q^{\tau}\sim\pi(Q)^{\tau}$.

Algorithm 1 Toy MCMC

```
Pick $q^{(0)} \sim P(Q)$
**repeat** $t = 1, 2 \dots, \tau$
    $q^{(t)} \sim T(Q, Q_{t-1} = q^{(t-1)})$
**until** happy
**return** $q^{(\tau)}$
```

However, often in practice, we cannot guarantee that T satisfies the detailed balance. Instead, we focus on the scenario, where we guarantee that T and the initial distribution and $P(Q|D)$ have the same support[2], and then devise a universal procedure such that for any such T the detailed balance holds. This leads to the famed Metropolis-Hastings Algorithm 2.

Let us now convince ourselves that the scheme in Algorithm 2 satisfies the detailed balance. The procedure defines a new transition matrix $\bar{T}$. Indeed, the new transition kernel with the help of notation $A(Q_{t+1},Q_t)=\frac{T(Q_t,Q_{t+1})R(Q_{t+1})}{T(Q_{t+1},Q_t)R(Q_t)}$ can be written as

$$\bar{T}(Q_{t+1},Q_t)=T(Q_{t+1},Q_t)A(Q_t,Q_{t+1})+\delta(Q_{t+1}-Q_t)\sum_{x\in\Omega}T(Q_{t+1}=x,Q_t)\left(1-A(Q_t,Q_{t+1}=x)\right)dx.$$

The first terms is associated with the acceptance step and the second is associated with the rejection step ($\delta(x)$ is a delta function where all of its mass is concentrated at 0 only). By construction this satisfies the detailed balance.

[2] Often it is the whole domain

Algorithm 2 Metropolis-Hastings algorithm

```
Require: $R(Q) = R(D|Q)P(Q)$ proportional to the posterior,$T$ transition kernel
Ensure:
  Pick $q^{(0)} \sim P(Q)$
  repeat $t = 1, 2 \ldots, \tau$
    Sample $u \sim$ Uniform$[0, 1]$
    Sample $c^{(t)} \sim T(Q, Q_{t-1} = q^{(t-1)})$
    if $u < \frac{T(q^{(t-1)}, c^{(t)}) R(c^{(t)})}{T(c^{(t)}, q^{(t-1)}) R(q^{(t-1)})}$ then
      $q^{(t)} = c^{(t)}$
    else
      $q^{(t)} = q^{(t-1)}$
    end if
  until happy
  return $q^{(\tau)}$
```

References

[Andrieu et al., 2003] Andrieu, C., De Freitas, N., Doucet, A., and Jordan, M. I. (2003). An introduction to mcmc for machine learning. *Machine learning*, 50(1-2):5–43.

[Ross, 2009] Ross, S. (2009). *A First Course in Probability 8th Edition.* Pearson.