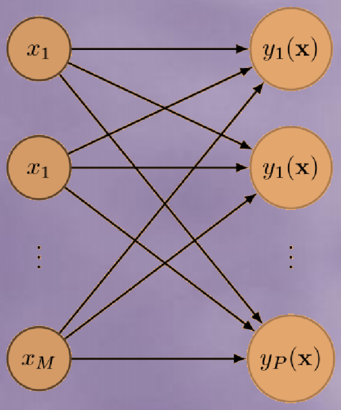


Sampling based learning

Daniel Khashabi [1]
KHASHAB2@ILLINOIS.EDU

0.1 Introduction

In all of the learning problems, after the parametric modelling we need to devise a way to learn the optimal parameters, using some training samplings, or some indirect rules, with respect to some criterion, e.g. a defined loss-function. Usually this can be cast as maximizing (or minimizing, with an additional negative sign) a function of parameters, training data, and the prior knowledge, commonly known as MAP or *maximum a posteriori*.

$$\mathcal{L} = \log p(\mathcal{D}|\Theta) \rightarrow \Theta^* = \max_{\Theta} \log p(\mathcal{D}|\Theta)$$

Since the posterior distribution (or function, if not normalized) is usually a complicated function, it is not straightforward to maximize it directly with respect to model parameters. One approach can be approximating this function and finding the sub-optimal parameters. The other approach which is mostly studied here, is statistical sampling methods, which take many samples of the model, to simulate the behaviour of the model. These methods are usually slow, and exact asymptotically (if they run long enough).

Before starting on learning based on sampling, we should first learn how to sample complicated distributions. Usually it could be assumed that we know how to sample a uniform distribution, and we aim at generalizing it to sampling other complicated distributions.

0.2 Sampling a proper distribution

In theory, there is an easy way to sample any distribution, by finding an invertible parametric form which converts the variables in two distributions. Let's say we know how to sample $p(x)$,

[1] This is part of my notes; to find the complete list of notes visit `http://web.engr.illinois.edu/~khashab2/learn.html`. This document is updated on August 10, 2013.

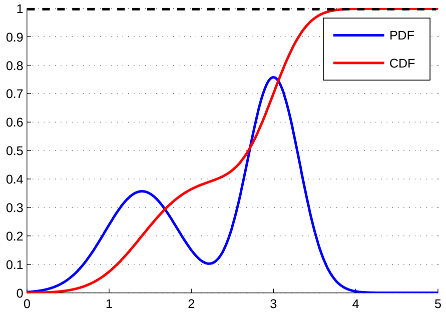


Figure 1: A sample distribution, and its cumulative distribution.

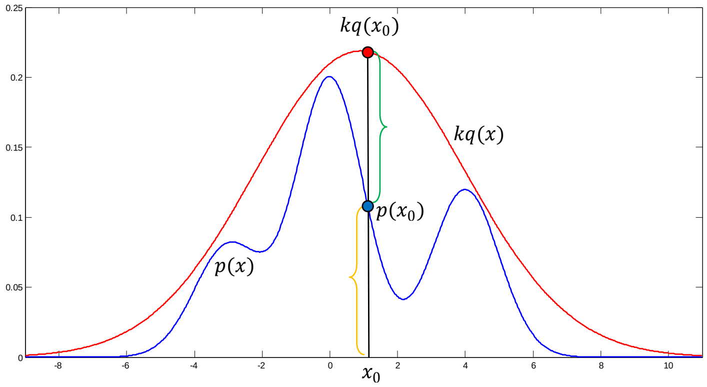


Figure 2: A sample distribution, and its cumulative distribution.

our goal is to get the distribution $p(z)$ by finding a parametric form for $x \rightarrow z$. We get the distribution $p(z)$ by the following conversion between two distributions,

$$p(x) = p(z)\left|\frac{dz}{dx}\right|.$$

In Figure 1 a probability distribution $p(x)$, and its cumulative distribution $P(X \leq x) = \int_{-\infty}^{x} p(x^{'})dx^{'}$ is depicted. If we sample the y-axis uniformly, find the corresponding points in the CDF curve, and map them on the the x-axis, the corresponding points are distributed according to $p(x)$. In mathematical form, this can be explained in the following form,

$$p(x) = p(z)\left|\frac{dz}{dx}\right| , p(z) = 1 \Rightarrow z = h(x) = P(X \leq x) = \int_{-\infty}^{x} p(x^{'})dx^{'} \Rightarrow x = h^{-1}(z).$$

0.2.1 Rejection sampling

Since in many modelling problems, it is not easy to find an analytical for the CDF, we prefer to find a way for sampling an arbitrary distribution, without the need for finding its CDF. One of these methods is called *rejection sampling*.

Let's say we want to sample a distribution $p(x)$ which has a complicated form, and we can't find its CDF. IJn rejection sampling, we find another distribution $q(x)$ which supports the target distribution, i.e. $p(x)$. In other words, for any $x^{'}$ in the domain of the distributions,

$q(x^{'}) > p(x^{'})$. Note that, in general $q(x)$ doesn't have to be a proper distribution, in the sense that the area under it sum up to one, but it needs to be of the forms which is easy to sample from. Then $kq(x),\ k \in \mathbb{R}_{++}$ which is easy to sample from, and supports $p(x)$ can be used. In Figure 2 a complicated target distribution$p(x)$, and a supporting distribution $kq(x)$ are shown.

The procedure for rejection sampling is as following: first we sample a point x_0 from the distribution $kq(x)$. Because $k > 0$, we know that $kq(x) > 0$. We create a uniform distributions on $[0, kq(x)]$, and sample a point from that. If the point is greater that $p(x_0)$ we accept it as a sample of $p(x)$, if not, we reject it. It can be shown that in long-run the accepted samples will have distribution according to $p(x)$ (proof?).
To decrease the ratio of the rejected samples it is necessary to choose the supporting distribution $kq(x)$ as close as possible to $p(x)$, though it might need might be hard to find such a distribution when handling high-dimensional distributions. Also it can be shown, roughly speaking, the probability of a sample being accepted diminishes exponentially with the number of the dimensions. This makes rejection sampling very hard to use in high-dimensional problems, and with a very complicated form, which are hard to visualize. There are a few works which aim at finding better supporting distribution adaptively by using peace-wise exponential functions, or log-concave families (see [Gilks and Wild(1992), Gilks et al.(1995)Gilks, Best, and Tan])

Usually in probabilistic inference problems, we are dealing with a real ratio of the target distribution $p(x)$. In other words, if we assume that $p(x) = \frac{1}{\mathcal{Z}}\tilde{p}(x)$, where $\mathcal{Z} = \int p(x)dx$ is a normalizing constant, we usually only have $\tilde{p}(x)$, and it is hard to normalize. Thus, there is a big motivation for finding methods which can use the unnormalized function $\tilde{p}(x)$, and give samples of $p(x)$ without directly having it.

0.2.2 Sampling for approximating integrals

Let's say we want to approximate the following integration, $\displaystyle\int f(z)p(z)dz$ which is equivalent to the following expectation, $\mathbb{E}_p[f]$. We can use sample mean as an estimator of the statistical mean, and we can approximate the above expectation by sampling from $p(x)$,

$$\mathbb{E}_p[f] \approx \frac{1}{L}\sum_{i=1}^{L} f(z^i), \quad x^i \sim p(x).$$

Note that in general this trick could be used for approximating any integration with a proper choice of $p(x)$. Also it can easily verified that sample mean is an unbiased estimator the statistical mean.

Let's say we don't know how to sample $p(x)$ and we want to approximate $\mathbb{E}_p[f]$. We choose a distribution $q(x)$ which we know how to sample from, and change the expectation using it,

$$\mathbb{E}_p[f] = \int f(z)p(z)dz = \int f(z)\frac{p(z)}{q(z)}q(z)dz \approx \frac{1}{L}\sum_{i=1}^{L} f(z^i)\frac{p(z^i)}{q(z^i)}, \quad z^i \sim q(z)$$

This let's us sample from $q(x)$ for approximating the value of the sample expectation. This trick is usually called *importance sampling*. Practical usage of this trick demands careful considerations. One important point is that, to get realistic answers, $q(x)$ must be non-zero (or not very small) wherever $p(x)f(x)$ is not zero. This trick has many interesting applications; for example one can use use this trick to calculate expectation of events happening when their probability is very small, e.g. calculating “bit error rate" in a communication system [Jeruchim(1984)].

Let’s consider the case where we don’t have the distribution $p(x)$ but we only have a positive ratio of that. In other words, if $p(z) = \frac{1}{Z}\tilde{p}(x)$, we only have $\frac{1}{Z}\tilde{p}(x)$ and calculation of the normalizing constant is too costly that we don’t want to do it. We can simplify the previous formulations as following,

$$\mathbb{E}_p\left[f\right] = \int f(z)p(z)dz = \frac{1}{Z_p}\int f(z)\tilde{p}(z)dz = \frac{1}{Z_p}\int f(z)\frac{\tilde{p}(z)}{q(z)}q(z)dz = \frac{1}{Z_p}\sum_{i=1}^{L} f(z^i)\frac{\tilde{p}(z^i)}{q(z^i)}, \quad z^i \sim q(z)$$

A similar thing can be done to find an estimation of the normalizing constant,

$$1 = \mathbb{E}_p\left[1\right] = \int p(z)dz = \frac{1}{Z_p}\int \tilde{p}(z)dz = \frac{1}{Z_p}\int \frac{\tilde{p}(z)}{q(z)}q(z)dz = \frac{1}{Z_p}\sum_{i=1}^{L} \frac{\tilde{p}(z^i)}{q(z^i)}, \quad z^i \sim q(z)$$

$$\Rightarrow Z_p = \sum_{i=1}^{L} \frac{\tilde{p}(z^i)}{q(z^i)}, \quad z^i \sim q(z)$$

Now using the above unbiased estimations, the estimation for the expectation is as following,

$$\Rightarrow \mathbb{E}_p\left[f\right] = \frac{\sum_{i=1}^{L} f(z^i)\frac{\tilde{p}(z^i)}{q(z^i)}}{\sum_{i=1}^{L} \frac{\tilde{p}(z^i)}{q(z^i)}}, \quad z^i \sim q(z).$$

This estimator is *biased* estimator of the target expectation (proof?), it is not always the case that the ratio of any two unbiased estimators is biased estimator (example?).

0.2.3 Gibbs sampling

Let’s say we want to sample from a multivariate distribution $p(x, y)$. Since sampling jointly sample from (x, y) we can sample for each variable, from the marginal distributions,

$$\begin{cases} x_t \sim p(x|y_{t-1}) \\ y_t \sim p(y|x_t) \end{cases}$$

In general this can be applied to any distribution with any number of the variables. More details on convergence proof and properties could be found at [Smith and Roberts(1993), Raftery and Lewis(1992)]. The idea of Gibbs sampling in statistics is very similar to “coordinate descent" optimization of multivariate objective functions in optimization(more?).

0.2.4 Markov Chain Monte Carlo(MCMC)

MCMC methods *implicitly* create makov chains which have the stationary distributions the same as that of the target distribution. At each step a new sample $x^{(i)}$ is proposed using a *transition distribution*, $\mathcal{P}(x, x^{'})$,

$$x^{(i-1)} \xrightarrow{\mathcal{P}} x^{(i)}.$$

There are many other names used to call this function, e.g. *Jumping Distribution*, *Proposal Distribution*, *Candidate Generating Distribution*.

```
Algorithm 1: Metropolis-Hastings algorithm
    [1] Start with random samples $z_0$, s.t. $p(z_0) > 0$. genrate random sample, $z^*$ from proposal distribution, $z_* \sim q(Z, z_t)$, given the sandom sample of the previous iteration $z_t$. Calculate:
    $\alpha = \min\left\{1, \frac{\tilde{p}(z_*)q(z_*, q_{t-1})}{\tilde{p}(z_{t-1})q(z_{t-1}, q_*)}\right\}$. $\alpha = \begin{cases} \geq 1 & : \text{Accept the sample: } x_t = z_* \\ < 1 & : \text{Accept the sample with probability of } \alpha. \end{cases}$
    TERMINATION-CONDITION
```

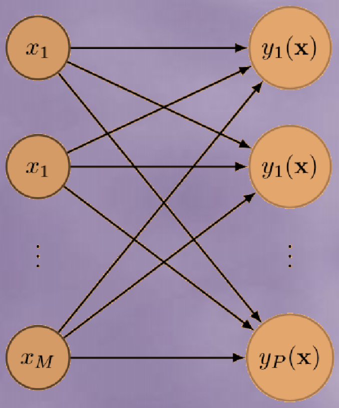
One of the families of MCMCs is Metopolis-Hastings methods which introduced in [Metropolis et al.(1953) Hastings(1970)]. Let's say we want to sample from $p(x) = \frac{1}{\mathcal{Z}}\tilde{p}(x)$, and let's assume that we don't have the normalization constant $\mathcal{Z}$. We define a transition distribution,

$$q(z_1, z_2) = \text{Pr}(z_1 \to z_2).$$

The steps of the algorithm are shown in Algorithm 1. To get a good approximation of the samples found from the above method, it is necessary to throw away the samples until a time *burn-in* period k where the samples $x_{k+1}, x_{k+2}, \ldots$ get closer to realistic samples of the target distribution, or the markov chain gets close enough to its stationary distribution. More proofs on the convergence and properties could be found at [Hastings(1970)].

0.3 Bibliographical notes

In preparation of this document I have used [Bishop(2006)].



Bibliography

[Beal(2003)] M.J. Beal. *Variational algorithms for approximate Bayesian inference*. PhD thesis, PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

[Bishop(2006)] C.M. Bishop. *Pattern recognition and machine learning*, volume 4. springer New York, 2006.

[Gilks and Wild(1992)] W.R. Gilks and P. Wild. Adaptive rejection sampling for gibbs sampling. *Applied Statistics*, pages 337–348, 1992.

[Gilks et al.(1995)Gilks, Best, and Tan] W.R. Gilks, NG Best, and KKC Tan. Adaptive rejection metropolis sampling within gibbs sampling. *Applied Statistics*, pages 455–472, 1995.

[Hastings(1970)] W.K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[Jeruchim(1984)] M. Jeruchim. Techniques for estimating the bit error rate in the simulation of digital communication systems. *Selected Areas in Communications, IEEE Journal on*, 2 (1):153–170, 1984.

[Metropolis et al.(1953)Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21:1087, 1953.

[Raftery and Lewis(1992)] A.E. Raftery and S. Lewis. How many iterations in the gibbs sampler. *Bayesian statistics*, 4(2):763–773, 1992.

[Smith and Roberts(1993)] A.F.M. Smith and G.O. Roberts. Bayesian computation via the gibbs sampler and related markov chain monte carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 3–23, 1993.