CSCI 590: Machine Learning

Lecture 25: Sampling and MCMC

Instructor: Murat Dundar

Acknowledgement:

1. PRML by C. Bishop
2. Markov Chain Monte Carlo Tutorial by Iain Murray

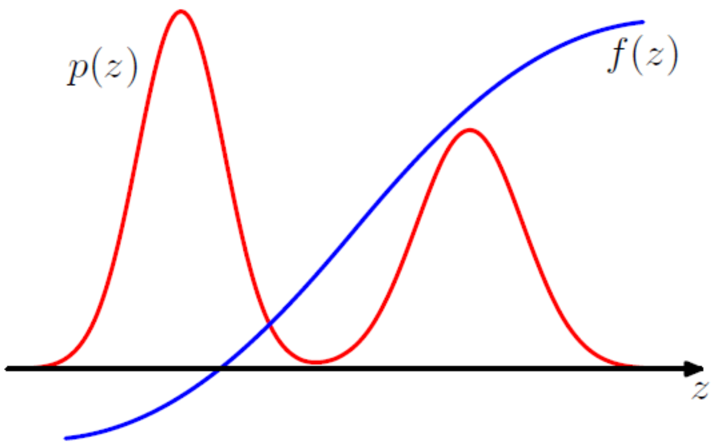
http://mlg.eng.cam.ac.uk/mlss09/

Monte Carlo integration (1)

We want to find the expectation of some function f(z) with respect to a probability distribution p(z).

$$\mathbb{E}[f] = \int f(\mathbf{z})p(\mathbf{z})\,\mathrm{d}\mathbf{z}$$

Schematic illustration of a function $f(z)$ whose expectation is to be evaluated with respect to a distribution $p(z)$.



Monte Carlo integration (2)

We obtain a set of samples $z^{(l)}$ where $l = 1, \ldots, L$ drawn independently from the distribution p(z) and approximate the expectation by the finite sum

$$\widehat{f} = \frac{1}{L} \sum_{l=1}^{L} f(\mathbf{z}^{(l)}).$$

$$\mathbb{E}[\widehat{f}] = \mathbb{E}[f]$$

$$\mathrm{var}[\widehat{f}] = \frac{1}{L} \mathbb{E}\left[(f - \mathbb{E}[f])^2\right]$$

We cannot always sample from p(z)!

Rejection sampling (1)

Suppose we wish to sample from a distribution p(z) whose inverse cdf does not exist in closed form

Suppose further that we can evaluate p(z) for any given value z, up to some normalizing constant

$$p(z) = \frac{1}{Z_p}\widetilde{p}(z)$$

where $\tilde{p}(z)$ can be evaluated but Z_p is unknown

Rejection sampling (2)

In rejection sampling we choose a proposal distribution q(z) from which we can easily draw samples.

We introduce a constant k whose value is chosen such that $kq(z) \geq \tilde{p}(z)$.

Each step of the rejection sampler involves generating two numbers.

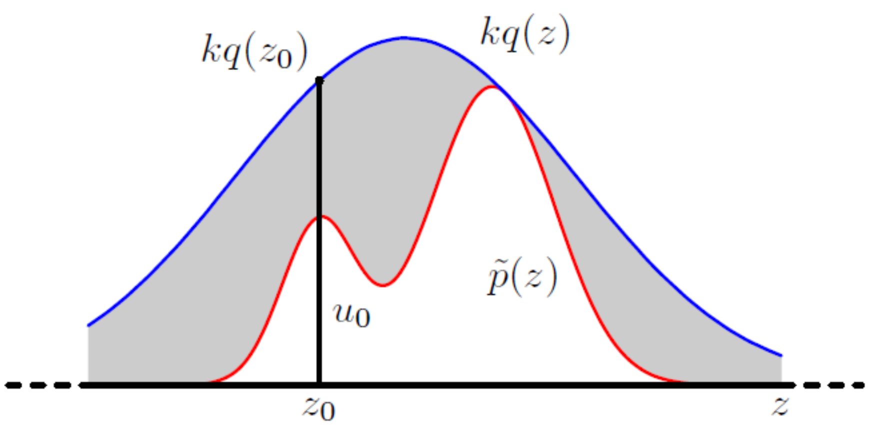
1. We generate a number z_0 from the distribution $q(z)$.
2. We generate a number u_0 from the uniform distribution $[0, kq(z_0)]$

Rejection sampling (3)

If $u_0 \geq \tilde{p}(z_0)$ reject else accept.

Thus, the pair (u_0, z_0) is rejected if it lies in the gray shaded region in the figure. The remaining pairs have uniform distribution under $\tilde{p}(z)$ and hence they are distributed according to p(z), which is the normalized verions of $\tilde{p}(z)$

In the rejection sampling method, samples are drawn from a simple distribution $q(z)$ and rejected if they fall in the grey area between the unnormalized distribution $\tilde{p}(z)$ and the scaled distribution $kq(z)$. The resulting samples are distributed according to $p(z)$, which is the normalized version of $\tilde{p}(z)$.



Important Sampling (1)

Rejection sampling can be very inefficient in approximating the expectation

$$\mathbb{E}[f] = \int f(\mathbf{z})p(\mathbf{z})\,\mathrm{d}\mathbf{z}$$

by the finite sum approximation to the expectation because only a very small proportion of samples drawn from a uniform distribution will make a significant contribution to the sum.

We want to choose the sample points where the product $f(z)p(z)$ is large.

Important Sampling (2)

We use a proposal distribution q(z) from which it is easy to draw samples.

$$\begin{aligned}\mathbb{E}[f] &= \int f(\mathbf{z})p(\mathbf{z})\,\mathrm{d}\mathbf{z} \\ &= \int f(\mathbf{z})\frac{p(\mathbf{z})}{q(\mathbf{z})}q(\mathbf{z})\,\mathrm{d}\mathbf{z} \\ &\simeq \frac{1}{L}\sum_{l=1}^{L}\frac{p(\mathbf{z}^{(l)})}{q(\mathbf{z}^{(l)})}f(\mathbf{z}^{(l)})\end{aligned}$$

The quantities $r_l = \frac{p(z^l)}{q(z^l)}$ are known as importance weights.

Unlike rejection sampling all samples are accepted but after correcting for the bias introduced by sampling from the wrong distribution.

Important Sampling (3)

In most cases p(z) can only be evaluated up to a normalization constant

$$p(z) = \frac{1}{Z_p}\widetilde{p}(z)$$

where $\tilde{p}(z)$ can be evaluated easily. Similarly,

$$q(\mathbf{z}) = \widetilde{q}(\mathbf{z})/Z_q$$

Important Sampling (4)

We then have

$$\begin{aligned}\mathbb{E}[f] &= \int f(\mathbf{z})p(\mathbf{z})\,\mathrm{d}\mathbf{z} \\ &= \frac{Z_q}{Z_p}\int f(\mathbf{z})\frac{\widetilde{p}(\mathbf{z})}{\widetilde{q}(\mathbf{z})}q(\mathbf{z})\,\mathrm{d}\mathbf{z} \\ &\simeq \frac{Z_q}{Z_p}\frac{1}{L}\sum_{l=1}^{L}\widetilde{r}_l f(\mathbf{z}^{(l)}).\end{aligned}$$

$$\tilde{r}_l = \frac{\tilde{p}(z^l)}{\tilde{q}(z^l)}$$

Important Sampling (5)

We can evaluate the ratio

$$\frac{Z_p}{Z_q} = \frac{1}{Z_q}\int \widetilde{p}(\mathbf{z})\,\mathrm{d}\mathbf{z} = \int \frac{\widetilde{p}(\mathbf{z})}{\widetilde{q}(\mathbf{z})} q(\mathbf{z})\,\mathrm{d}\mathbf{z}$$

$$\simeq \frac{1}{L}\sum_{l=1}^{L} \widetilde{r}_l$$

and hence

$$\mathbb{E}[f] \simeq \sum_{l=1}^{L} w_l f(\mathbf{z}^{(l)})$$

where

$$w_l = \frac{\widetilde{r}_l}{\sum_m \widetilde{r}_m} = \frac{\widetilde{p}(\mathbf{z}^{(l)})/q(\mathbf{z}^{(l)})}{\sum_m \widetilde{p}(\mathbf{z}^{(m)})/q(\mathbf{z}^{(m)})}$$

MCMC (1)

A first order Markov chain is a series of RVs such that the following conditional independence property holds

$$p(\mathbf{z}^{(m+1)}|\mathbf{z}^{(1)},\ldots,\mathbf{z}^{(m)}) = p(\mathbf{z}^{(m+1)}|\mathbf{z}^{(m)})$$

We can specify the Markov chain by the conditional probabilities in the form of transition probabilities

$$T_m(\mathbf{z}^{(m)},\mathbf{z}^{(m+1)}) \equiv p(\mathbf{z}^{(m+1)}|\mathbf{z}^{(m)})$$

The marginal probability for a particular variable in terms of the marginal probability for the previous variable in the chain

$$p(\mathbf{z}^{(m+1)}) = \sum_{\mathbf{z}^{(m)}} p(\mathbf{z}^{(m+1)}|\mathbf{z}^{(m)})p(\mathbf{z}^{(m)})$$

MCMC (2)

Homogeneity: A Markov chain is called homogeneous if the transition probabilities are the same for all m.

Invariance: A distribution is said to be invariant, or stationary, with respect to a Markov if each step in the chain leaves that distribution invariant.

$$p^{\star}(\mathbf{z}) = \sum_{\mathbf{z}'} T(\mathbf{z}', \mathbf{z}) p^{\star}(\mathbf{z}')$$

Detailed balance: Transition probabilities satisfy detailed balance when

$$p^{\star}(\mathbf{z}) T(\mathbf{z}, \mathbf{z}') = p^{\star}(\mathbf{z}') T(\mathbf{z}', \mathbf{z})$$

MCMC (3)

A sufficient (but not necessary) condition for ensuring that the required distribution p(z) is invariant is to choose the transition probabilities to satisfy the property of detailed balance.

$$p^\star(\mathbf{z})T(\mathbf{z},\mathbf{z}') = p^\star(\mathbf{z}')T(\mathbf{z}',\mathbf{z})$$

$$\sum_{\mathbf{z}'} p^\star(\mathbf{z}')T(\mathbf{z}',\mathbf{z}) = \sum_{\mathbf{z}'} p^\star(\mathbf{z})T(\mathbf{z},\mathbf{z}') = p^\star(\mathbf{z})\sum_{\mathbf{z}'} p(\mathbf{z}'|\mathbf{z}) = p^\star(\mathbf{z})$$

A Markov chain with detailed balance is *reversible*.

MCMC (4)

Our goal is to use Markov chains to sample from a given distribution. We can achieve this if we set up a Markov chain such that the desired distribution is invariant.

Ergodicity: We must also require that for $\mathrm{m} \to \infty$, the distribution $p\left(z^{(m)}\right)$ converges to $p^{*}(z)$ irrespective of the choice of initial distribution $p\left(z^{(0)}\right)$. This property is called ergodicity.

Metropolis-Hastings

$$A_k(\mathbf{z}^\star, \mathbf{z}^{(\tau)}) = \min\left(1, \frac{\widetilde{p}(\mathbf{z}^\star) q_k(\mathbf{z}^{(\tau)}|\mathbf{z}^\star)}{\widetilde{p}(\mathbf{z}^{(\tau)}) q_k(\mathbf{z}^\star|\mathbf{z}^{(\tau)})}\right)$$

p(z) is invariant distribution of the Markov chain defined by the Metropolis-Hastings algorithm because it satisfies the detailed balance property.

$$\begin{aligned} p(\mathbf{z}) q_k(\mathbf{z}|\mathbf{z}') A_k(\mathbf{z}', \mathbf{z}) &= \min\left(p(\mathbf{z}) q_k(\mathbf{z}|\mathbf{z}'), p(\mathbf{z}') q_k(\mathbf{z}'|\mathbf{z})\right) \\ &= \min\left(p(\mathbf{z}') q_k(\mathbf{z}'|\mathbf{z}), p(\mathbf{z}) q_k(\mathbf{z}|\mathbf{z}')\right) \\ &= p(\mathbf{z}') q_k(\mathbf{z}'|\mathbf{z}) A_k(\mathbf{z}, \mathbf{z}') \end{aligned}$$

Gibbs sampling (1)

Gibbs Sampling

1. Initialize $\{z_i : i = 1, \ldots, M\}$
2. For $\tau = 1, \ldots, T$:
   - Sample $z_1^{(\tau+1)} \sim p(z_1 | z_2^{(\tau)}, z_3^{(\tau)}, \ldots, z_M^{(\tau)})$.
   - Sample $z_2^{(\tau+1)} \sim p(z_2 | z_1^{(\tau+1)}, z_3^{(\tau)}, \ldots, z_M^{(\tau)})$.

     $\vdots$
   - Sample $z_j^{(\tau+1)} \sim p(z_j | z_1^{(\tau+1)}, \ldots, z_{j-1}^{(\tau+1)}, z_{j+1}^{(\tau)}, \ldots, z_M^{(\tau)})$.

     $\vdots$
   - Sample $z_M^{(\tau+1)} \sim p(z_M | z_1^{(\tau+1)}, z_2^{(\tau+1)}, \ldots, z_{M-1}^{(\tau+1)})$.

Gibbs sampling (2)

Gibbs sampling draws from the right distribution because:

1. p(z) is invariant because conditional distributions together define the joint distribution.
2. Markov chain is ergodic

Gibbs sampling can be obtained as a special case of the Metropolis Hastings algorithm.