Sampling according to the multivariate normal density

Ravi Kannan
School of Computer Science
Carnegie-Mellon University
Pittsburgh, PA 15213*
ravi.kannan@cs.cmu.edu

Guangxing Li
Department of Mathematics
Carnegie-Mellon University
Pittsburgh, PA 15213

April 2, 1999

Abstract

This paper [1] deals with the normal density of n dependent random variables. This is a function of the form :

$$ce^{-x^T Ax}$$

where A is an $n \times n$ positive definite matrix, x is the $n-$vector of the random variables and c is a suitable constant.

The first problem we consider is the (approximate) evaluation of the integral of this function over the positive orthant

$$\int_{x_1=0}^{\infty} \int_{x_2=0}^{\infty} \cdots \int_{x_n=0}^{\infty} ce^{-x^T Ax}.$$

This problem has a long history and a substantial literature . Related to it is the problem of drawing a sample from the positive orthant with probability density (approximately) equal to $ce^{-x^T Ax}$. We solve both these problems here in polynomial time using rapidly mixing Markov Chains. For proving rapid convergence of the chains to their stationary distribution, we use a geometric property called the Isoperimetric Inequality. Such an inequality has been the subject of recent papers for general log-concave functions. We use these techniques, but the

*Supported in part by NSF grant CCR 9528973.

[1] To appear in the Proceedings of the IEEE Symposium on the Foundations of Computer Science, 1996

main thrust of the paper is to exploit the special property of the normal density to prove a stronger inequality than for general log-concave functions.

We actually consider first the problem of drawing a sample according to the normal density with A equal to the identity matrix from a convex set K in $\mathbf{R}^n$ which contains the unit ball. This problem is motivated by the problem of computing the volume of a convex set in a way we explain later. Also, the methods used in the solution of this and the orthant problem are similar.

1 Introduction

The problem of integrating the multivariate normal density over certain convex sets, especially orthants is a classical problem. It is related to other problems in probability and statistics. The approximation of orthant probabilities of the normal distribution has an extensive literature, see for example the results and bibliographies in [15], [10], [16] and [8]. The problem of integration is well-known to be related to the problem of sampling and many methods are known for estimating the integral based on a number of samples. (For example, see [15] , [8]). However, the difficulty with the known methods is that one needs a number of samples that grows exponentially with the number of dimensions n. For this reason, it has been noted ([15],[8]) that as a rule, the estimation of the integral is only feasible for very modest n.

There is an even more fundamental problem : that of sampling according to e^{-x^TAx} in, say, the positive orthant quickly. (I.e., in time bounded by a polynomial in n.)

[We use the phrase "sampling according to a (nonnegative real valued function) $F(x)$ in a region R" to mean : pick $x \in R$ at random from the probability distribution over R with density $F(x)/\int_R F$.]

It is easy to sample in the whole space according to e^{-x^TAx} : one simply draws a sample according to the standard normal density $e^{-|x|^2}$ by any one of a number of schemes and applies a (suitable) linear transformation to it. (We omit the details.) But, then if one wishes to sample in the positive orthant according to this function, by say Rejection Sampling, we expect to draw

$$\frac{\int_{\mathbf{R}^n} e^{-x^TAx}}{\int_{\mathbf{R}^n_+} e^{-x^TAx}}$$

samples before one lands in the positive orthant $\mathbf{R}^n_+$. This ratio in general will grow exponentially with n.

We use instead a Markov Chain with the desired steady state density - $e^{-x^TAx}/\int_{\mathbf{R}_+^n} e^{-x^TAx}$ in $\mathbf{R}_+^n$. The chain itself is a natural "Metropolis" algorithm. The Metropolis algorithm can be used in a wide range of contexts - see [5] for a recent survey. In our cases, the chain is easily described : Suppose we wish to sample from a set R according to function F. We first choose a "step size" δ suitably. (To be described later.) Suppose the Chain is at a point $x \in R$. Then we choose a random point y from a ball of radius δ around x uniformly. If y is not in R, we stay at x. Otherwise, if $F(y) \geq F(x)$, we move to y, else, we move to y with probability $F(y)/F(x)$. Formally, the probability density P_{xy} of moving from x to y is given by :

$$P_{xy} = \begin{cases} \frac{1}{\text{vol}\delta B} \min\{1, \frac{F(y)}{F(x)}\} & \text{if } x \neq y \in R;\ |x-y| \leq \delta; \\ 0 & \text{otherwise .} \end{cases}$$

$$P_{xx} = 1 - \int_{y \neq x} P_{xy}\ dy.$$

We call this the **Metropolis Chain according to** F. It is easily see that its steady state distribution is $F(x)/\int_R F$.

Our proof that this chain "mixes rapidly" (i.e., converges to the steady state fast) uses techniques developed recently. These techniques were developed to solve the problem of estimating the volume of a convex set (this problem itself motivates another sampling question in this paper) in [7],[14],[3],[6] (see also [12] for a survey.) Using these techniques, one relates the convergence to a purely geometric property called the Isoperimetric Inequality. In this paper, we exploit the special nature of the normal density (specifically the fact that its logarithm declines quadratically in every direction) to prove a sharper Isoperimetric Inequality than the one for general log-concave functions proved in [3], [9], [6] and other papers. We remind the reader that a function is said to be log-concave if its logarithm is concave.

The Markov Chain yields samples (nearly) according to $F(x)$. We show how to use a polynomial (in n and an error parameter) number of samples to estimate the integral of F over the positive orthant. To avoid the problem of high variance, we use a trick often used in volume estimation algorithms - the integral is estimated as a product of polynomially many quantities, each of low variance.

Our sampling algorithm uses $O^*(\Lambda^2 n^3)$ steps (of the random walk) for the first sample and $O^*(\Lambda^2 n^2)$ steps for each subsequent nearly independent sample, where Λ is the condition number of A, namely the ratio of the largest eigenvalue to the smallest. We also show that we can estimate the orthant integral with relative error ε with $O^*(\Lambda^2 n^2/\varepsilon^2)$ samples.

Before tackling the orthant problem, we first consider the problem of drawing a sample according to the standard normal (we use this phrase here to mean the case of $A = I$) from a convex set K containing the unit ball. For this problem, our algorithm finds the first sample in $O^*(n^4)$ steps and each subsequent (nearly independent) sample in $O^*(n^3)$ steps. [The O^* notation hides constant factors, log factors and factors depending upon the accuracy needed.]

A crucial parameter in the algorithm is the "step size" δ. In the present paper, we use $\delta = \Theta(\frac{1}{n})$ for the latter problem (of sampling from a convex set containing the unit ball according to the standard normal density) to achieve the bounds on the number of steps claimed. By using the techniques of [13], we hope to show that a step size $\delta = \Theta(\frac{1}{\sqrt{n}})$ may be used for the subsequent samples, which would then improve the number of steps needed for each subsequent sample in this problem to $O^*(n^2)$ steps. Now by direct integration, it is not difficult to show that we may discard the part of K beyond a ball of radius $O(\sqrt{n})$ from the origin as this has negligible measure. Thus in this case we have that $B \subseteq K \subseteq c\sqrt{n}B$. The number of steps needed in such a situation for the algorithm in [13] to draw each subsequent sample is $O^*(n^3)$ (see Theorem 2.2 of [13]). So, if the improvement holds, then at least under the normal density, samples come faster than the best known for the uniform density. This may ultimately help improve the complexity of the difficult problem of estimating the volume of a convex set.

2 Isoperimetric Inequalities

Isoperimetric inequalities play an important role in analysis of the mixing time of Markov chains. This is because they can be related to the conductance of the chain, a notion introduced by Jerrum and Sinclair [11] which is very useful in the analysis of convergence.

Roughly speaking an isoperimetric inequality in this context says that for any partition of a convex set K into two parts, the measure (according to the desired steady state function F) of the dividing surface is not too small compared with the minimum of the measures of the two parts. The version we use here is slightly different - here, we consider a division into three parts K_1, K_2, K_3 where K_1, K_2 correspond to the two main parts described above and K_3 may be thought of as the set of points within a certain distance

$\sigma/2$ of the dividing surface above. We now state precisely our isoperimetric inequality; this is the central new theorem we prove and use in the paper. One important feature of this is that there is no dependence on the diameter of the convex set K.

Theorem 2.1 *Let $F(x) = e^{-|x|^2} g(x)$, where $g(x)$ is an arbitrary log-concave function. Suppose K is a convex set in $\mathbf{R}^n$. K_1 and K_2 are two measurable subsets of K, and the distance between K_1 and K_2 (i.e., $\inf_{x\in K_1, y\in K_2} |x-y|$) is σ. $K_3 = K\backslash(K_1 \cup K_2)$. The following inequality holds.*

$$\int_{K_3} F(x)\,dx \geq \frac{2\sigma e^{-\sigma^2}}{\sqrt{\pi}} \min\{\int_{K_1} F(x)\,dx,\ \int_{K_2} F(x)\,dx\} \tag{1}$$

Idea of Proof : We use a powerful technique due to Lovász and Simonovits [14] to prove this result. This technique uses their Localization Lemma to reduce the $n-$ dimensional case to essentially a 1 dimensional problem. In this situation, this lemma implies that the assertion in the Theorem holds provided for each line segment $[a\ b]$ and each linear function $\ell(\cdot)$, we have the following :

$$\int_{K_3\cap[a\ b]} F\ell^{n-1} \geq \frac{2\sigma e^{-\sigma^2}}{\sqrt{\pi}} \min\{\int_{K_1\cap[a\ b]} F\ell^{n-1}dx, \int_{K_2\cap[a\ b]} F\ell^{n-1}\} \quad (1')$$

where now all the integrals involved are just one dimensional integrals. (Intuitively, it says that the inequality asserted in the Theorem holds provided it holds for every pointed cone with an infinitismal cross section.) As we move along the line $[a\ b]$, x is a function of one parameter t. We abuse notation and let all the functions involved be functions of t. Also, we let $G(t) = g(t)\ell(t)^{n-1}$ and note that G is log-concave. Then, we have that $F(t) = e^{-t^2+\delta_1 t+\delta_2}G(t)$ for some constants δ_1, δ_2. In this sketch of the proof, we also deal only with the case when $K_3 \cap [a\ b]$ is an interval; the more general case can be dealt with the same ideas and the full proof is deferred to the final paper. By the hypothesis of the Theorem, we have in this case that $K_3 \cap [a\ b]$ is an interval of length at least σ. It will suffice to prove that for every real number ν :

$$\int_{\nu}^{\nu+\sigma} e^{-t^2+\delta_1 t+\delta_2} G(t)dt \geq \frac{2\sigma e^{-\sigma^2}}{\sqrt{\pi}}$$
$$\min\{\int_{-\infty}^{\nu} e^{-t^2+\delta_1 t+\delta_2} G(t)dt, \int_{\nu+\sigma}^{\infty} e^{-t^2+\delta_1 t+\delta_2} G(t)dt\}. \tag{2}$$

There are positive reals c_1 and c_2 such that $G(\nu) = c_1 e^{c_2 \nu}$ and $G(\nu+\sigma) = c_1 e^{c_2(\nu+\sigma)}$. By using the log-concavity of $G(t)$, we can see that the following is true:

$$G(t) \begin{cases} \leq c_1 e^{c_2 t} & t \in (-\infty, \nu) \cup (\nu+\sigma, \infty) \\ \geq c_1 e^{c_2 t} & t \in (\nu, \nu+\sigma) \end{cases}$$

So we can replace $G(t)$ in (2) by $e^{c_2 t}$ and prove the resulting inequality :

$$\int_\nu^{\nu+\sigma} e^{-t^2+\delta_1' t+\delta_2'} dt \geq \frac{2\sigma e^{-\sigma^2}}{\sqrt{\pi}} \min\{\int_{-\infty}^{\nu} e^{-t^2+\delta_1' t+\delta_2'} dt, \int_{\nu+\sigma} e^{-t^2+\delta_1' t+\delta_2'} dt\} \quad (2')$$

After cancelling common factors, it suffices to prove the following.

Lemma 2.2 *The following inequality holds:*

$$\int_\nu^{\nu+\sigma} e^{-t^2} dt \geq \frac{2\sigma e^{-\sigma^2}}{\sqrt{\pi}} \min\{\int_{-\infty}^{\nu} e^{-t^2} dt, \int_{\nu+\sigma}^{\infty} e^{-t^2} dt\} \tag{3}$$

Proof. Without loss of generality, we assume that $\nu > 0$. The other case can be dealt with exactly symmetrically. Consider the following function of x:

$$\phi(x) = \frac{\int_x^{x+\sigma} e^{-t^2} dt}{\int_{x+\sigma}^{\infty} e^{-t^2} dt}$$

This function is increasing for $x > 0$, as seen by differentiating it. This means that it suffices to consider the case when $\nu = 0$. In this case, we have

$$\int_0^\sigma e^{-t^2} dt \geq \sigma e^{-\sigma^2}.$$

Also, we have

$$\int_\sigma^\infty e^{-t^2} dt \leq \frac{\sqrt{\pi}}{2}$$

Therefore the lemma follows and the theorem is proved.

For the general normal density function, we have similar result and the proof of it is similar to that of the standard case.

Theorem 2.3 *Let $F(x) = e^{-xAx^T} g(x)$, where A is a positive definite matrix and $g(x)$ is an arbitrary log-concave function. K is a convex set in R^n. K_1 and K_2 are two subsets of K. The distance between K_1 and K_2 is σ.*

$K_3 = K \backslash (K_1 \cup K_2)$. *The following inequality holds, where λ_1 is the smallest eigenvalue of A :*

$$\int_{K_3} F(x)dx \geq \frac{2\sqrt{\lambda_1}\sigma e^{-\sigma^2}}{\sqrt{\pi}} \min\{\int_{K_1} F(x)dx, \int_{K_2} F(x)dx\} \tag{4}$$

Proof. Again, with the Localization Lemma, we can reduce it to the one dimensional case. Let v be the direction of this $[a\ b]$, then

$$\begin{aligned} F(t) &= e^{-(a+tv)A(a+tv)^T} g(a+tv) \\ &= e^{-(t^2\sigma_1 + t\sigma_2 + \sigma_3)} g(a+tv) \end{aligned}$$

where $\sigma_1 = vAv^T \geq \lambda_1$, $\sigma_2 = aAv^T + vAa^T$ and $\sigma_3 = aAa^T$. Note that g is again log-concave. By following the same procedure as in the proof of the last Theorem, we can see that it suffices to prove the following

$$\int_{\nu}^{\nu+\sigma} e^{-\sigma_1 t^2} dt \geq \frac{2\sqrt{\lambda_1}\sigma e^{-\sigma^2}}{\sqrt{\pi}} \min\{\int_{-\infty}^{\nu} e^{-\sigma_1 t^2} dt, \int_{\nu+\sigma} e^{-\sigma_1 t^2} dt\}$$

With a simple variable substitution $u = \sqrt{\sigma_1} t$ and (3), as well as the fact that $\sigma_1 \geq \lambda_1$ we can derive (4). This completes the proof.

3 Rapid Mixing

3.1 Local conductance, step size and conductance

A central quantity of interest in rapid mixing proofs is the notion of conductance introduced by Jerrum and Sinclair. Suppose we have a Markov Chain in K with steady state probability density $\pi(\cdot)$. The conductance of the chain is defined as the infimum over all (measurable) partitions of K into two sets S_1, S_2 of the quantity below :

$$\frac{\int_{x \in S_1} \int_{y \in S_2} \pi(x) P_{xy}\ dy\ dx}{\min\{\int_{S_1} \pi(x)\, dx,\ \int_{S_2} \pi(y)\, dy\}}. \tag{5}$$

If $\pi(S_1) \leq 1/2$, then quantity is the conditional probability of going from S_1 to S_2 in one step conditioned on starting in the steady state in S_1. The central result of Jerrum and Sinclair is that under mild conditions on the chain (which hold for us), high conductance implies rapid convergence. Below we will use a more directly relevant version of this from [14].

In the analysis of convergence rates of geometric random walks, a lower bound on conductance has often been proved using an Isoperimetric Inequality. (See [7], [14],[3].) The connection between conductance and Isoperimetry is intuitively not difficult to see. But to carry out the argument, we need another quantity, namely the local conductance.

The local conductance at a point x in space is defined as the probability of making a proper move from x in one step. For example, if only an exponentially small volume of the δ ball around x is contained in the convex set, then the local conductance of the Metropolis random walk is exponentially small. In this case, we have to make exponentially many steps before moving away from x, so we cannot guarantee rapid convergence. [In a sense, one may view local conductance as conductance specialized to single element sets.]

We establish the following result about the conductance in terms of local conductance and step size. In a manner similar to the proof of Theorem 3.2 in [14], we relate conductance to Isoperimetric Inequality, but in order to achieve better coefficient, we will carry out the estimation more carefully. Then we use our Isoperimetric Inequality (Theorem 2.3). The full proof is deferred to the final paper.

Theorem 3.1 *Suppose*

$$F(x) = e^{-x^T Ax} g(x),$$

where g is a log-concave function and A is a positive definite matrix with least eigenvalue λ_1. Suppose we are doing Metropolis random walk with respect to $F(x)$. Also, suppose the step size is δ and the local conductance everywhere is at least a certain $l > 0$. Then the conductance is at least

$$\Phi \geq \frac{l^2 \delta \sqrt{\lambda_1}}{2\sqrt{n\pi}}.$$

3.2 M-distance

Now we relate conductance to convergence. To do so, we use a quantity called the "M-distance" which measures how far apart two distributions are. The M-distance between two distributions P and Q is defined as (it is really not a distance, but is however a convenient measure of the difference) :

$$M(P,Q) = \sup_{S:0<Q(S)\leq 1/2} \frac{|P(S) - Q(S)|}{\sqrt{Q(S)}}.$$

See [13] for a note about why the M distance is convenient to use. The more common measure used is the total variation distance between P and Q which is denoted $|P-Q|_{\text{tv}}$ and defined as

$$\sup_S |P(S) - Q(S)|.$$

Clearly the M-distance is an upper bound on the total variation distance.

In the context of this paper, we run a Markov chain starting from distribution Q_0 and achieve distribution Q_k at step k. Our steady state distribution is denoted Q. Therefore we want to know the bound for $M(Q_k, Q)$. Our main tool is the following theorem from [14].

Theorem 3.2 *Suppose Φ is the conductance of the Markov chain, we have*

$$M(Q_k, Q) \leq M(Q_0, Q)\left(1 - \frac{\Phi^2}{2}\right)^k \tag{6}$$

In many contexts, it is not sufficient to get one sample. We like many independent samples all from (nearly) the same distribution Q. To get this, one may of course restart the random walk from a different (independently chosen) point. We will see that this is not necessary. In fact, in general, for geometric chains as in this paper, it turns out that it is easy to start from distributions Q_0 which have $M(Q_0, Q) \leq e^{O^*(n)}$ (as we show later) and so applying the above Theorem, one sees that in $O^*(n \log(1/\varepsilon)/\Phi^2)$ steps, we can get a sample from a distribution with M-distance at most ε from the desired Q. Intuitively, now for the second sample, it is better to start from the sample we already have, but this introduces dependence. However, we will argue below that after $O(1/\Phi^2)$ steps, we will have a second sample which is "nearly" independent of the first and is from a distribution with M-distance at most ε again to Q. This improves steps needed for subsequent samples by a factor of $O^*(n)$. Something akin to this happens for general Markov Chains, as first pointed out in [1] for discrete chains. Here, we will follow the simple proof in [13]. First, we need a definition of a particular weak form of independence that turns out to be sufficient for us.

Definition : Suppose X, Y are two random variables (with values in two possibly different $\sigma-$ algebras) and ε any positive real. We say that X, Y are $\varepsilon-$independent if

$$|\text{Prob}(X \in A\ ,\ Y \in B) - \text{Prob}(X \in A)\text{Prob}(Y \in B)| \leq \varepsilon$$

for all measurable sets A, B.

Corollary 3.3 *Suppose we start from a point w_0 drawn according to distribution Q_0 and run the Metropolis random walk with conductance Φ for t steps to get a point w_t. Then w_o and w_t are $\tau-$independent where $\tau = [2M(Q_0, Q) + 1]\left(1 - \frac{1}{\Phi^2}\right)^t$.*

Proof. To see the τ-independence, we argue as follows. Let A, B be any two measurable sets. Let $f(A, B)$ be defined as

$$\begin{aligned} l&|\text{Prob}(w_0 \in A, w_t \in B) - \text{Prob}(w_0 \in A)\text{Prob}(w_t \in B)| \\ &= Q_0(A)|\text{Prob}(w_t \in B \mid w_0 \in A) - Q_t(B)| \end{aligned} \tag{7}$$

Let Q_0' be the distribution of w_0 conditioned on it being in A, i.e., $Q_0'(S) = Q_0(S \cap A)/Q_0(A)$ for any measurable S. Then, $\text{Prob}(w_t \in B|w_0 \in A)$ is the distribution of w_t if we start with w_0 drawn from Q_0'. So, applying the theorem above, we get

$$\begin{aligned} &|\text{Prob}(w_t \in B \mid w_0 \in A) - Q(B)| \\ &\leq \sqrt{Q(B)}M(Q_0', Q)\left(1 - \frac{1}{\Phi^2}\right)^t. \end{aligned} \tag{8}$$

For any measurable S,

$$Q(S) = \frac{1}{Q_0(A)}Q(S \cap A) - \left(\frac{1}{Q_0(A)} - 1\right)Q(S \cap A) + Q(S \setminus A),$$

and hence

$$\begin{aligned} |Q_0'(S) - Q(S)| \;\; \leq \;\; & \frac{1}{Q_0(A)}|Q_0(S \cap A) - Q(S \cap A)| \\ & + \left(\frac{1}{Q_0(A)} - 1\right)Q(S \cap A) + Q(S \setminus A). \end{aligned}$$

So, we have $M(Q_0', Q) \leq \frac{1}{Q_0(A)}[M(Q_0, Q) + 1]$ and now using (7) and (8), and the theorem, we get the claimed τ-independence.

In the rest of this write-up, we will make assertions of the form "each subsequent nearly independent sample takes ..... number of steps". The specifications of how nearly independent are postponed to the final paper. But such assertions will all be based on the Corollary above and independence will mean the ε independence we have above.

4 The standard normal density case

Now we consider the case: sampling according to the standard normal density $e^{-|x|^2}$ in a convex set K with $B(0,1) \subseteq K$. This is split to several steps: penalty function, step size and local conductance, conductance and sampling.

4.1 Penalty function

If we run the Metropolis random walk in K, according to just F, we run into the following problem : K may have very sharp corners and the local conductance may be exponentially small in such a corner.

One way to handle sharp boundary is introducing penalty function of the convex set as used in [3],[14] etc. The idea is: instead of restricting the random walk to K, we allow it to go outside of K, but "dampen" the probability outside more as we go away from K . We dampen by using some appropriate penalty function to penalize being outside.

Definition 4.1 *Given a convex set K with the origin in its interior , the gauge function of it is defined as follows:*

$$\varphi_K(x) = \begin{cases} 1 & \text{if } x \in K; \\ \min\{t : x \in tK\} & \text{if } x \notin K. \end{cases}$$

It is well-known that $g(x) = e^{-c(\varphi_K(x)-1)}$ is a log-concave function for any positive constant c. We use this as the penalty function.

We will do a Metropolis random walk in all of $\mathbf{R}^n$ according to the function

$$F(x) = e^{-|x|^2 - c(\varphi_K(x)-1)}$$

for a certain number of steps to be specified and reject if the resulting sample is not in K. We will choose c big enough to guarantee that the probability of being outside of K is small, so that we do not reject too many times before we accept a sample.

It can be shown that with $c = \frac{3}{2}n$, the probability outside is at most 3 times of that of inside. [This is shown by showing that it holds in every infinitismal pointed cone emanating from the origin.] The proof is deferred to the final paper.

From now on, we sample in the whole of R^n according to the following function:

$$F(x) = e^{-|x|^2 - \frac{3}{2}n(\varphi_K(x)-1)} \tag{9}$$

Proposition 4.2 *The probability inside K of distribution defined by (9) is at least* 0.25.

4.2 Step size and local conductance

Choosing appropriate step size is tricky. Usually, small step size gives large local conductance. But if it is too small, it lowers the lower bound on conductance that we use (namely from Theorem 3.1). So, there is a balance between small and large step size. Here it turns out that we want to choose step size with order $\frac{1}{n}$. In order to achieve the best bound for conductance, we use a parameter $1 \leq c \leq 3$, and choose $\frac{1}{cn}$ as the step size. c will be identified when we discuss conductance. We prove in the final paper that :

Lemma 4.3 *With step size $\delta = \frac{1}{cn}$, the local conductance at every point is at least*

$$\frac{1}{2}e^{-\frac{1}{c^2n^2}-\frac{3}{2c}}.$$

4.3 Conductance and convergence rate

We can plug in the above value for l (and $\delta = 1/(cn)$) into Theorem 3.1 and choose the best value for c. It is convenient to ignore the $e^{-\frac{1}{c^2n^2}}$ term and then to differentiate with respect to a new variable $\lambda = 1/c$. This tells us to choose $c = 3$.

Now to describe the random walk, one only needs to specify the starting distribution, for which we just use the uniform distribution over the unit ball. Then we get the following :

Algorithm Starting from a point w_0 picked according to the uniform density from the unit ball, do the Metropolis random walk in all of $\mathbf{R}^n$ according to the function $F(x)$ defined in (9) using step size $\delta = 1/(3n)$. [We do the walk for a number of steps chosen to ensure that the M distance to the steady state is within a desired bound (see below); if the resulting point is not in K, we reject and repeat.]

In the final paper, we show the following facts about this walk.

Theorem 4.4 *The conductance of the above Markov chain is at least*

$$\frac{1}{20e\sqrt{\pi}n^{3/2}} \tag{10}$$

Proposition 4.5

$$M(Q_0, Q) \leq 1 + e\pi^{n/2}n!.$$

From (10) and (6), we get that we can get the first sample in number of steps

$$O^*(n^4).$$

Also, from Corollary 3.3 the number of steps for each subsequent (nearly independent) sample is

$$O^*(n^3).$$

5 Orthant case with general normal distribution

Now we consider the general normal distribution $f(x) = e^{-x^T Ax}$ in the orthant $\Omega = \{x : x_i \geq 0, i = 1, 2, ..., n\}$. A is a positive definite matrix with eigenvalues $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. Let $\Lambda = \lambda_n/\lambda_1$.

5.1 Penalty function

Again we introduce a penalty function to overcome the sharp boundary problem. It is :

Definition 5.1 *Function $\varphi(x)$ is defined as follows:*

$$\varphi(x) = \begin{cases} 0 & \textit{if } x_i \geq 0 \textit{ for each } i; \\ \max\{-x_i\sqrt{n} : \; x_i < 0\} & \textit{otherwise.} \end{cases}$$

In other words, for x not in the orthant, if we draw a line starting from it and along direction $\tau = (\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, ..., \frac{1}{\sqrt{n}})$, then $\varphi(x)$ is the distance between x and the intersecting point of this line with the boundary of the orthant. It is easy to show that this function is convex.

The proof of the following proposition is technically complicated and long and is deferred to the final paper.

Proposition 5.2 *Suppose that $c = 4\alpha\lambda_n\sqrt{n}/\sqrt{\lambda_1}$ and $\alpha \geq 1$. Then with density function proportional to the following F in all of $\mathbf{R}^n$, the probability inside the orthant is at least 0.25.*

$$F(x) = e^{-xAx^T - c\varphi(x)} \tag{11}$$

5.2 Step size, local conductance and conductance

From now on, we sample in the whole R^n according to the following function:

$$F(x) = e^{-xAx^T - (4\alpha\lambda_n\sqrt{n}/\sqrt{\lambda_1})\varphi(x)}.$$

The step size used is

$$\delta = \frac{\sqrt{\lambda_1}}{\gamma\lambda_n\sqrt{n}\ln n}$$

where $\gamma \geq 2$ is a positive real. It can be proved that with this step size, the local conductance is at least

$$(\frac{1}{2} - \frac{1}{4\gamma\sqrt{n}\ln n})\omega$$

where ω is a positive real. The proof involves many details, and thus will be presented in the final paper.

We then optimize the parameters as in the other problem. Here is the algorithm with the parameters chosen (nearly) optimally.

Algorithm Start at w_0 chosen according to the uniform density on the unit ball intersected with the positive orthant. Then do the Metropolis random walk in all of $\mathbf{R}^n$ according to the function

$$F(x) = e^{-x^TAx - 5.5\frac{\sqrt{n}\lambda_n}{\sqrt{\lambda_1}}\varphi(x)}$$

and with step size

$$\delta = \frac{\sqrt{\lambda_1}}{16\lambda_n\sqrt{n}\log n}.$$

We do the random walk for a number of steps dictated by the desired M distance to steady state as determined by Proposition 5.3, Theorem 5.4 , Theorem 3.2 and Corollary 3.3.

We prove the following about this walk.

Proposition 5.3

$$M(Q_0, Q) \leq 1 + 2^n\frac{e^{\lambda_n}}{\lambda_1}\pi^{n/2}n!.$$

Theorem 5.4 *The conductance of this chain is at least*

$$\frac{1}{900\Lambda n\log n}. \tag{12}$$

Just like the standard normal case that we discussed in the last section, we can give bounds for the first sample mixing time and the subsequent sample mixing time. The first sample takes $O^*(n^3\Lambda^3)$ steps. Each subsequent (nearly independent) sample takes number of steps

$$O^*(n^2\Lambda^2).$$

6 From Sampling to Integration

In this section we show how the ability to sample from $\mathbf{R}_+^n$ according to functions of the form e^{-x^TAx} helps us estimate the integrals of such functions over the orthant.

The connection between sampling and integration is in a sense a natural and old one, but the point here is to describe a scheme which given a polynomial time sampling algorithm, can estimate the integral in polynomial time.

For $p \in [0 \;\; 1]$, define

$$f(p) = \int_{\mathbf{R}_+^n} e^{-px^TAx-(1-p)|x|^2}\, dx = \int_{\mathbf{R}_+^n} e^{-x^TMx}\, dx,$$

where $M = pA + (1-p)I$ is a positive definite matrix. $f(0)$ is known and we are interested in $f(1)$.

Let X be a ($n-$vector) random variable with values in $\mathbf{R}_+^n$ and with density

$$e^{-px^TAx-(1-p)|x|^2}/f(p).$$

Let ν be a positive real number to be chosen later. Let $Y = e^{-\nu X^TAX+\nu|X|^2}$. Then, clearly, we have

$$E(Y) = \frac{f(p+\nu)}{f(p)}.$$

We will use this to estimate $\frac{f(p+\nu)}{f(p)}$ and we will then find $f(1)/f(0)$ as the telescoping product of several such ratios. First, let us bound the number of samples of X needed.

We have

$$\mathrm{Var}(Y) \le E(Y^2) = \frac{f(p+2\nu)}{f(p)}.$$

So, we have

$$\frac{\mathrm{Var}(Y)}{(E(Y))^2} \le \frac{f(p)f(p+2\nu)}{f(p+\nu)f(p+\nu)}.$$

We will use the following Lemma to bound this ratio.

Lemma Assuming $\lambda_1 = 1$ after rescaling if necessary, we have

$$\left|\frac{d(\log f)}{dp}\right| \le \frac{n}{2}\Lambda \qquad \forall p.$$

Proof. We see by differentiating that

$$\frac{df}{dp} = \int_{\mathbf{R}_+^n} [-x^T Ax + |x|^2] e^{-x^T Mx}\ dx,$$

Since $x^T Ax \le \lambda_n |x|^2$, we have

$$\left|\frac{df}{dp}\right| \le \lambda_n \int_{\mathbf{R}_+^n} |x|^2 e^{-x^T Mx}\ dx.$$

Define a function $g : \mathbf{R}_+ \to \mathbf{R}$ by

$$g(t) = \int_{\mathbf{R}_+^n} e^{-tx^T Mx}\ dx.$$

Then with the change of variables $y_i = \sqrt{t} x_i$, we see that $g(t) = t^{-n/2} g(1)$. Differentiating we get $dg/dt = -n/2$ at $t = 1$. But differentiating under the integral we get (noting that the eigenvalues of M are all also at least 1),

$$-\frac{dg}{dt} = \int_{\mathbf{R}_+^n} (x^T Mx) e^{-tx^T Mx}\ dx \ge \int_{\mathbf{R}_+^n} |x|^2 e^{-tx^T Mx}\ dx.$$

So we get

$$|\frac{df}{dp}| \le \Lambda f(p),$$

from which the lemma follows.

From the Lemma we get

$$\left|\log\left(\frac{f(p+2\nu)}{f(p+\nu)}\right)\right| = \int_{q=p+\nu}^{p+2\nu} \left|\frac{d(\log f)}{dp}\right| dq \le n\nu\Lambda/2.$$

The same bound holds for $\left|\log\left(\frac{f(p+\nu)}{f(p)}\right)\right|$. So, choosing

$$\nu = \frac{1}{n\Lambda},$$

we see that

$$\frac{\mathrm{Var}(Y)}{(E(Y))^2} \leq 3.$$

Thus then the ratio $f(p+\nu)/f(p)$ can be estimated to relative error at most δ with $O(1/\delta^2)$ samples. Let $m = n\Lambda$.

We use the notation introduced in this section to describe the algorithm, stating the number of samples used etc.

Algorithm to estimate the integral

- Given $A; 0 \leq \varepsilon \leq 1/8$ and $\delta > 0$, the algorithm will find a real number T such that with probability at least $1-\delta$, we have :

$$Te^{-\varepsilon} \leq \int_{\mathbf{R}_+^n} e^{-x^TAx} \leq Te^{\varepsilon}.$$

- For $p = 0, \nu, 2\nu, \ldots$ (upto 1), we do the following :
  - Draw $s = 20m/(\varepsilon^2\delta)$ (pairwise nearly independent) samples $X_1, X_2, \ldots X_s$ (nearly) according to the density

$$e^{-px^TAx-(1-p)|x|^2}/f(p).$$

  - Let

$$T_p = \frac{1}{s}\sum_{i=1}^{s} e^{-\nu|X_i^TAX_i|+\nu|X_i|^2}.$$

- Return the product of all the T_p 's as T.

Remark : The samples drawn for a particular value of p will be totally independent of samples drawn for other values of p.

To give the essential idea of the analysis of the algorithm, we will assume here that

(i) $E(T_p) = \frac{f(p+\nu)}{f(p)}$ and

(ii) $\mathrm{Var}(T_p) \leq \varepsilon^2\delta/(4m)$.

Both these are valid if the s samples used for finding T_p are pairwise completely independent (i.e., 0-independent) and each exactly according to the desired density. This is not quite true, but the errors cause only minor technical problems and the final paper will contain the full analysis. Under the assumptions, we have first that

$$\mathrm{Prob}(T_p \leq E(T_p)/\sqrt{2}) \leq \delta/(16m).$$

So with probability at least $1 - (\delta/16)$, we have that for every $p : T_p \geq E(T_p)/\sqrt{2}$. This technical fact is only used in the second case below.

Now, consider the random variable

$$t = \log \frac{T}{\prod_p (E(T_p))} = \sum_p \log \frac{T_p}{E(T_p)}.$$

First consider the case when $t \geq 0$. Since $\log x \leq x - 1$ for all real x, we have

$$t = \sum_p \log \frac{T_p}{E(T_p)} \leq \sum_p \frac{T_p - E(T_p)}{E(T_p)}.$$

The random variables $\frac{T_p - E(T_p)}{E(T_p)}$; $p = 0, \nu, \ldots$ are independent, so the variance of the sum above is just the sum of the variances. Thus we have by Markov inequality that

$$\text{Prob}(t \geq \varepsilon) \leq \delta/4.$$

The second case when $t < 0$ is a little more complicated. In this case, we use the inequality $\log x \geq (x-1) - (x-1)^2$ which is valid for all $x \geq 1/\sqrt{2}$. We will defer the full argument to the final paper.

We saw that we can produce the first sample for each p in $O^*(\Lambda^2 n^3)$ steps and subsequent samples in $O^*(\Lambda^2 n^2)$ steps per sample. So the total number of steps needed to estimate the integral to relative error ϵ is at most

$$O^*(\Lambda^4 n^4).$$

7 Open Problems

Several interesting open problems remain. The first will be to improve the number of steps needed. In this connection, it should be possible to use the results of [13], especially the notion of proper moves. A proper move of the Metropolis random walk is what the name indicates - it is a move where we move from x to $y \neq x$. If the local conductance is low at a point, we have to make many moves before we make a proper move. But in certain situations, like sampling from the orthant, it may be easy to run a different chain whose transition probability density function is :

$$P_{xy} = \frac{1}{\int_{\mathbf{R}_+^n \cap (x + \delta B)} F} \min\left(1, \frac{F(y)}{F(x)}\right).$$

In other words, we may be able to make proper moves with the correct probabilities without having to "try" out (possibly) improper moves. In [13], better convergence bounds are proved in such cases and it will be interesting to exploit this. There is a hope that for such an improvement for the case of the convex set containing the unit ball (as in section 4) may help improve the volume algorithm.

Our arguments do not carry over for the case when the orthant is shifted - i.e., when the region is of the form : $\{x : x_i \geq p_i\}$ where not all the p_i are zero. The main difficulty is the cumbersome penalty function and the argument that the positive orthant has sizable (here at least 1/4) measure. Perhaps this can be overcome and a better penalty function or a better argument can be found.

It is also possible that these methods here extend to other log-concave densities.

Acknowledgements : We thank Persi Diaconis for bringing the problem to our attention. Thanks to Satish Iyengar for helpful discussions.

References

[1] D. Aldous, "On the Markov chain simulation method for uniform combinatorial distributions and simulated annealing," *Probability in the Engineering and Informational Sciences 1*, pp. 33-46, 1987.

[2] . Aldous, "Random Walks on Finite Groups and Rapidly Mixing Markov Chains", *Séminaire de Probabilités XVII*, Springer Lecture Notes in Mathematics 986, pp. 243-297, 1981/82. Times",

[3] D. Applegate and R. Kannan, "Sampling and integration of near log-concave functions", *Proceedings of the 23rd ACM Symposium on Theory of Computing*, IEEE, 1991.

[4] A. Z. Broder, "How hard is it to marry at random? (On the approximation of the permanent)", *Proceedings of the 18th ACM Symposium on Theory of Computing*, pp. 50-88, 1986.

[5] P. Diaconis and Saloff-Coste, "What do we know about Metropolis Algorithm", *Symposium on the Theory of Computing*, ACM (1995).

[6] M. Dyer and A. Frieze, "Computing the volume of convex bodies: a case where randomness provably helps", *Probabilistic Combinatorics and Its Applications* **44**, 1991.

[7] M. Dyer, A. Frieze and R. Kannan, "A random polynomial time algorithm for approximating the volume of convex bodies", *Journal of the ACM* **38**, 1991.

Tables",

[8] M. Evans, and T. Swartz, "Monte Carlo Computation of Some Multivariate Normal Probabilities", *J. Statist. Comput. Simul.*, Vol. 30, pp. 117-128, 1988.

[9] A. Frieze, R. Kannan and N. Polson, "Sampling from Log-concave Distributions", *Annals of Applied Probability*, Vol. 4, No. 3, (1994) pp 812-837.

[10] S. Iyengar, Evaluation of Normal Probabilities of Symmetric Regions, *SIAM J. Sci. Stat. Comput.*, Vol.9, No.3, May 1988.

[11] M. Jerrum and A. Sinclair, "Approximating the permanent", *SIAM J. Comp.* **18**, 1989.

[12] R. Kannan, Markov Chains and Polynomial Time Algorithms, *Proceedings of the 35th Annual Symposium on Foundations of Computer Science*, pp.656-671, 1994.

[13] R. Kannan, L. Lovász and M. Simonovits, *Random walks and an* $O^*(n^5)$ *volume algorithm for convex sets* Preprint (1996).

[14] L. Lovász and M. Simonovits, "Random walks in a convex body and an improved volume algorithm", *Random Structures and Algorithm* **4**, 1993.

[15] G. Martynov, Evaluation of the Normal Distribution Function, *J. Soviet Math.*, 77 (1980), pp. 1857-1875.

[16] P. Moran, "Calculation of Multivariate Normal Probabilities—Another Special Case", *Austral. J. Statist.*, 27 (1), pp. 60-67, 1985.