DRAFT WORKING PAPER

How to Use Markov Chain Monte Carlo to Do Difficult Monte-Carlo Integrals (Including Those for Normalizing Constants)

William H. Press

(last revised 8/9/2004)

1. Background

The task we set ourselves is simply that of Monte Carlo integration, namely to estimate

$$D_1 \equiv \frac{1}{V} \int F \, dV \tag{1}$$

where V is a given volume, generally in a high dimensional space. (The notation D_1 will be explained below.) We are interested in the hard case where F has, effectively, very small support in V, so that uniform sampling of V *almost never* hits a point where F contributes to the integral.

To tackle this problem, we want to use the machinery of Markov Chain Monte Carlo (MCMC). MCMC provides a (reasonably) efficient way to sample a multidimensional distribution P^* (the star denoting an unnormalized distribution), *even when* it has very small support (as did our F above). That is, MCMC is able to estimate efficiently by the Monte Carlo formula

$$\int F(P \, dV) \approx \langle F \rangle_P \pm \left(\langle F^2 \rangle_P - \langle F \rangle_P^2 \right)^{1/2} / \sqrt{N} \tag{2}$$

where $<>_P$ denotes MCMC sampling over the distribution P, and N is the equivalent number of independent samples. (Notation: We will write $\approx$ to mean equality in the limit of infinite sampling, and ± to indicate the standard deviation when the sample is finite.) This is fine when the support of F is not much smaller than that of P, and is the usual domain of applicability for MCMC. If the contrary is the case, then the MCMC is very inefficient.

The magic of MCMC is that it enables the sampling over P even when we can only compute an unnormalized, but proportional, P^*, where the two are related by

$$P = P^* / \int (P^* \, dV) \tag{3}$$

With MCMC, we never actually need to know the integral, or normalizing constant, in the denominator of equation (3). Indeed it would be hard to know it even if we wanted to, since this is just the same hard problem as equation (1).

MCMC seems superficially related to importance sampling. We might be tempted to write

$$\int F \, dV = \int (F/P)(P \, dV) \approx \langle F/P \rangle_P \pm \left[\langle (F/P)^2 \rangle_P - \langle F/P \rangle_P^2 \right]^{1/2} / \sqrt{N} \tag{4}$$

and use MCMC to perform the sampling over P. As usual in importance sampling, we would choose P so as to resemble F in a way that reduces the variance of the integrand and thus improve convergence. The problem is that the approximate equality in (4) holds only if P is already normalized. If it isn't, then, we need the normalization. Since P resembles F in its support, computing the normalization of P is just as hard as the original problem.

The extreme form of this misuse of MCMC would be to use F (assuming for the moment an F that is everywhere nonnegative) as its own template, i.e., $F = P$, so that the variance of the integrand in (4) is reduced to zero. Then we can either estimate the normalizing constant by sampling over F, namely

$$\frac{1}{V}\int F\,dV = \frac{\int (F/F)(F\,dV)}{\int (1/F)(F\,dV)} \approx 1\Big/\left\langle \frac{1}{F}\right\rangle_F \tag{5}$$

(a really bad idea since $1/F \sim \infty$ almost everywhere!) or else pull it out for evaluation in some other way, which gives the completely circular,

$$\frac{1}{V}\int F\,dV = 1 \times \left[\frac{1}{V}\int F\,dV\right] \tag{6}$$

2. Sampling a Function over Powers of Itself

The above excursion into importance sampling led to a (bad) example, equation (5) where the desired integral D_1 was estimated by averaging one power of F over the sampling by another power, that is,

$$D_1 \approx \left[\left\langle F^{-1}\right\rangle_F\right]^{-1} \tag{7}$$

It was the negative powers that led us astray. We will now show how to use only positive powers F^β, with $0 \le \beta \le 1$, to advantage.

Define

$$D_\beta \equiv \frac{1}{V}\int F^\beta dV \tag{8}$$

Note that D_1 is our desired result, while $D_0 = 1$. For $0 \le \alpha \le \beta \le 1$ we can write

$$\frac{D_\beta}{D_{\beta-\alpha}} = \frac{\int F^\beta dV}{\int F^{\beta-\alpha} dV} = \frac{\int F^\alpha (F^{\beta-\alpha} dV)}{\int (F^{\beta-\alpha} dV)} \approx \langle F^\alpha\rangle_{F^{\beta-\alpha}} \tag{9}$$

So, any ratio of D's can be estimated by an MCMC sampling over a power of F.

Using equation (9) we can write a sequence like

$$\begin{aligned} D_1 &\approx \left\langle F^{\frac{1}{2}}\right\rangle_{\frac{1}{2}} D_{\frac{1}{2}} \\ &\approx \left\langle F^{\frac{1}{2}}\right\rangle_{\frac{1}{2}} \left\langle F^{\frac{1}{4}}\right\rangle_{\frac{1}{4}} D_{\frac{1}{4}} \\ &\approx \left\langle F^{\frac{1}{2}}\right\rangle_{\frac{1}{2}} \left\langle F^{\frac{1}{4}}\right\rangle_{\frac{1}{4}} \left\langle F^{\frac{1}{8}}\right\rangle_{\frac{1}{8}} \cdots \left\langle F^{\frac{1}{1024}}\right\rangle_{\frac{1}{1024}} D_{\frac{1}{1024}} \end{aligned} \tag{10}$$

(where we have simplified the notation in an obvious way). Eventually we reach a point where the power ϵ is so small that $F^\epsilon \approx 1$ everywhere in V, so that we can terminate the sequence using (e.g.)

$$\left\langle F^{\frac{1}{1024}} \right\rangle_{\frac{1}{1024}} D_{\frac{1}{1024}} \approx 1 \times V = V \tag{11}$$

Actually, it is not necessary to go all the way to the limit of equation (11): Eventually F^ϵ will have reasonable support in V so that D_ϵ can be estimated by the uniform sampling over V implied by its defining equation (8).

Let us answer in passing the objection that F might be so small (or zero) outside of its effective support that no practical power of it ever approaches unity. This would be true far out on a Gaussian tail, for example. However, in such a case, we can from the very start replace F by $F + \delta$, with some very small δ, and then subtract $V\delta$ from the final estimate of the integral D_1. As long as $\delta \ll F$ (for values of F returned by the MCMC), it will not seriously distort the MCMC sampling, and no additional error is introduced.

Also in passing, let us note that if F is not everywhere positive, the first sequence step can be replaced by

$$\frac{D_1}{D_{1-\alpha}} = \frac{\int \mathrm{sgn}(F)|F|^\alpha(|F|^{1-\alpha}dV)}{\int |F|^{1-\alpha}dV} \approx \langle \mathrm{sgn}(F)F^\alpha \rangle_{1-\alpha} \tag{12}$$

and then replace F by $|F|$ in all subsequent steps.

Let's summarize what we have so far: We have replaced a single Monte Carlo sampling (of F uniformly over V, say), by a sequence of many MCMC samples of different powers of F. While doing many separate MCMC's instead of one Monte Carlo integral sounds like a bad idea, the MCMC's are only logarithmically many; and they are all practical to do, since the integrand has roughly the same support as the MCMC sampling. So the whole process is harder than a single MCMC, but only by a logarithmic factor. By contrast, getting the desired integral of F by uniform sampling over V is entirely impractical, since the support of F in V is assumed to be truly infinitesimal.

In the next section we find sequences that are more optimized than equation (10), and estimate the workload of achieving convergence.

3. Optimal Sequences

There was nothing magical about the powers of 2 in the exponents in equation (10). The goal is just to drive the remaining index ϵ in D_ϵ to zero. Instead of equation (10), one could use the sequence

$$\begin{aligned} D_1 &\approx \left\langle F^{0.8} \right\rangle_{0.2} D_{0.2} \\ &\approx \left\langle F^{0.8} \right\rangle_{0.2} \left\langle F^{0.16} \right\rangle_{0.04} D_{0.04} \\ &\approx \left\langle F^{0.8} \right\rangle_{0.2} \left\langle F^{0.2} \right\rangle_{0.04} \left\langle F^{0.032} \right\rangle_{0.008} \cdots \end{aligned} \tag{13}$$

where the index is reduced a factor 5 in each step. And what about the sequence

$$\begin{aligned} D_1 &\approx \left\langle F^{0.99} \right\rangle_{0.01} D_{0.01} \\ &\approx \left\langle F^{0.99} \right\rangle_{0.01} \left\langle F^{0.0099} \right\rangle_{0.0001} \cdots \end{aligned} \tag{14}$$

where we would apparently gain a factor of 100 with each step? Equations (10), (13), and (14) are equally good consequences of the basic equation (9).

We will now show that, parameterized by the factor that one gains at each step (2, 5, and 100 in the above examples), there is actually an optimal factor that minimizes the variance per sampling workload.

The variance of one of the sampling factors is given by

$$\mathrm{Var}\left[\left\langle F^R \right\rangle_\rho\right] = \frac{1}{N}\left(\left\langle F^{2R}\right\rangle_\rho - \left[\left\langle F^R\right\rangle_\rho\right]^2\right) \tag{14}$$

so the fractional variance (variance divided by square of value) is

$$\begin{aligned}\frac{\mathrm{Var}\left[\left\langle F^R \right\rangle_\rho\right]}{\left[\left\langle F^R\right\rangle_\rho\right]^2} &= \frac{1}{N}\left(\frac{\left\langle F^{2R}\right\rangle_\rho}{\left[\left\langle F^R\right\rangle_\rho\right]^2} - 1\right)\\ &= \frac{1}{N}\left(\frac{\int F^{2R}(F^\rho\, dV)/D_\rho}{[\int F^R(F^\rho\, dV)/D_\rho]^2} - 1\right)\\ &= \frac{1}{N}\left(\frac{\int F^\rho\, dV \int F^{2R}(F^\rho\, dV)}{[\int F^R(F^\rho\, dV)]^2} - 1\right)\end{aligned} \tag{15}$$

We can evaluate this explicitly, as a function of R and ρ, in the case of multivariate Gaussian in M dimensions. Without loss of generality taking the Gaussian as having unit shape, we have

$$I_\alpha \equiv \int \exp[-(x_1^2 + x_2^2 + \cdots + x_M^2)/2]^\alpha \, dx_1 dx_2 \cdots dx_M = \left(\frac{1}{\alpha}\right)^{M/2} I_1 \tag{16}$$

so that

$$\begin{aligned}\frac{\mathrm{Var}\left[\left\langle F^R \right\rangle_\rho\right]}{\left[\left\langle F^R\right\rangle_\rho\right]^2} &= \frac{1}{N}\left(\left[\frac{(R+\rho)^2}{\rho(2R+\rho)}\right]^{M/2} - 1\right)\\ &= \frac{1}{N}\left(\left[\frac{(1+\rho/R)^2}{(1+\rho/R)^2 - 1}\right]^{M/2} - 1\right) \equiv \frac{1}{N} V_M(\rho/R)\end{aligned} \tag{17}$$

is a function only of M and the ratio ρ/R.

Equation (17) relates the value of ρ/R to the variance of a single factor in a sequence like equation (10). However, the number of such factors is also a function of ρ/R. Suppose p factors are just sufficient to make the residual factor be D_ϵ for some desired ϵ. (That is, suppose we know independently that we can evaluate D_ϵ efficiently by direct sampling.) Then, from equation (9), evident in the sample cases in equations (10), (13), and (14), we have

$$\left(\frac{R}{\rho} + 1\right)^{-p} \approx \epsilon \tag{18}$$

and

$$p \approx \frac{(-\ln \epsilon)}{\ln\left(\frac{R}{\rho}+1\right)} \tag{19}$$

Next, suppose we want a fractional accuracy δ in overall estimation of the integral D_1. Let N_s be the *total* number of samples taken, with therefore $N = N_s/p$ samples allocated to each factor. Then the fractional accuracy of the product of the p factors is p times that for a single sample, or

$$\delta = p\left(\frac{p}{N}V_M(\rho/R)\right)^{1/2} \tag{20}$$

Equations (19) and (20) give the estimate for N_s, the total required number of samples, in terms of ϵ, δ, and ρ/R,

$$N_s = \frac{(-\ln \epsilon)^3}{\delta^2}\left(\frac{V_M(\rho/R)}{[\ln(R/\rho+1)]^3}\right) \equiv \frac{(-\ln \epsilon)^3}{\delta^2}\widetilde{V}_M(\rho/R) \tag{21}$$

Thus, the optimal value of ρ/R is that which minimizes $\widetilde{V}_M$ as defined in equation (21).

Figures 1-4 plot $\widetilde{V}_M$ as a function of ρ/R for $M = 2, 5, 10, 20$. Indeed, there is a minimum in all cases. Figure 5 plots the value of ρ/R at the minimum for M, the number of dimensions, in the range of 1 to 20. The value of ρ/R increases approximately linearly with M in this range, and is reasonably well fit by

$$(\rho/R)_{min} \approx 0.1(M-1) \tag{22}$$

The value of $\widetilde{V}_M$ at the minimum is plotted in Figure 6. It increases roughly quadratically for M in this range, and can be fit by

$$(\widetilde{V}_M)_{min} \approx -0.73 + 0.34M + 0.070M^2 \tag{23}$$

We can now see that the original sequence of samplings in equation (10), with $\rho/R = 1$ would be near optimal for $M = 10$, while the sequence in equation (13), with $\rho/R = 0.25$, would be near optimal for two- or three-dimensional problems.

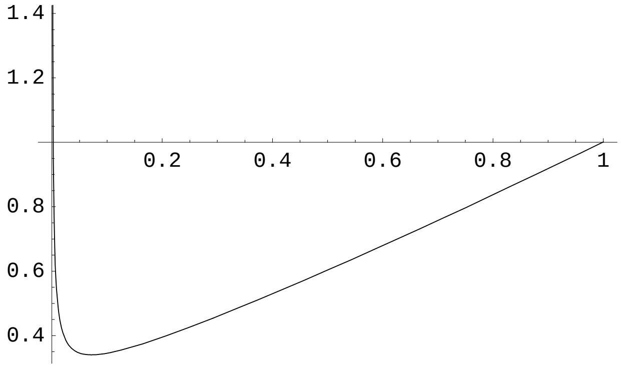
For high-dimensional problems, $M > 10$, the optimal ρ/R is perhaps counter-intuitively greater than 1, implying even more separate MCMC samplings than equation (10). The explanation is that although there are more samplings (one for each factor in the sequence), the variance of each is kept small enough to make this strategy pay – evidently.

```
f[x_, M_] := (((1 + x)^2 / ((1 + x)^2 - 1))^(M / 2) - 1) / (Log[1 / x + 1])^3
```

```
N[f[1 / 2, 2]]
```

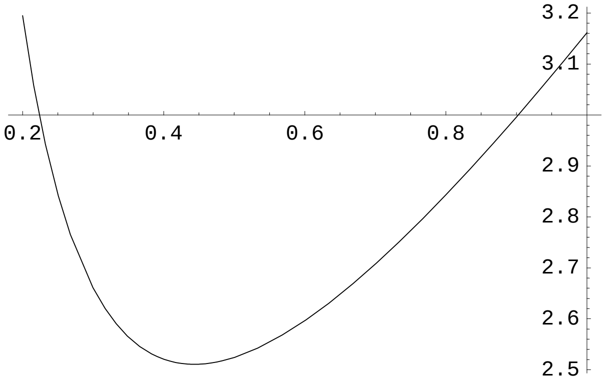
```
0.603332
```

```
Plot[f[x, 2], {x, 0, 1}]
```



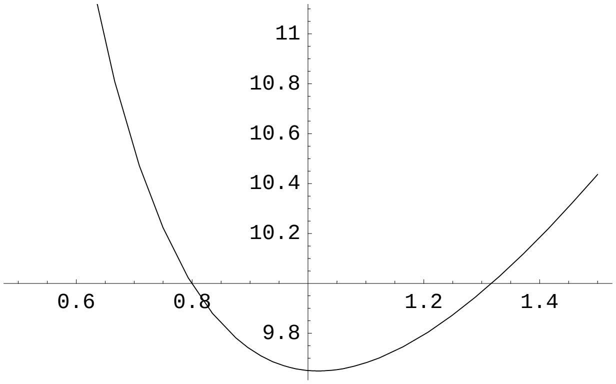
```
- Graphics -
```

```
Plot[f[x, 5], {x, 0.2, 1}]
```



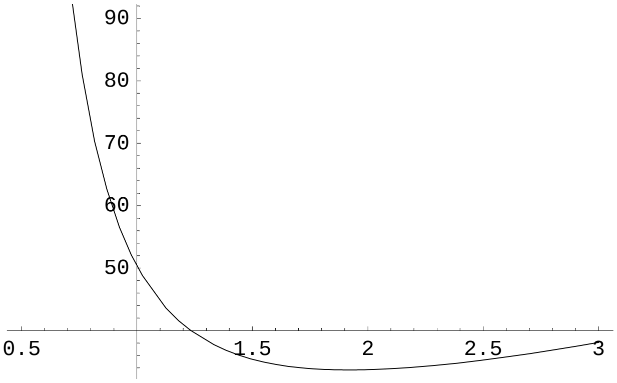
```
- Graphics -
```

```
Plot[f[x, 10], {x, 0.5, 1.5}]
```

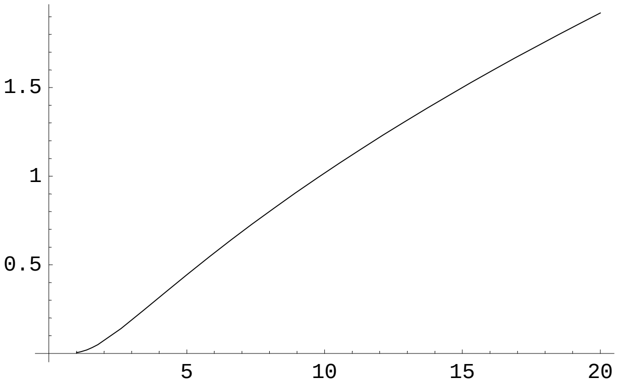


```
- Graphics -
```

```
Plot[f[x, 20], {x, 0.5, 3}]
```

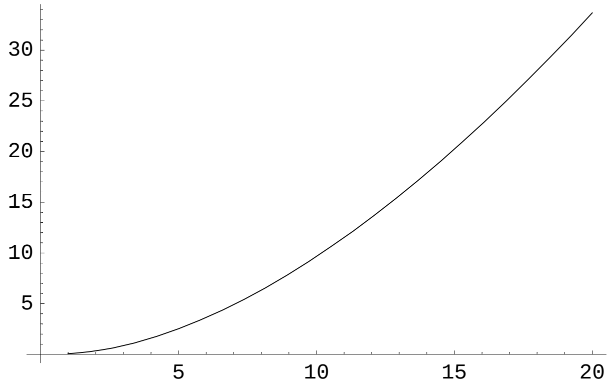


```
- Graphics -

MyXmin[M_] := FindMinimum[f[x, M], {x, 0.1, 0.001, 3}][[2]][[1]][[2]]



xmintrue = Plot[MyXmin[M], {M, 1, 20}]
```



```
- Graphics -

MyY[M_] := FindMinimum[f[x, M], {x, 0.1, 0.001, 3}][[1]]
```

```
ytrue = Plot[MyY[M], {M, 1, 20}]
```



```
- Graphics -
```

```
Off[FindMinimum::lstol]
```

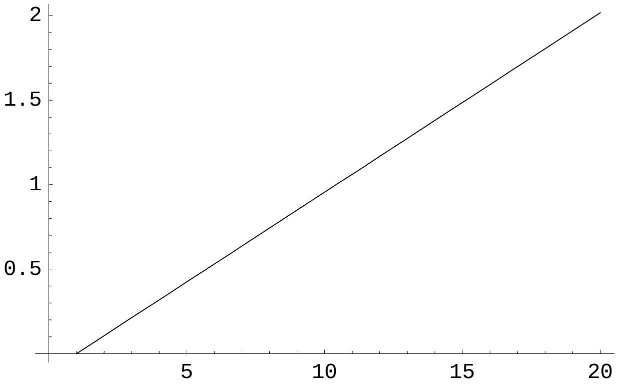
```
Fit[Table[MyXmin[M], {M, 1, 20}], {(M - 1)}, M]
```

$0.106135\,(-1+\mathrm{M})$

```
Fit[Table[MyY[M], {M, 1, 20}], {1, M, M^2}, M]
```

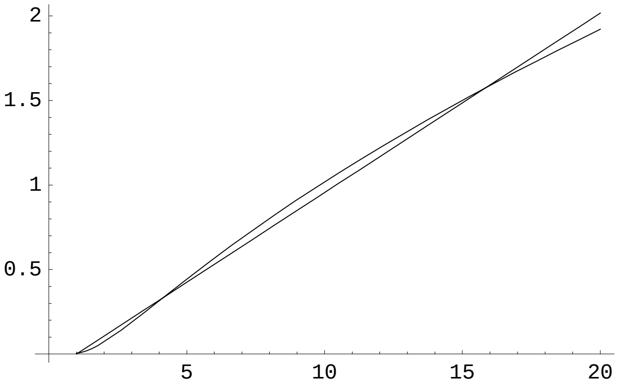
$-0.731859+0.342289\,\mathrm{M}+0.0695944\,\mathrm{M}^2$

```
xminfit = Plot[0.10613547640041451 (-1 + M), {M, 1, 20}]
```



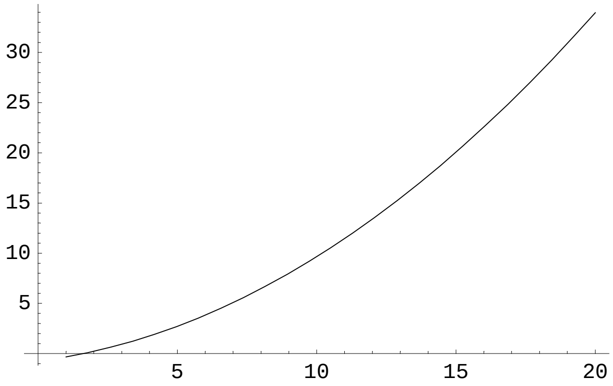
```
- Graphics -
```

```
Show[xmintrue, xminfit]
```



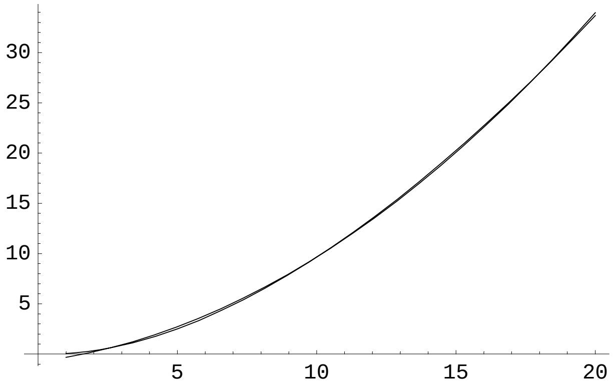
- Graphics -

```
yfit =
 Plot[-0.7318592286584759 + 0.342289149948013 M + 0.06959442495386041 M^2, {M, 1, 20}]
```



- Graphics -

```
Show[ytrue, yfit]
```



- Graphics -