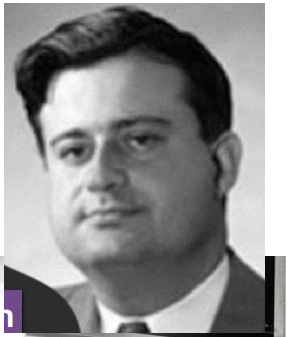
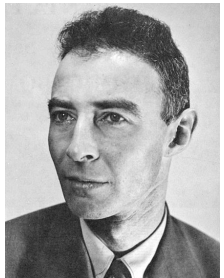
(Metropolis)-Hastings

Fabian Bull


15




X 169



Different methods from Metropolis

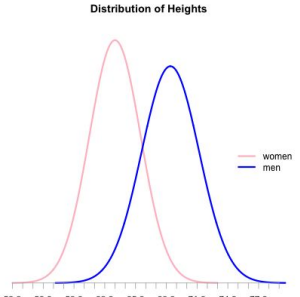
Difficult and expensive to sample multidimen (2021 very fast!)

General methods:

(i) if possible factorize P(x1,x2,...,xn) = P(x1) P(x2|x1) P(x3|x2,x1)....

(ii) Importance sampling, but it does not work that good in high dimen:
q(x) or p(x) can yield very low probability

(iii) Simulate from an simpler dist q, but if q has low prob in regions of interest then it will take long time to get reliable estimates.



Markov chain

Markov property $P(x^t \mid x^{(t-1)}, \ldots, x^0) = P(x^t \mid x^{(t-1)})$

Irreducible (can go from any state to any other state in a finite number of steps)

Aperiodic (it does not continue in a cyclic pattern, RNG has to cycle...)

Monte Carlo

The first thoughts and attempts I made to practice [the Monte Carlo Method] were suggested by a question which occurred to me in 1946 as I was playing solitaires. The question was what are the chances that a solitaire laid out with 52 cards will come out successfully? After spending a lot of time trying to estimate them by pure combinatorial calculations, I wondered whether a more practical method than "abstract thinking" might not be to lay it out say one hundred times and simply observe and count the number of successful plays

Stanislaw Ulam

Metropolis Method

(a) We dont need to know the normalization factor
$P(x) = c*f(x)$ ---> $P(x')/P(x) = c*f(x')/c*f(x) = f(x')/f(x)$

(b) The sequence of a Markov chain are correlated and that may affect estimates.

Markov chain Monte Carlo

Simulate the N samples from the Markov chain x^1,...,x^N

Monte carlo estimate of f on p

$$\hat{I} = \sum_{t=1}^{N} f\{X(t)\}/N.$$

But be careful of the variance because there may exist correlations between samples

$$\mathrm{var}\,(\bar{Y}) = \frac{\sigma^2}{N} \sum_{j=-N+1}^{N-1} \left(1 - \frac{|j|}{N}\right) \rho_j$$

Hastings assures us that there exist a chain=target if:

Choose $\mathbf{P}$ so that $\boldsymbol{\pi}$ is its unique stationary distribution $\boldsymbol{\pi = \pi P}$.

So the chain converges to the target. Check reversibility condition

$$\pi_i p_{ij} = \pi_j p_{ji}.$$

Assume $p_{ij} = q_{ij}\alpha_{ij}$ where $\mathbf{Q} = \{q_{ij}\}$ is the transition matrix $\alpha_{ij} = \dfrac{s_{ij}}{1 + \dfrac{\pi_i}{\pi_j}\dfrac{q_{ij}}{q_{ji}}}$,

Since sij is symmetric

$$\pi_i p_{ij} = \frac{(\pi_i q_{ij}) s_{ij}}{1 + \frac{\pi_i q_{ij}}{\pi_j q_{ji}}}$$

Readily verified...

$$= \frac{(\pi_i q_{ij}) s_{ij}}{\frac{\pi_j q_{ji} \pi_i q_{ij}}{(\pi_j q_{ji})}} = \frac{(\pi_j q_{ji}) s_{ij}}{\frac{\pi_j q_{ji} \pi_i q_{ij}}{(\pi_i q_{ij})}}$$

$$= \frac{(\pi_j q_{ji}) s_{ij}}{1 + \frac{\pi_j q_{ji}}{\pi_i q_{ij}}} = \pi_j p_{ji}$$

Metropolis-Hastings

(i) assume that X(t) = i and select at state j using the proposal distribution qi

(ii) take X(t+1) = j with prob α_{ij} and X(t+1) =i with $1-\alpha_{ij}$.

Where $$\alpha_{ij} = \frac{s_{ij}}{1+\frac{\pi_i q_{ij}}{\pi_j q_{ji}}},$$

More generally, we may choose

$$s_{ij} = g[\min\{(\pi_i q_{ij})/(\pi_j q_{ji}), (\pi_j q_{ji})/(\pi_i q_{ij})\}],$$

$g(x)$ is chosen so that $0 \leqslant g(x) \leqslant 1+x$ for $0 \leqslant x \leqslant 1$,

Metropolis-Hastings

More generally, we may choose

$$s_{ij} = g[\min\{(\pi_i q_{ij})/(\pi_j q_{ji}),\ (\pi_j q_{ji})/(\pi_i q_{ij})\}],$$

$g(x)$ is chosen so that $0 \leqslant g(x) \leqslant 1+x$ for $0 \leqslant x \leqslant 1$,

$$\alpha_{ij} = \frac{s_{ij}}{1 + \frac{\pi_i q_{ij}}{\pi_j q_{ji}}},$$

Shows that Metropolis has g(x) = 1+x

if $\frac{\pi_i q_{ij}}{\pi_j q_{ji}} \leq \frac{\pi_j q_{ji}}{\pi_i q_{ij}}$

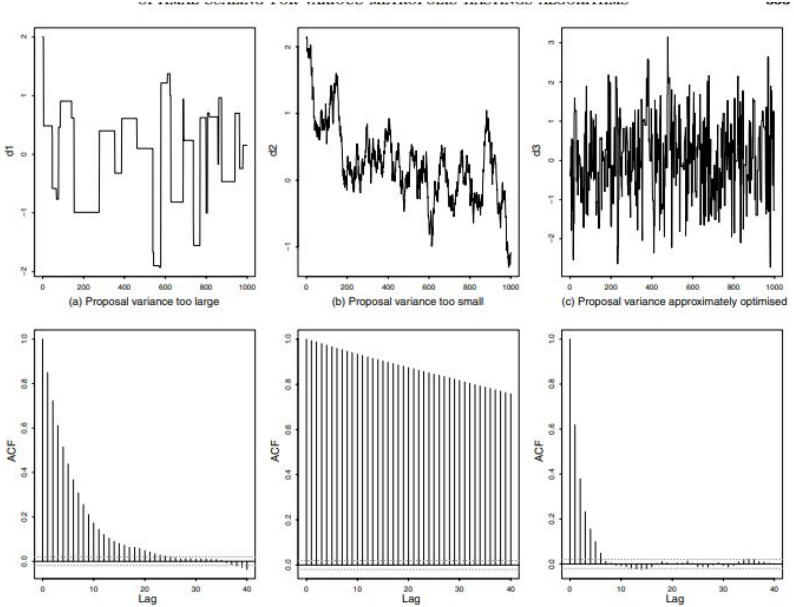
$$\alpha_{ij} = \frac{1 + \frac{\pi_i q_{ij}}{\pi_j q_{ji}}}{1 + \frac{\pi_i q_{ij}}{\pi_j q_{ji}}} = 1$$

if $\frac{\pi_i q_{ij}}{\pi_j q_{ji}} > \frac{\pi_j q_{ji}}{\pi_i q_{ij}}$

$$\alpha_{ij} = \frac{1 + \frac{\pi_j q_{ji}}{\pi_i q_{ij}}}{1 + \frac{\pi_i q_{ij}}{\pi_j q_{ji}}} = \frac{\frac{\pi_i q_{ij} + \pi_j q_{ji}}{\pi_i q_{ij}}}{\frac{\pi_j q_{ji} + \pi_i q_{ij}}{\pi_j q_{ji}}} = \frac{\pi_j q_{ji}}{\pi_i q_{ij}}$$

Rejection rate the proportion of times t for which X(t+1)=X(t)

if Rejection rate = 0 or 1 is a bad choice of Q (1970)



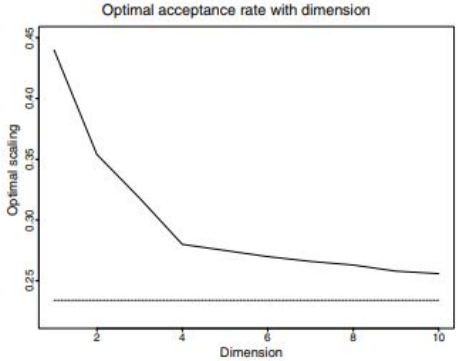


FIG. 4. *Optimal scaling as a function of acceptance rate, using the minimum autocorrelation criterion, as dimension increases for the case of Gaussian target densities. This analysis comes from a simulation study on standard Gaussian target densities.*

as d->infty optimal AR = 0.234

Gareth O. Roberts. Jeffrey S. Rosenthal. "Optimal scaling for various Metropolis-Hastings algorithms." Statist. Sci. 16 (4) 351 - 367, November 2001. https://doi.org/10.1214/ss/1015346320

Roberts, G. O., Gelman, A. and Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. Ann. Appl. Probab. 7 110–120.

Multidimensional distributions

(3)Change one coordinate along axis 1,2,..d in succession
Ehrman, Fosdick & Handscomb (1960)

$$\mathbf{P} = \mathbf{P}_1 \dots \mathbf{P}_d. \qquad \mathbf{P}_k \text{ is constructed so that } \boldsymbol{\pi}\mathbf{P}_k = \boldsymbol{\pi},$$

$$\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}\mathbf{P}_1 \dots \mathbf{P}_d = \boldsymbol{\pi}\mathbf{P}_2 \dots \mathbf{P}_d = \dots = \boldsymbol{\pi}.$$

Observe the chain only after the coordinate has been updated along all axes 1,..k,. d

Then $\boldsymbol{\pi}$ will be the stationary distribution of the chain (0, d, 2d, …, Sd)

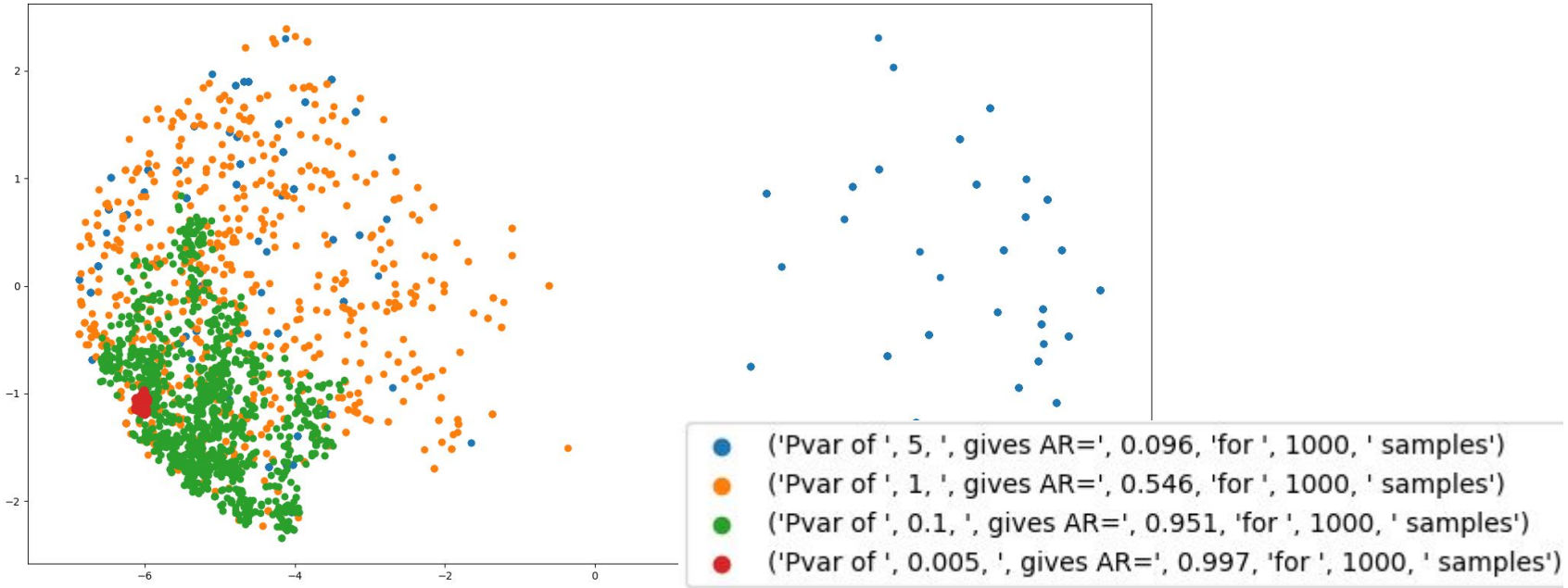
Not sure if G. O. Roberts considered (1) change all coordinates or (3)

Short two-dimensional example.

Target is a difficult function only up to a constant.

Smaller density closer to 0

Proposal is $x_1^*, x_2^* \sim N((x_1, x_2), Pvar)$

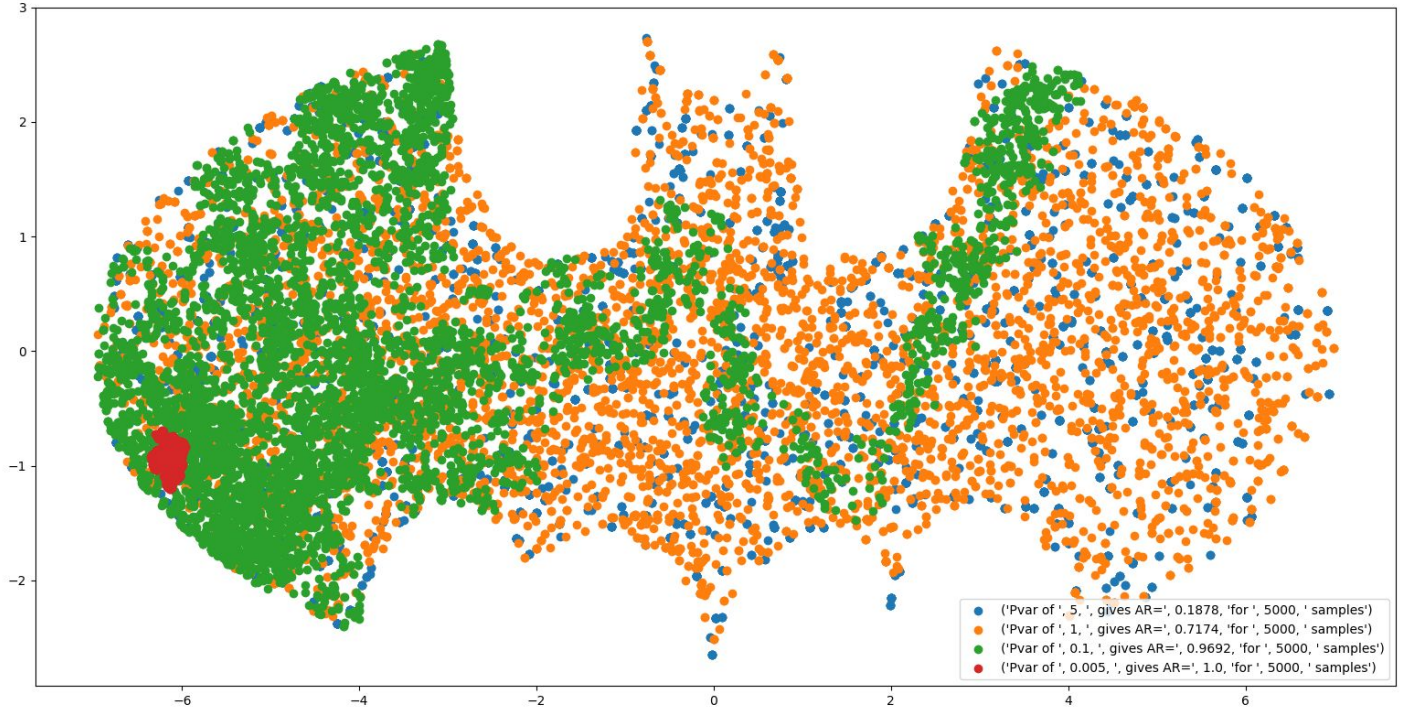


Importance sampling

use a new target pi' that gives more probability of the area closer to 0

For Inference correct for the new pi' with

$$\hat{I} = \frac{\sum_{t=1}^{N} [f\{X(t)\}\pi_{X(t)}/\pi'_{X(t)}]/N}{\sum_{t=1}^{N} \{\pi_{X(t)}/\pi'_{X(t)}\}/N}.$$



Error in the methods

(i) RNG

(ii) The complex estimated dist (multidimensional and multimodal)

(iii) Computational errors in the estimate

(iv) Computational limits of the computer

(v) Error from a small sample size

Sol:
Choose a Q that can sample a point in the sample space (not limited)

Smart choice of initialization (avoid exploring an uninteresting region)
test method on similar problem. Start in region of interest. Compare segments in the chain.

Other things that might be of interest in 2021

Discard the first B samples, called the burn in period

Extend the run length (it is 2021 after all)

Run multiple chains from different starts to see if the traces explore the same area

Autocorrelation/lag

Subsampling