Multicanonical Monte Carlo and Importance Sampling simulation techniques: How are they related?

A short course for PhD Students in Information Engineering

Alberto Bononi

Universita' degli Studi di Parma, Parma, Italy, Jan. 15-26, 2007

OUTLINE of the course

Part I

- Standard Monte Carlo (MC) estimation
- Importance Sampling (IS)
- the Multicanonical idea: Flat Histogram importance sampling (FHIS)
- Berg's standard Multicanonical Monte Carlo algorithm (MMC)
- Berg's Smoothed MMC

Part II

- CDF method
- Von Neumann's Rejection method
- the Markov Chain Monte Carlo (MCMC) method
  - Hastings MCMC
  - Metropolis MCMC
  - Convergence issues
- MMC using MCMC
- Output-warping importance sampling (OWIS)
- On the origin of the name MMC: Berg's physical problem.

FOREWORD

These notes contain my personal synthesis of a vast body of published literature on a new topic in fast computer simulation known as multicanonical monte carlo (MMC). MMC is a truly innovative algorithm that is connected to the well known method of Importance Sampling (IS), but does not need a special knowledge of the physical problem at hand in order to find a good biasing distribution, which is the true limit of IS.

The key tool used by MMC in order to adaptively generate biasing distributions with a desired density is the markov chain monte carlo (MCMC) method. All papers I consulted on MMC, reported in the Bibliography, delve into the machinery of the MCMC method, as if the true heart of the MMC algorithm were the MCMC biasing scheme.

My view is that instead one can explain MMC without the need of MCMC, so that all the attention can be focused on the analytical connections between MMC and IS. Later MCMC enters into the play, but its function within MMC is now clear and the reader can better appreciate the subtleties connected with its use within MMC.

For this work I am indebted to Dr. Armando Vannucci, who introduced me to the MMC algorithm, and his former student Ing. Nicola Rossi, who actively contributed together with Dr. Alessandra Orlandini to many useful applications of the MMC to nonlinear optical communications. N. Rossi also pepared the computer code for the numerical examples for this short course.

I would like to mention that this course came out of my research notes on MMC, and that I soon realized the importance this tool could have for the research of my fellow researchers and their graduate students. That's why I gladly accepted the invitation of Prof. C. Morandi, coordinator of the PhD program in Information Technology at the University of Parma, to extend the short lectures I gave to my small research group to a larger interested audience, notwithstanding my still incomplete understanding of the global topic.

Hence please bear with me if many typos, some imprecisions, and perhaps even some mistakes will be there to plague these handwritten notes. After all, it is from our very mistakes that we all truly learn.

Parma, January 15, 2007 Alberto Bononi

Historical Background

(see e.g. E. Veach, PhD Thesis, Stanford 1997)

▷ Monte Carlo (MC) methods originated at Los Alamos Nat. Lab. soon after World War II. There, the first computer (ENIAC) was built, and scientists were trying to use it for the design of the H-Bomb.

1946: S. Ulam suggested use of RANDOM SAMPLING for simulation.
1947: J. Von Neumann: a first proposal.
1949: N. Metropolis, S. Ulam: first published paper
"The Monte-Carlo method" J. Am. Stat. Assoc., V. 44, p 335-341.

Name, due to Metropolis, connected to famous casino in Monaco.

The Problem: evaluation of Multiple integrals:

$$I = \int_\Gamma f(\underline{x})\, d\underline{x} \qquad \underline{x} \in \mathbb{R}^n$$

for some domain $\Gamma \subset \mathbb{R}^n$

When integrand is not well behaved (discontinuities, peaks, ...), can be convenient to evaluate I as follows.
Consider a prob. density $p(\underline{x})$ on Γ such that $f(\underline{x}) \neq 0 \Rightarrow p(\underline{x}) > 0$.
Then write I as:

$$I = \int_\Gamma \left[\frac{f(\underline{x})}{p(\underline{x})}\right] p(\underline{x})\, d\underline{x} = E\left[\frac{f(\underline{X})}{p(\underline{X})}\right]$$

Recall: for a random Variable (RV) X with PDF $p(x)$

$$E[g(X)] = \int_\Gamma g(x) \cdot p(x)\, dx$$

LAW UNCONSCIOUS STATISTICIAN (LUS)

The random sampling idea is to estimate the value of $I = E\left[\frac{f(\underline{x})}{p(\underline{x})}\right]$ as

$$\hat{I} = \frac{1}{N} \sum_{i=1}^{N} \frac{f(\underline{x}_i)}{p(\underline{x}_i)}$$

(Horvitz-Thompson, 1952 : J. Am. Stat. Assoc., V. 47, p 663-685)

i.e. as the sample mean in a series of N independent trials ($\equiv$ simulations $\equiv$ evaluations of $f(\underline{x})/p(\underline{x})$) in which samples $\underline{X}_i$ are RVs generated from density $p(\underline{x})$.

The 1949 MC proposal used a uniform density $p(\underline{x}) \equiv U(\underline{x})$, hence the name of "random sampling".

With non-uniform $p(\underline{x})$, the technique is now known as "Importance Sampling" (IS).

The choice of a "good" density $p(\underline{x})$ for a desired unknown integral I is the crux of IS: it is more an art than a science. Choice is usually done by trial-and-error.

▷ Flat Histogram (FH) methods

In 1992 B.A. Berg invented the MHC method, first published in

B.A. Berg, T. Neuhaus, Phys. Rev. Lett, V. 68, p. 9-12.

which soon revolutionized the way of doing fast simulations. Berg's papers are hard to read for non-physicists, and the probability theory ideas are hidden by the physical details of their problem.

A similar method is due to Wang and Landau [Phys Rev Lett, v. 86, p. 2050-2053, 2001] which is also based on the flat histogram concept.

The flat histogram methods took a long time to escape from the physics community literature, and only recently the statistics community is starting to study the novel class of algorithms (see e.g. Y.F. Atchade, J.S. Liu "the Wang-Landau Algorithm for MC computation in general state spaces", Technical Report, 2004 www.mathstat.uottawa.ca/~yatch436/gwl.pdf).

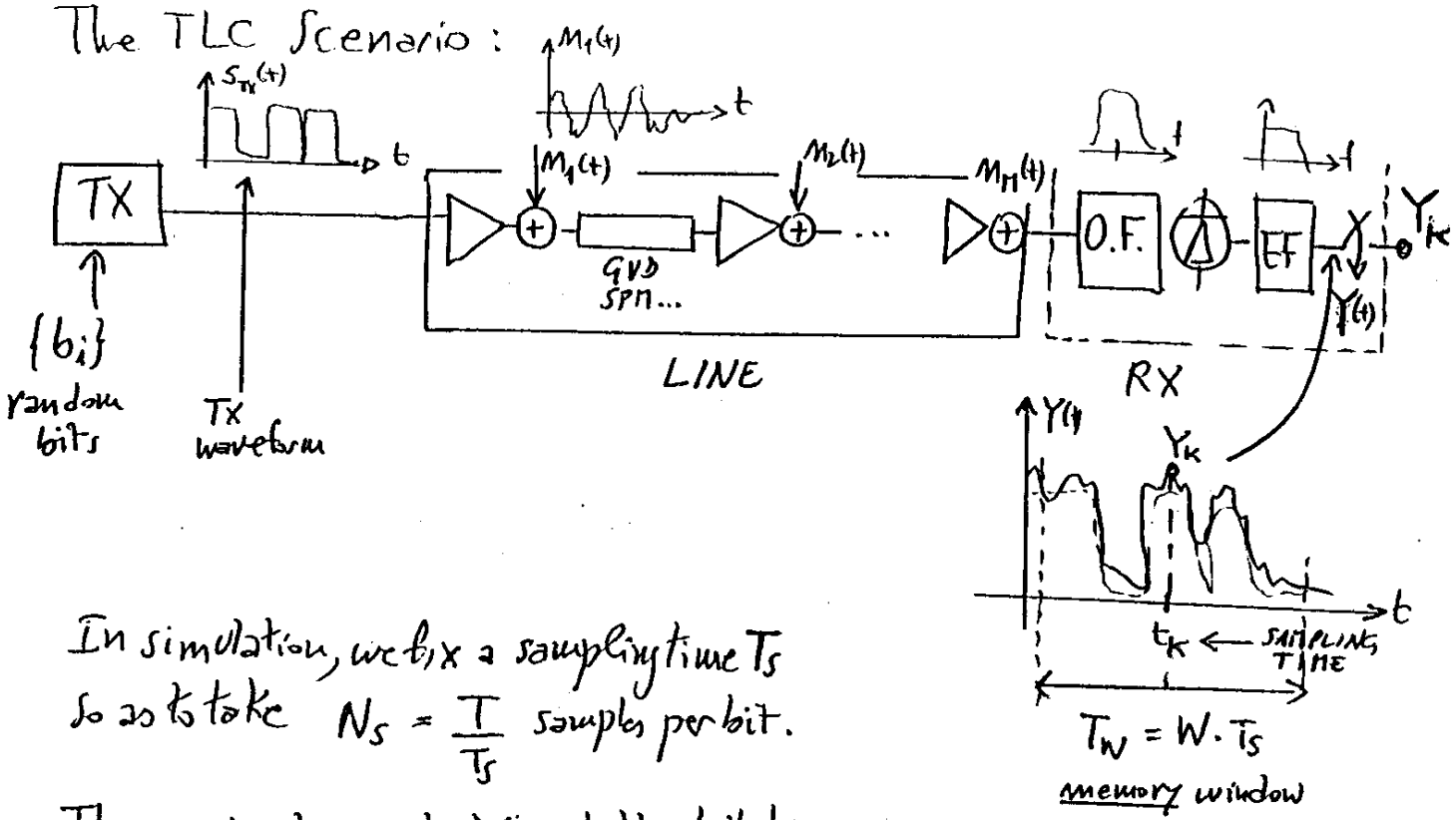
▷ In this Course, I will give my contribution in clarifying the statistical concepts behind MMC.

My first exposure to the MMC idea comes from simulation problems in optical communications, related to polarization mode dispersion (D. Yevick, Photon Technol Lett, v. 14, p. 1512-1514, 2002) (R. Holzlöhner and C.R. Menyuk, Opt. Lett., v. 28, p. 1894-1896, 2003).

▷ In this text I will explain the theory in the context of estimation of the probability of events, and will avoid specific physical applications in order to highlight the general theoretical principles. But first, let me give you my motivation for studying the problem.

Motivation

The TLC scenario:



In simulation, we fix a sampling time T_s so as to take $N_s = \frac{T}{T_s}$ samples per bit.

The received sample Y_k at the bit time of interest, upon which a decision must be made is affected by received waveform time samples

$$Y(t) = s_{RX}(t) + n(t)$$

("DISTORTED" signal; ↑ all noise sources)

$$\left(\begin{array}{ll} Y_k > \theta \Rightarrow & \hat{b}_k = 1 \\ < \theta & \hat{b}_k = 0 \end{array} \right) \quad \uparrow \text{Threshold}$$

within a memory window of W samples. If M = # amplifiers (or noise sources $n_1(t), \ldots, n_M(t)$) then Y_k depends on $W \cdot M$ noise samples $[N_1 \ldots\ldots N_{W \cdot M}]$ and on signal random bits $[b_1, \ldots, b_{\lceil W/N_s \rceil}]$

N $\qquad$ B

(excluding bit of interest, upon which must decide). Hence, if $\underline{X} = [\underline{N}, \underline{B}]$,

then $(Y_k \equiv Y)$ $\qquad \boxed{Y = g(\underline{X})}$

decision variable is a (in general) nonlinear function of all random "noise" samples $\underline{X}$ within system memory.

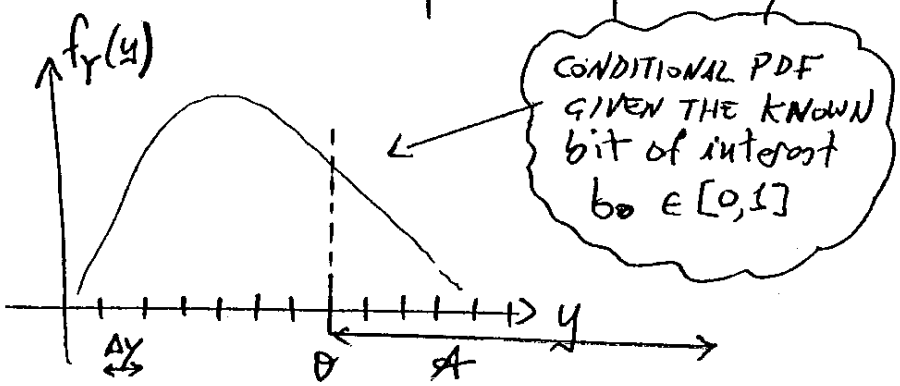
$\underline{X} \in \Gamma$ = state space (or input space)

Motivation

Objective is to run N simulations, i.e. select N samples from state space $(\underline{X}_1, \ldots, \underline{X}_N)$ and then compute

$$\left[Y_1 = g(\underline{X}_1), \ldots, Y_N = g(\underline{X}_N)\right]$$

And from samples $\{Y_i\}$ estimate either the probability density function (PDF) of Y



obtained by "binning" the y values (bin width = Δy) and counting samples in each bin,

Or estimate the probability of a desired "rare" event $\{Y \in A\}$ (e.g. the tail probability, or error probability) by counting samples $\{Y_i\}$ falling over set A.

In the rest of the talk, to simplify math, I will concentrate on the scalar case $\boxed{Y = g(X)}$.
Results will nicely extend to $\underline{X}$ a vector.

Before starting, let me recall a few concepts from basic statistics.

Statistics Refresher

Consider a sequence of RVs $\{X_1, X_2, X_3, \dots\}$ having same mean η and variance σ^2.

Let
$$\begin{cases} \overline{X}_N \triangleq \frac{1}{N} \sum_{m=1}^{N} X_m & \text{sample mean} \\ V_N \triangleq \frac{1}{(N-1)} \sum_{m=1}^{N} \left(X_m - \overline{X}_N\right)^2 & \text{sample variance} \end{cases}$$

It is easy to prove that
$$E[\overline{X}_N] = \eta \quad , \quad E[V_N] = \sigma^2 \quad \text{for all } N$$

For this reason $\overline{X}_N$ and V_N are said to be UNBIASED ESTIMATORS of η, σ^2, respectively.
Also, under fairly mild conditions on correlations among $\{X_i\}$, one can also prove that

$$\begin{cases} \lim_{N\to\infty} \mathrm{Var}[\overline{X}_N] = 0 \\ \lim_{N\to\infty} \mathrm{Var}[V_N] = 0 \end{cases}$$
and these estimators are thus said to be CONSISTENT

Ex if $\{X_i\}$ IID (indep., identically distributed) then
$$\mathrm{Var}[\overline{X}_N] = \frac{1}{N^2} \sum_{n=1}^{N} \mathrm{Var}[X_n] = \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{N} \xrightarrow[N\to\infty]{} 0$$

From their Definition, it would seem that storage of all samples X_n is needed to compute $\bar{X}_N$, V_N.

This is not so; they can be computed from the following recursion [Biondini et.al, JLT Apr.04]:

For $n = 1, \dots, N$:

$$\begin{cases} \bar{X}_n = \frac{n-1}{n} \bar{X}_{n-1} + X_n \frac{1}{n} \\ S_n = S_{n-1} + \frac{n-1}{n} \left(X_n - \bar{X}_{n-1} \right)^2 \end{cases}$$

and finally $V_N = \frac{S_N}{N-1}$.

Proof

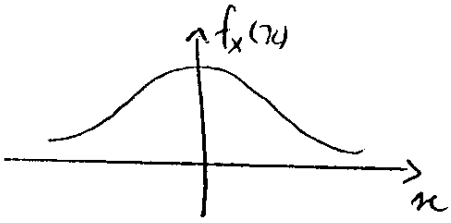
(Exercise 1) Hints:

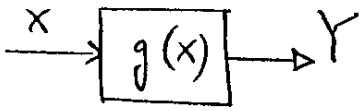
First part is trivial. For second part, just note that

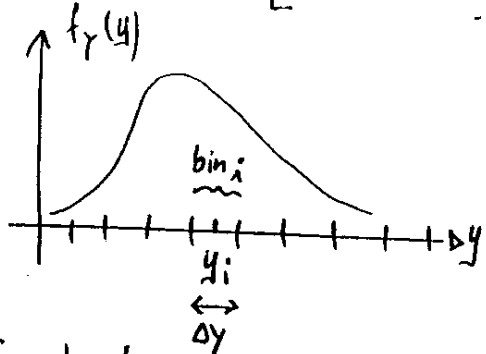
$$S_N \triangleq \sum_{n=1}^{N} \left(X_n - \bar{X}_N \right)^2 = \sum_{n=1}^{N} X_n^2 - N \bar{X}_N^2$$

and find how to compute both pieces recursively.

0 | Standard Monte-Carlo (MC) Estimation | [Jeruchim, JSAC Jan 8(







We know $f_X(x)$ and have a method to generate 110 Samples $\{x_1, x_2, x_3 \ldots\}$ from f_X. We cover the output space of Y with bins of width Δy, centered at points $\{y_i\}$.

Define the event

$$\{Y \in \text{bin}_i\} = \left\{y_i - \frac{\Delta y}{2} < Y < y_i + \frac{\Delta y}{2}\right\} \triangleq \{Y \approx y_i\} \qquad (1)$$

Define the prob. mass function (PMF) of the quantized $\tilde{Y}$ as $(i = 1, 2, \ldots)$

$$P_Y(y_i) \triangleq P\{Y \approx y_i\} \cong f_Y(y_i)\Delta y \quad \left(\text{for } \Delta y \text{ "small"} \text{ and } Y \text{ continuous RV}\right) \qquad (2)$$

Define the indicator of event $\{Y \approx y_i\}$ as

$$I_i(Y) = \begin{cases} 1 & \text{if } \{Y \approx y_i\} \\ 0 & \text{else} \end{cases} \qquad (3)$$

Then the expected value of $I_i(Y)$ is $P\{Y \approx y_i\}$ and thus from (2)

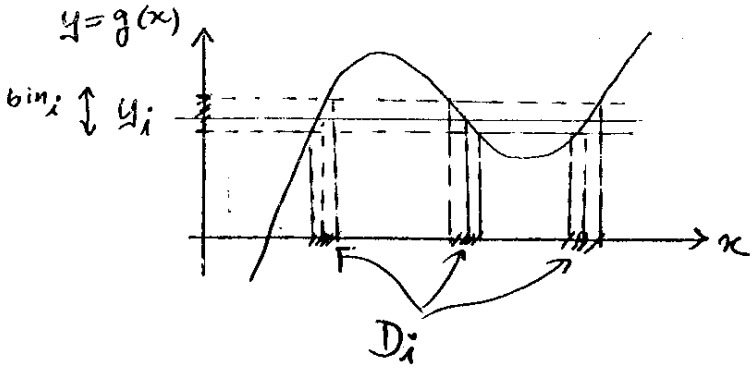
$$P_Y(y_i) = E[I_i(Y)] = \int_{-\infty}^{\infty} I_i(y) f_Y(y)\, dy$$

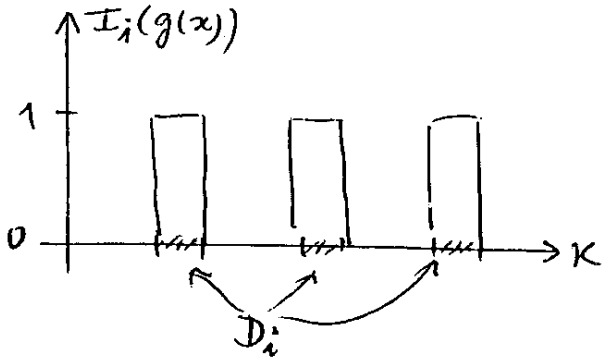
law of unconscious statistician (LUS)

$$= \int_{-\infty}^{\infty} I_i(g(x)) f_X(x)\, dx = E[I_i(g(X))] \qquad (4)$$

Also: $$P_Y(y_i) = \int_{D_i} f_X(x)\, dx \qquad (5)$$

where $D_i = \{x \in \mathbb{R} : g(x) \approx y_i\}$ set of states that map into bin i through $g(\cdot)$





MC estimation of $P_Y(y_i)$

Run N independent simulations, i.e. generate X_i and calculate $Y_i = g(X_i)$, $i = 1, \ldots, N$. We estimate $P_Y(y_i)$ as per (4):

$$\widehat{P}_Y(y_i) = \frac{1}{N} \underbrace{\sum_{J=1}^{N} I_i(g(X_J))}_{\triangleq N_i} \qquad (6)$$

N_i = # of Y samples that fall on bin i.

Now, $I_i(g(X_J)) \sim \text{Bernoulli}(P_Y(y_i))$ by construction. For different J's they are INDEPENDENT random variables (RVs). Hence

$$N_i \sim \text{Binomial}(N, P_Y(y_i)) \qquad [= \text{number of successes on } N \text{ trials w/ success prob. } P_Y(y_i)]$$

Hence

$$(6b) \quad \begin{cases} E[N_i] = N P_Y(y_i) \\ \text{Var}[N_i] = N P_Y(y_i)\left(1 - P_Y(y_i)\right) \end{cases}$$

And thus

$$E[\widehat{P}_Y(y_i)] = P_Y(y_i) \quad \leftarrow \text{UNBIASED ESTIMATOR} \qquad (7)$$

$$\sigma^2(\widehat{P}_Y(y_i)) \triangleq \text{Var}[\widehat{P}_Y(y_i)] = \frac{P_Y(y_i)\left(1 - P_Y(y_i)\right)}{N} \cong \frac{P_Y(y_i)}{N}$$

if bin ($\Delta y \to 0$) is "small" = ($P_Y(y_i)$ small)

▷ Reliability of MC estimator

Define the COEFFICIENT OF VARIATION of the MC estimator (6) as

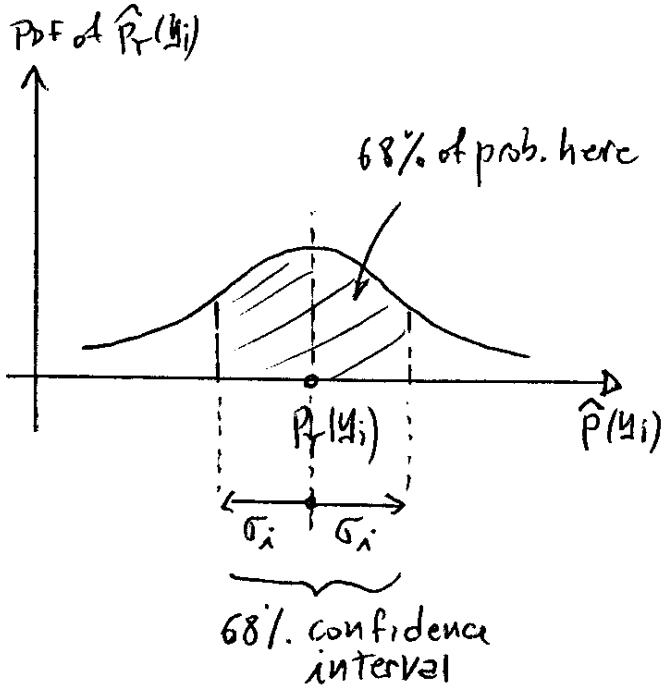
$$C_i^{MC} \triangleq \sqrt{\frac{Var\left[\hat{P}_T(y_i)\right]}{E^2\left[\hat{P}_T(y_i)\right]}} = \frac{\sigma\left(\hat{P}_T(y_i)\right)}{P_T(y_i)} \qquad (8)$$

hence $\pm C_i^{MC}$ is the RELATIVE ERROR of the estimator.

More precisely, when N large, by the central limit Theorem (CLT)

$$(9) \qquad \hat{P}_T(y_i) \longrightarrow \mathcal{N}\left(\eta = P_T(y_i),\ \sigma^2 = \frac{P_T(y_i)\left(1-P_T(y_i)\right)}{N}\right)$$

Gaussian PDF



CONFIDENCE LEVEL

$$P\left\{\hat{P}_T(y_i) \in \left(P_T(y_i) \pm \sigma_i\right)\right\} \cong 0.68$$

$$\pm 2\sigma_i \cong 0.95$$

$$\pm 3\sigma_i > 0.99$$

So, with 95% confidence,

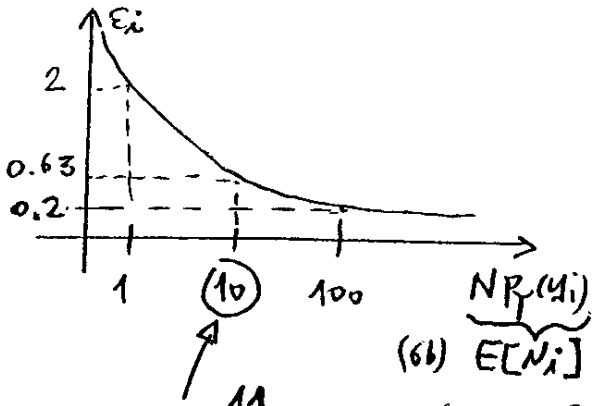
$$\hat{P}_T(y_i) = P_T(y_i) \pm 2\sigma_i$$

$$= P_T(y_i)\left[1 \pm 2C_i^{MC}\right] \qquad (10)$$

$\triangleq \varepsilon_i$ = relative error of MC estimator for a $(\pm 2\sigma)$ confidence interval

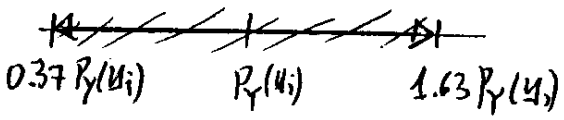
How does (95% confidence) percent error decrease with N?

(11) $\varepsilon_i = 2C_i^{nc} \overset{(7)}{\approx} \frac{2}{\sqrt{N P_f(y_i)}}$ For P_f "small"

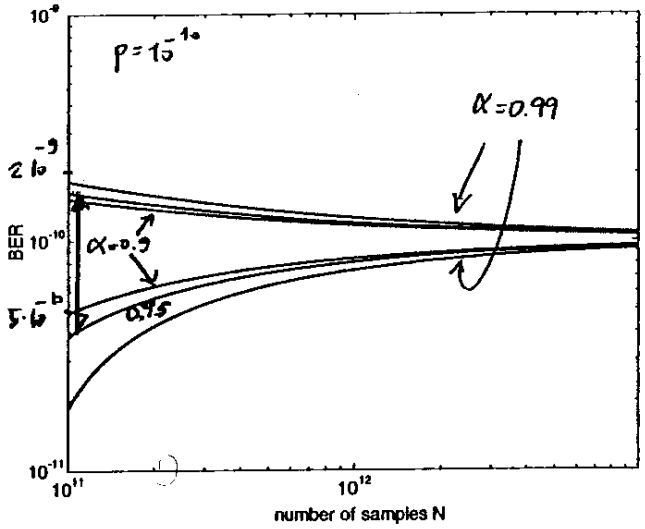


So if want to estimate $P_f(y_i) = 10^{-10}$ and use $N = 10^{11}$ samples ($NP_f = 10$)

then $\varepsilon_i = 0.63$, then your 95% confidence interval is



which is not too bad if plotted on a log scale:



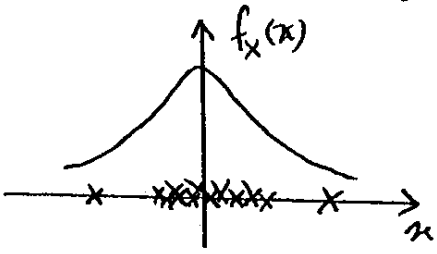
α = confidence level

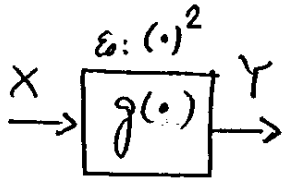
.... so to have a decent estimate you should count (on average) $E[N_i] = 10$ points in bin i, a good estimate $E[N_i] = 100$ (which means an error of $\pm 20\%$)

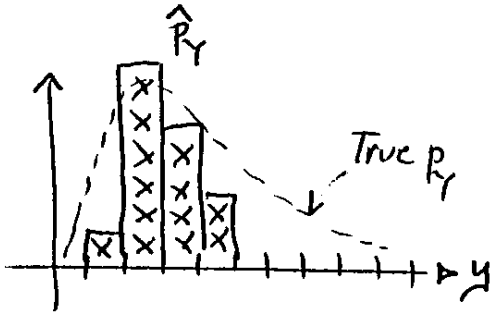
◦ Importance Sampling

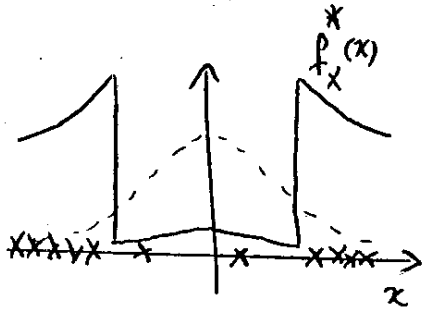
(Shanmugan et al, Trans. Commun. Nov '80)

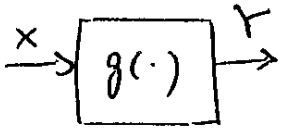
To estimate more accurately the PMF $P_Y(y_i)$ ON SPECIFIC BINS, one can use the following trick:

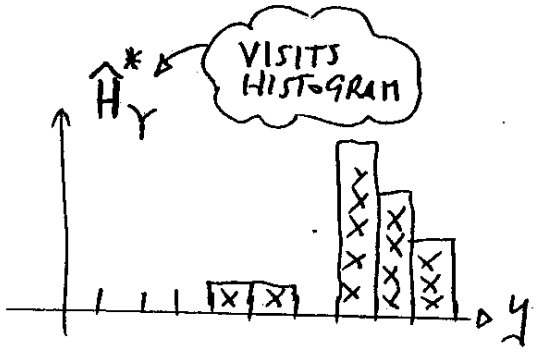












if we WARP the PDF of X....

....we estimate a different (normalized) histogram (PMF)
(I use the letter H to remind us of that)

We have more counts on tails: by suitably "weighting" every count (×), we can now "correctly" estimate $\hat{P}_Y(y_i)$ on tails, where with standard MC I would not get any counts!

Let's see how to "weight".

In the run with f_X^*, define

This reminds us that Expect. done w.r.t. f_X^*

$$H_Y^*(y_i) \triangleq P^*\{Y \ge y_i\} = E^*[I_i(Y)] \overset{LUS}{=} \int_{-\infty}^{\infty} I_i[g(x)] f_X^*(x)\, dx \qquad (12)$$

and we estimate it as

$$\hat{H}_Y^*(y_i) = \frac{1}{N}\underbrace{\sum_{j=1}^{N} I_i(g(X_j))}_{\triangleq N_i^*} = \frac{N_i^*}{N} \qquad (13)$$

(13) N_i^* = # visits of bin i in the run with f_X^*

When I instead simulate samples from $f_X(x)$, the true one, instead of (12) I get

$$P_Y(y_i) = P\{Y \simeq y_i\} \overset{(4)}{=} E[I_i(g(X))] = \int_{-\infty}^{\infty} I_i[g(x)]\, f_X(x)\, dx \qquad (14)$$

But (14) can also be obtained as $\left(f_X^*(x) = 0 \Rightarrow f_X(x) = 0\right)$

$$P_Y(y_i) = \int_{-\infty}^{\infty} I_i[g(x)] \underbrace{\left[\frac{f_X(x)}{f_X^*(x)}\right]}_{\triangleq W(x)\ \text{"weight"}} f_X^*(x)\, dx = E^*\left[I_i(g(X))W(X)\right] \qquad (15)$$

which can be estimated by sampling with f_X^*:

$$\hat{P}_Y(y_i) = \frac{1}{N}\sum_{J=1}^{N} \underbrace{I_i(g(X_J))W(X_J)}_{\triangleq W_{iJ}} = \frac{Z_i}{N} \qquad (16)$$

$\triangleq Z_i$

Z_i: # of visits to bin i, suitably weighted

At each visit of bin i, $I_i(g(X_J)) = 1$, and there are N_i^* such visits.
Hence I can rewrite (16) as:

$$\hat{P}_Y(y_i) = \underbrace{\left(\frac{N_i^*}{N}\right)}_{\hat{H}_Y^*(y_i)} \cdot \underbrace{\left[\frac{1}{N_i^*}\sum_{n=1}^{N_i^*} W(X_{J_n})\right]}_{\{J_n:\, g(X_{J_n}) \simeq y_i\}} \qquad (17)$$

$\triangleq \hat{W}_i^*$

$\hat{W}_i^*$: Avg weight of all Y samples (from f^*) that fall in bin i

Eqs (16),(17) are the "weighting rule" to convert the visits histogram (PMF) into an estimate of the true histogram (PMF) $P_Y(y_i)$ [cfr Shanmugan p 1918, eq. (11)].

Interpretation: Eq. (15) can also be written as

$$P_Y(y_i) \overset{(15)}{=} \int_{D_i} f_X(x)\,dx = \int_{D_i} \frac{f_X}{f_X^*} f_X^*(x)\,dx = \underbrace{\left[\int_{D_i} f_X^*(x)\,dx\right]}_{\overset{(12)}{=} H_Y^*(y_i)} \cdot \int_{D_i} \underbrace{\frac{f_X}{f_X^*}}_{w(x)} \underbrace{\frac{f_X^*(x)}{\left[\int_{D_i} f_X^*(x)\,dx\right]}}_{\text{by def. this is } f_X^*(x \mid X \in D_i), \text{ hence}} dx$$

$$= E^*[w(X) \mid X \in D_i] \triangleq \overline{w}_i^*$$

Therefore we get:

$$\boxed{P_Y(y_i) = H_Y^*(y_i) \cdot \overline{w}_i^*} \quad (18)$$

which justifies the estimator (17).

▸ Reliability of IS estimator

We now verify over which bins the IS estimator (16), (17) has lower variance than the MC estimator (6). We need $E^*[\hat{P}_Y(y_i)]$, $\mathrm{Var}^*[\hat{P}_Y(y_i)]$ in (16):

$$E^*[\hat{P}_Y(y_i)] \overset{(16)}{=} \frac{1}{N} \sum_{j=1}^{N} E^*\underbrace{[I_i(g(x_j)w(x_j)]}_{w_{ij}} \overset{(15)}{=} \frac{N\,p_Y(y_i)}{N} = P_Y(y_i) \quad \text{UNBIASED!}$$

Then

$$\mathrm{Var}^*[\hat{P}_Y(y_i)] \overset{(IND)}{=} \frac{1}{N^2} \sum_{j=1}^{N} \mathrm{Var}^*[w_{ij}] = \frac{1}{N}\left(E^*[w_{ij}^2] - \left(E^*[w_{ij}]\right)^2\right)$$

$$\hookrightarrow = E^*\left[I_i(g(x_j))\,w^2(x_j)\right] = \int_{D_i} f_X^*(x)\,\underbrace{w^2(x)}_{\left(\frac{f_X}{f_X^*}\right)^2} dx = \int_{D_i} f_X(x) \cdot \underbrace{\left[\frac{f_X}{f_X^*}\right]}_{w} dx$$

$I_i^2 \equiv I_i$

Hence

$$\mathrm{Var}^*\left[\hat{P}_\Gamma(y_i)\right] = \frac{1}{N_{IS}}\left(\int_{D_i} f_X(x)\underbrace{\left[\frac{f_X(x)}{f_X^*(x)}\right]}_{w(x)} dx - P_\Gamma^2(y_i)\right) \qquad (19a)$$

Rewrite this way

$$\ldots = \frac{1}{N_{IS}}\left(\underbrace{\left[\int_{D_i} f_X(x)\,dx\right]}_{P_\Gamma} \cdot \underbrace{\int_{D_i} \frac{f_X(x)}{\left[\int_{D_i} f_X dx\right]} w(x)\,dx}_{\bar{W}_i \triangleq E\left[w(x)\,|\,X\in D_i\right]} - P_\Gamma^2\right)$$

($f_X(x|X\in D_i)$; AVG weight on D_i)

Hence

$$\underbrace{\mathrm{Var}^*\left[\hat{P}_\Gamma(y_i)\right]}_{\triangleq \sigma^2_{IS}} = \frac{P_\Gamma(y_i)}{N_{IS}}\left(\bar{W}_i - P_\Gamma(y_i)\right) \qquad (19b)$$

To be checked against that of the MC estimator (6):

$$\underbrace{\mathrm{Var}\left[\hat{P}_\Gamma(y_i)\right]}_{\triangleq \sigma^2_{MC}} \overset{(7)}{=} \frac{1}{N_{MC}}\left(\int_{D_i} f_X(x)\,dx - P_\Gamma^2(y_i)\right) \qquad (20a)$$

$$= \frac{P_\Gamma(y_i)}{N_{MC}}\left(1 - P_\Gamma(y_i)\right) \qquad (20b)$$

✗ Comparing (19) and (20) we see that IS has lower variance (for same # of runs $N = N_{IS} \equiv N_{MC}$) over those "bins" D_i in state space for which

$$f_X^*(x) > f_X(x) \quad \left(\text{hence } w(x) < 1\right)$$

on average (cfr Figure at page 13) over D_i.

Exercise: Prove that

$$\boxed{\overline{w}_i = \overline{Wi}^{*}\left(1 + \frac{\mathrm{Var}^{*}[w(x)|x\in\mathcal{D}_i]}{\overline{w}_i^{*2}}\right)} \qquad (21)$$

Exercise: Using same technique leading to (19), and using (18), prove that:

$$\sigma^2_{IS} = \underbrace{\frac{H^{*}_Y(y_i)}{N}}_{E^{*}[\hat{H}^{*}(y_i)]}\mathrm{Var}^{*}[w(x)|x\in\mathcal{D}_i] + \underbrace{\frac{H^{*}_Y(y_i)\left(1-H^{*}_Y(y_i)\right)}{N}}_{\mathrm{Var}^{*}[\hat{H}^{*}_Y(y_i)]}\overline{w}_i^{*2} \qquad (22)$$

Exercise: Using (17) and the conditional variance formula (CFV)⊙, give an alternative proof of (22).

⊙ Cond. Variance formula: Given $G = f(X,Y)$

$$\mathrm{Var}[G] = E_X\left[\mathrm{Var}_Y[G|X]\right] + \mathrm{Var}_X\left[E_Y[G|X]\right] \qquad (23)$$

Proof: $\forall$ RV F, $V[F] = E[F^2] - E[F]^2$. Hence r.h.s. of (23) written in brief is:

$$E_X[V_Y[G]] + V_X[E_Y[G]] = E_X[E_Y[G^2] - E_Y[G]^2] + E_X[E_Y[G]^2] - E_X[E_Y[G]]^2$$
$$= E_X[E_Y[G^2]] - E_X[E_Y[G]]^2$$
$$= V[G]$$

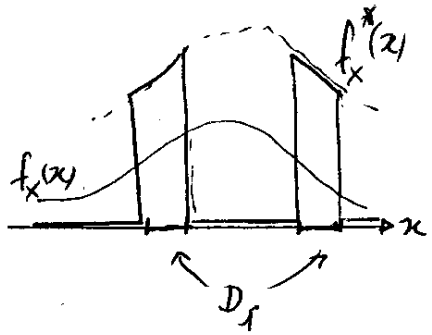
□

(E. Veach, PhD Thesis, p. 37)

Zero-Variance estimator

Rewrite (16) as $$\hat{P}_Y(y_i) = \frac{1}{N}\sum_{j=1}^{N} \frac{I_{A_i}(g(X_j))f_X(X_j)}{f_X^*(X_j)} \qquad (24)$$

If we choose



$$f_X^*(x) = \begin{cases} \frac{f_X(x)}{P_Y(y_i)} & \text{if } x \in D_i \\ 0 & \text{else} \end{cases} = \frac{f_X(x)}{P_Y(y_i)} \cdot I_{A_i}(g(x)) \qquad (25)$$

then substitution in (24) gives

$$\hat{P}_Y(y_i) = \frac{1}{N}\sum_{j=1}^{N} P_Y(y_i) = P_Y(y_i)$$

deterministic! Since all X_j samples fall into D_i.

Hence $\mathrm{Var}^*[\hat{P}_Y(y_i)] = 0$ for estimation in bin i : ZERO VARIANCE!

In this case, in fact, we have constant weight within D_i :

$\forall x \in D_i$, $$w(x) = \frac{f_X(x)}{f_X^*(x)} = P_Y(y_i)$$

Hence $$\bar{w}_i \triangleq \int_{D_i} \frac{f_X(x)}{\left[\int_{D_i} f_X(x)dx\right]} \cdot w(x)\,dx = P_Y(y_i) \cdot \frac{\int_{D_i} f_X(x)dx}{\left[\int_{D_i} f_X(x)dx\right]} = P_Y(y_i)$$

Thus $\left(\bar{w}_i - P_Y(y_i)\right) = 0$ which gives zero variance in (19).

Similarly, $$\bar{w}_i^* \triangleq \int_{D_i} \frac{f_X^*(x)}{\left[\int_{D_i} f_X^*(x)dx\right]} \cdot w(x)\,dx = P_Y(y_i)$$

Hence here $\boxed{\bar{w}_i = \bar{w}_i^*}$. Relation (21) is valid, since $\mathrm{Var}^*[w \mid x \in D_i] = 0$

Clearly, zero-variance estimator is useless, since we should know in advance $P_Y(y_i)$! However, the important fact is that weight should be constant within D_i □

Coefficient of variation & efficiency

From (7) and def(8) we have for MC estimator (cfr approx(11)):

$$C_i^{MC} = \sqrt{\frac{1 - P_r(y_i)}{N_{MC} P_r(y_i)}} \qquad (26a)$$

Similarly, for IS coeff. of variation on bin i is:

$$C_i^{IS} \overset{(8)}{=} \frac{\sigma_*\left(\hat{P}_r(y_i)\right)}{P_r(y_i)} \overset{(19)}{=} \sqrt{\frac{\bar{w}_i - P_r(y_i)}{N_{IS} P_r(y_i)}} \qquad (26b)$$

($P_r(y_i)$: unbiased!)

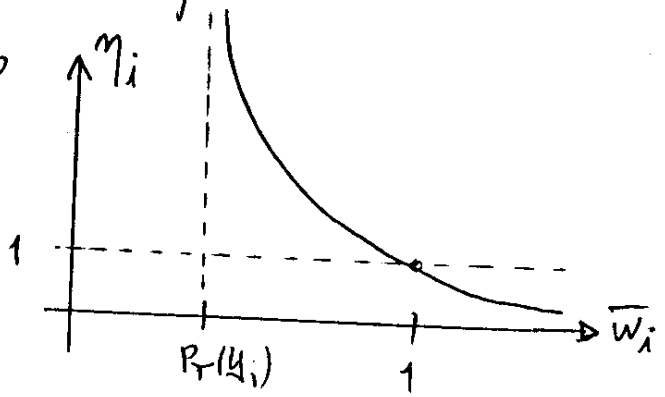
which, for the same number of runs is smaller than C_i^{MC} whenever $\bar{w}_i < 1$.

▷ It is customary to define the efficiency of IS versus MC as

$$\eta_i \triangleq \frac{N_{MC}\, \sigma_{MC}^2}{N_{IS}\, \sigma_{IS}^2} = \frac{N_{MC}}{N_{IS}} \left[\frac{C_i^{MC}}{C_i^{IS}}\right]^2 \overset{(26)}{=} \frac{1 - P_r(y_i)}{\bar{w}_i - P_r(y_i)} \qquad (27)$$

For fixed $N_{MC} = N_{IS}$ it gives the variance reduction factor; for fixed $\sigma_{MC}^2 = \sigma_{IS}^2$ it gives the computation reduction factor.

A Plot of (27) gives



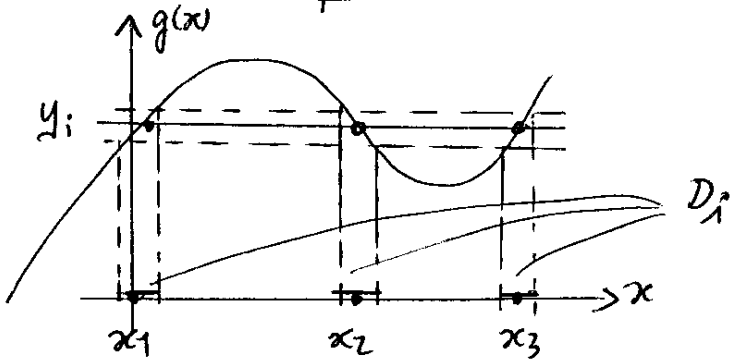
$\eta_i \to \infty$ when $\bar{w}_i \to P_r(y_i)$ (zero-variance estimator).

$\bar{w}_i$ cannot clearly be smaller than $P_r(y_i)$.

When $P_Y(y_i) \ll \overline{w}_i \ll 1$, from (27) get

$$\eta_i \cong \frac{1}{\overline{w}_i} = \frac{\int_{D_i} f_x(x)\,dx}{\int_{D_i} f_x(x) \cdot \left[\frac{f_x(x)}{f_x^*(x)}\right] dx} \qquad (28)$$

Let's reason in the simplest case of scalar X in order to get more intuition on the power and limits of IS:



K solutions $x_1, \dots, x_K$ to $y_i = g(x)$

For small Δy, thus small $|\Delta x_j| = \frac{\Delta y}{|g'(x_j)|}$, get

$$\eta_i \cong \frac{\sum_{j=1}^{K} f_x(x_j) \frac{\Delta y}{|g'(x_j)|}}{\sum_{j=1}^{K} f_x(x_j) \left[\frac{f_x(x_j)}{f_x^*(x_j)}\right] \frac{\Delta y}{|g'(x_j)|}} \qquad (29)$$

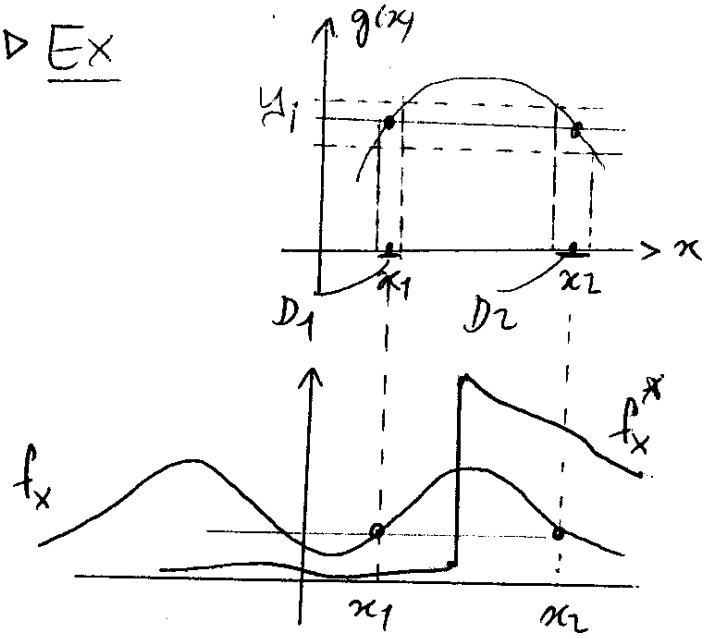
▷ When $g^{-1}(\cdot)$ exists (single solution, $K=1$) (29) gives

$$\eta_i = \frac{1}{w(x_1)} = \frac{f_x^*(x_1)}{f_x(x_1)}$$

Ex: if $f_x(x_1) = 10^{-30}$ and we use $f_x^*(x_1) = 10^{-5}$, then $\eta_i = 10^{25}$!

Gain can be impressive, and is large when weight f/f^* is small. This is the power of IS.

▷ Ex



Assume that

$$\begin{cases} |g'(x_1)| = |g'(x_2)| = g' \\ \underbrace{f_x(x_1)}_{f_1} = \underbrace{f_x(x_2)}_{f_2} \end{cases}$$

$$W(x) = \begin{cases} w_1 \gg 1 & \text{if } x \in D_1 \text{ (bad)} \\ w_2 \ll 1 & \text{if } x \in D_2 \text{ (Good)} \end{cases}$$

Then from (29)

$$\eta_i = \frac{\frac{f_1+f_2}{g'}}{\frac{w_1 f_1 + w_2 f_2}{g'}} = \frac{2}{w_1+w_2} \cong \frac{2}{w_1} \ll 1 \quad \underbrace{\text{bad!!}}$$

□

This example illustrates the real limit of IS.
In our problems we do not know the domain D_i associated with $\{Y \hat{>} y_i\}$.
Hence, when we "blindly" warp, may correctly decrease $w(x)$ on some parts of D_i, and enhance it ($W(x) > 1$) on others.
In IS, to correctly bias, you should have an idea of where your domain D_i is in state space, then "warp" to increase $f^* \gg f$ on it.

Note Without weight-variation on D_i, have $Var^*[W(X) | X \in D_i] = 0$,
hence σ_{IS}^2 is seen from (22) to have ⊘ first addendum:
"zero weight variance reduces σ_{IS}^2"

Also with uniform weight (UW) $w(x) = \bar{w}_i \equiv \bar{w}_i^* \quad \forall x \in D_i$

have from (22):
$$\sigma_{IS}^2 = Var^*\left[\frac{N_i^*}{N_{IS}}\right] \cdot \bar{w}_i^2 \qquad (30)$$

while
$$\sigma_{MC}^2 \overset{(6)}{=} Var\left[\frac{N_i}{N_{MC}}\right]$$

Even with same accuracy of histogram count $\left(Var\left[\frac{N_i}{N_{MC}}\right] = Var^*\left[\frac{N_i^*}{N_{IS}}\right]\right)$
IS variance reduction is due to weight factor $\bar{w}_i^2$!!

Estimating Variance from Simulations

Some authors suggest to verify the reliability of estimated probability $\hat{p}$ from the sample variance of the samples used to evaluate $\hat{p}$.

For MC estimation $\hat{p}_{MC} = \frac{1}{N}\sum_{J=1}^{N} \underbrace{I_i(g(X))}_{\triangleq N_{iJ} = \begin{cases} 1 & \text{if } X_J \in D_i \\ 0 & \text{else} \end{cases}} = N_i/N$

then $\sigma_{MC}^2 = Var[\hat{p}_{MC}] \overset{(11b)}{=} \frac{Var[N_{iJ}]}{N}$, and its estimate is

$\hat{\sigma}_{MC}^2 = \frac{\hat{Var}[N_{iJ}]}{N}$, where

$$\hat{Var}[N_{iJ}] \triangleq \frac{1}{N-1}\sum_{J=1}^{N}\left[N_{iJ} - \hat{p}_{MC}\right]^2$$ is the sample variance defined at p. 7.

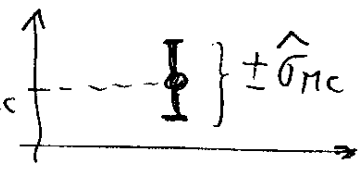
Exercise: Prove that

$$\hat{Var}[N_{iJ}] = \frac{N}{N-1}\hat{p}_{MC}(1-\hat{p}_{MC})$$

so that
$$\hat{\sigma}_{MC}^2 = \frac{\hat{p}_{MC}(1-\hat{p}_{MC})}{N-1} \qquad (31)$$

□

Then error bars are placed around estimated point:



For IS estimation,

$$\hat{P}_{IS} = \frac{1}{N}\sum_{J=1}^{N} \underbrace{I_i(g(X_J))w(X_J)}_{\triangleq W_{IJ} = \begin{cases} w(X_J) & \text{if } X \in D_i \\ 0 & \text{else} \end{cases}} = z_i/N$$

and for IID samples, estimate

$$\hat{\sigma}_{IS}^2 = \frac{\hat{Var}[W_{IJ}]}{N}, \text{ where}$$

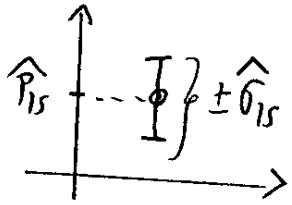
$$\hat{Var}[W_{IJ}] \triangleq \frac{1}{N-1}\sum_{J=1}^{N}\left[W_{IJ} - \hat{P}_{IS}\right]^2$$

Exercise: for UW, prove that

$$\hat{Var}[W_{IJ}] = \frac{N}{N-1}\hat{P}_{IS}\left(\bar{W}_i - \hat{P}_{IS}\right)$$

so that

$$\hat{\sigma}_{IS}^2 = \frac{\hat{P}_{IS}\left(\bar{W}_i - \hat{P}_{IS}\right)}{(N-1)} \qquad (32)$$



□

The trouble: Refer to example @ p. 21. When enough samples N_{IS} are collected from f^* in figure so that we have samples both from D_1 and D_2, variance is large (ctr. eq (22)). However, if N_{IS} is not "too large", we may have NO SAMPLES from D_1 (those with large weight $W_1 \gg 1$) and thus we estimate from samples in D_2 only (Good ones)

$P_2 = f_2 \Delta x$
$P_1 = f_1 \Delta x$

$$\hat{\sigma}_{IS}^2 \cong \frac{P_2(W_2 - P_2)}{N-1} \cong \frac{P_2 W_2}{N-1} \text{ very small!}$$

Hence we are happy with our simulation, but, alas, error is instead large!!

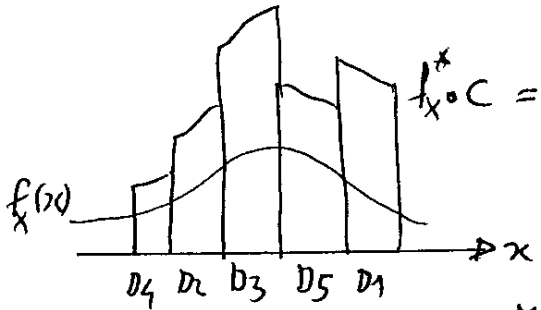
Flat Histogram IS (FHIS)

Let's go back to the zero-variance estimator, p. 18. The f_x^* in (25) is good to estimate the ith bin $P_\Upsilon(y_i)$, when knowing the location D_i in state space, but not the remaining bins.

Suppose we know the locations D_i for all bins $i=1,\dots,M$, and the associated PMF $\{P_\Upsilon(y_i)\}_{i=1}^M$.

Let's try warping with [use (25) for all bins and add up]:

$$f_X^*(x) = \frac{1}{C}\sum_{i=1}^{M} \frac{I_i(g(x))}{P_\Upsilon(y_i)} f_X(x) \qquad (33.a)$$



Norm. constant is $C = \sum_{i=1}^{M} \int \frac{I_i(g(x))}{P_\Upsilon(y_i)} f_X(x)\,dx \overset{(4)}{=} \sum_{i=1}^{M} \frac{\int_{D_i} f_X(x)\,dx}{P_\Upsilon(y_i)} \overset{(5)}{=} P_\Upsilon(y_i) = \underbrace{M}_{\#\text{ bins}}$

$\left[\int \equiv \int_{\Gamma^M \text{ sample space}}\right]$

and (33) becomes: $\sum_{i=1}^{M} \frac{I_i(g(x))}{P_\Upsilon(g(x))} = 1$

$$f_X^*(x) = \frac{1}{M}\sum_{i=1}^{M} \frac{I_i(g(x))}{P_\Upsilon(g(x))} f_X(x) = \frac{f_X(x)}{M\,P_\Upsilon(g(x))} \qquad (33.b)$$

$g(x) = y_i \;\forall x \in D_i$

IS estimator (16) becomes:

$$\hat{P}_\Upsilon(y_i) = \frac{1}{N}\sum_{j=1}^{N} \frac{I_i(g(X_j))\,f_X(X_j)}{f_X^*(X_j)} \overset{(33.b)}{=} \frac{M}{N}\underbrace{\left[\sum_{j=1}^{N} I_i(g(X_j))\right]}_{N_i^*,\ \text{a RV!}} \cdot P_\Upsilon(y_i) = \left(\frac{N_i^*}{N}\right) M P_\Upsilon(y_i) \qquad (33.c)$$

since now not all samples fall on bin i, we do have variance!

Now, $N_i^* \sim \text{Binomial}\left(N, H_Y^*(y_i)\right)$, with

(33d) $H_Y^*(y_i) \overset{(12)}{=} \int_{D_i} f_x^*(x)\,dx \overset{(33b)}{=} \frac{1}{M} \underbrace{\int_{D_i} \frac{f_x(x)\,dx}{P_Y(y_i)}}_{} = \frac{1}{M}$ Flat Histogram!

Then $E^*\left[\hat{P}_Y(y_i)\right] = \frac{M}{N} P_Y(y_i) \underbrace{E[N_i^*]}_{N H_Y(y_i)} = P_Y(y_i)$ unbiased!

And $\text{Var}^*\left[\hat{P}_Y(y_i)\right] \overset{(33c)}{=} \left(\frac{M P_Y(y_i)}{N}\right)^2 \cdot \underbrace{\text{Var}^*[N_i^*]}_{N H_Y^*(1-H_Y^*)}$

$$= \frac{P_Y^2(y_i)}{N} M\left(1 - \frac{1}{M}\right) \qquad (33e)$$

Therefore relative error is (many bins $M \gg 1$)

$$C_i^{\text{FHIS}} = \sqrt{\frac{\left(1-\frac{1}{M}\right)}{\underbrace{\left(\frac{N}{M}\right)}_{\triangleq N_b}}} \cong \frac{1}{\sqrt{N_b}} \qquad (33f)$$

N_b: AVG # of points per bin, $E^*[N_i^*] = \frac{N}{M}$

▷ Hence its fundamental properties are:

a) $E\left[\hat{H}_Y^*(y_i)\right] = H_Y^*(y_i) = \frac{1}{M}$ Histogram of visits on average is flat

b) relative error is <u>SAME</u> for all bins:
all bins estimated with same accuracy!

c) weight $w(x) = \frac{f(x)}{f^*(x)} = M P_Y(g(x))$
is <u>constant</u> for all $x \in D_i$
(a special case of UWIS)

Output warping in IS

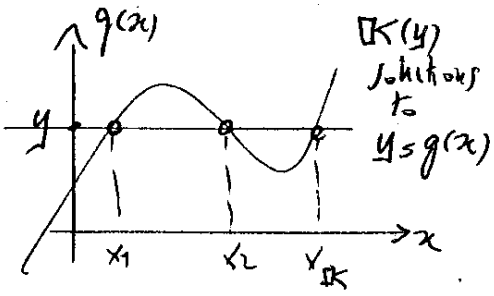
$$f_X(x) \longrightarrow \boxed{g(\cdot)} \longrightarrow f_Y(y), \quad P_Y(y_i) = f_Y(y_i)\,\Delta y$$

$$\underbrace{f_X^*(x)}_{?} \longrightarrow \boxed{g(\cdot)} \longrightarrow f_Y^*(y) = \frac{f_Y(y)}{c \cdot d(y)}, \qquad H_Y^*(y_i) = \frac{P_Y(y_i)}{c \cdot d(y_i)}$$

Consider the problem of finding the input warped PDF $f_X^*(x)$ such that the output PDF is warped according to the output weight $w_Y(y) = \frac{f_Y(y)}{f_Y^*(y)} = c \cdot d(y)$, where $d(y)$ is a PDF >0 and c a suitable normal. constant for f_Y^*.

A sufficient condition is that

$$f_X^*(x) = \frac{f_X(x)}{c \cdot d(g(x))}$$



Proof For simplicity we give it for scalar X.
Consider $Y = g(X)$.
By the fundamental theorem (FT):

$$f_Y^*(y) = \sum_{k=1}^{K} \frac{f_X^*(x_k)}{|g'(x_k)|} = \sum_{k=1}^{K} \frac{f_X(x_k)}{|g'(x_k)|} \cdot \frac{1}{c\, d(\underbrace{g(x_k)}_{=y})} =$$

$$= \underbrace{\left(\sum_{k=1}^{K} \frac{f_X(x_k)}{|g'(x_k)|} \right)}_{\overset{FT}{=} f_Y(y)} \cdot \frac{1}{c \cdot d(y)} = \frac{f_Y(y)}{c \cdot d(y)} \qquad \square$$

In Summary

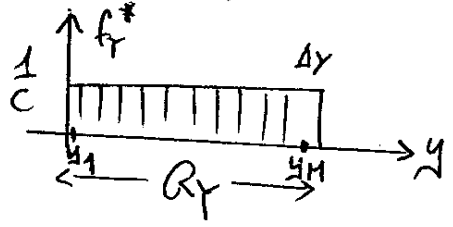
$$f_X^*(x) = \frac{f(x)}{c\, d(g(x))} \longrightarrow \boxed{g(\cdot)} \longrightarrow f_Y^*(y) = \frac{f_Y(y)}{c\, d(y)}$$

Hence choosing $d(y) \equiv f_Y(y)$ get ($f_Y(y_i) = P_Y(y_i)\,\Delta y$ after binning)

(33g) $$f_X^*(x) = \frac{f_X(x)}{\frac{c}{\Delta y} P_Y(g(x))} \longrightarrow \boxed{g(\cdot)} \longrightarrow f_Y^*(y) = \frac{1}{c}$$

$$H_Y^*(y_i) = f_Y^*(y_i)\,\Delta y = \frac{\Delta y}{c}$$ FLAT HISTOGRAM

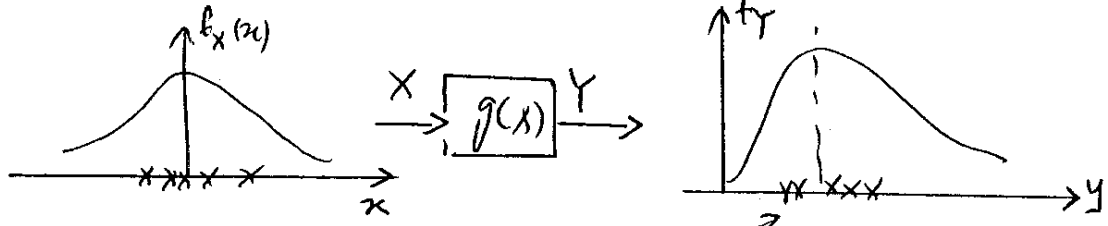
From $1 = \int f_Y^*(y)\,dy$ get



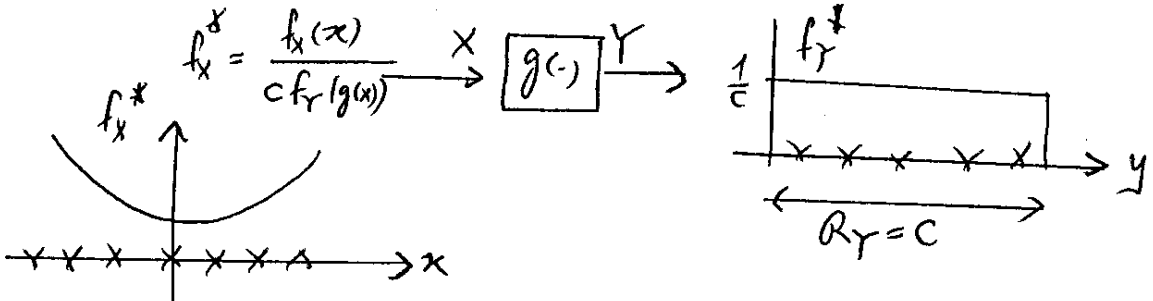
$$\boxed{c = \underbrace{R_Y}_{\text{range of } Y} = M \cdot \Delta y} \quad (33h)$$

thus obtaining again result (33b).

Note



when sampling directly from f_X, most samples fall on modal bins



Philosophy of flat histogram:

- Reduce f_X^* where it would produce too many samples on f_Y^*
- Increase f_X^* " " " few " " "),

so that our N samples are <u>evenly spread</u> on output space.

Uniform Weight IS (UWIS)

Let $\underline{\theta} \triangleq \{\theta(y_i)\}_{i=1}^n$ be a (possibly un-normalized) PMF, with $\underline{\theta(y_i) > 0}$ (strictly).

A biasing scheme as in (33g):

$$f_X^*(x) = \frac{f_X(x)}{C_\theta \theta(g(x))} \overset{(33g)}{=} \frac{1}{C_\theta} \sum_{i=1}^n \frac{I_i(g(x))}{\theta(y_i)} f_X(x) \qquad (33i)$$

is a UWIS, with constant weight on each D_i equal to

$$w(x) = \frac{f_X}{f_X^*} = C_\theta \theta(y_i) \qquad \forall x \in D_i$$

The normalization constant is

$$C_\theta = \int_\Gamma \sum_{i=1}^n \frac{I_i(g(x))}{\theta(y_i)} f_X(x)\,dx = \sum_{i=1}^n \frac{\int_{D_i} f_X(x)\,dx}{\theta(y_i)} = \sum_{i=1}^n \frac{P_Y(y_i)}{\theta(y_i)} \qquad (33l)$$

The expected visits histogram is

$$H_Y^*(y_i) \triangleq \int_{D_i} f_X^*(x)\,dx = \frac{1}{C_\theta} \sum_{j=1}^n \int_{D_i} \frac{I_j(g(x))}{\theta(y_j)} f_X(x)\,dx = \frac{1}{C_\theta} \frac{\left(\int_{D_i} f_X(x)\,dx\right)}{\theta(y_i)}$$

($P_Y(y_i)$; all zero, except for $i=j$)

So:

$$H_Y^*(y_i) = \frac{\frac{P_Y(y_i)}{\theta(y_i)}}{\sum_{j=1}^n \frac{P_Y(y_j)}{\theta(y_j)}} \qquad (33m)$$

$\overline{w_i} = \overline{w_i}^*$ constant weight!

Estimator is $\hat{P}_Y(y_i) \overset{(17)}{=} \left(\hat{H}_Y^*(y_i)\right) \hat{\overline{w}}_i^* = \left(\frac{N_i^*}{N}\right) \cdot \underbrace{C_\theta \theta(y_i)}$ (33n)

As per p.15 it is unbiased: $E^*[\hat{P}_Y(y_i)] = P_Y(y_i)$ with variance

$$\sigma^2_{UWIS}(y_i) \overset{(20)}{=} \frac{P_Y(y_i)}{N}\left(\overbrace{C_\theta \theta(y_i)}^{\overline{w_i}} - P_Y(y_i)\right) \qquad (33o)$$

on binomial $\overset{(22)}{=} \mathrm{Var}^*[\hat{H}_Y^*(y_i)] \cdot [C_\theta \theta(y_i)]^2 \qquad (33p)$

Exercise: show that the UWIS that minimizes

$$A = \sum_{j=1}^{M} \overline{W}_i = C_\theta \left[\sum_{j=1}^{M} \theta(y_i) \right]$$

subject to constraint $\sum_{j=1}^{M} \theta(y_i) = 1$ (proper PMF) is, $\forall k = 1, .., M$

$$\theta(y_k) = \frac{\sqrt{R_Y(y_k)}}{\sum_{j=1}^{M} \sqrt{R_Y(y_j)}}$$

and $\min_{\underline{\theta}} C_\theta = 1$ .. (Hint: Use Method of Lagrange Multipliers).

Exercise: Prove that for a UWIS we have a coeff. of variation for bin i:

$$C_i^{UWIS} = \sqrt{\frac{1 - H_Y^*(y_i)}{N\, H_Y^*(y_i)}} \simeq \frac{1}{\sqrt{N H_Y^*(y_i)}} = \frac{1}{\sqrt{E[N_i^*]}} \qquad (33q)$$

Exercise: Prove that, among all PMFs $\underline{\theta}$ for UWIS, the FHIS ($\theta(y_i) = R_Y(y_i)\ \forall i$) is the one that minimizes the worst value of the relative error over all bins, $\max_i C_i^\theta$.

In other words, prove that, $\forall$ proper $\underline{\theta} > 0$

$$\max_i \left(C_i^\theta \right)^2 \geq \left(C_i^{FHIS} \right)^2 \overset{(33f)}{=} \frac{M-1}{N} \qquad (33r)$$

Flat Histogram Methods

FH methods are a family of algorithms which perform an adaptive UWIS: starting from an initial guess $\underline{\theta}_0$, one builds a sequence

$\underline{\theta}_0, \underline{\theta}_1, \dots, \underline{\theta}_n$ such that the exp. visits histogram converges to the flat histogram:

$$H_Y^{(n)}(y_i) \longrightarrow \frac{1}{M} \quad \forall \text{bin } i \quad \text{as } n \text{ increases.}$$

$$\underbrace{\frac{P_Y(y_i)}{C_n \theta_n(y_i)}}_{\text{indep. of } i} \longrightarrow \frac{1}{M} \quad (\text{as } n \to \infty) \Rightarrow \begin{cases} C_n = M \\ \theta_n(y_i) = P_Y(y_i) \quad \forall i = 1, \dots M \end{cases}$$

(assuming $\underline{\theta}$ is a proper PMF)

Hence, flatness of the expected output histogram, which is checked by the visits histogram $\hat{H}^{(n)}(y_i)$, is evidence of convergence of $\theta_n(y_i)$ to the true $P_Y(y_i)$.

▷ Flat Histogram methods include MMC, Wang Landau, and others [F. Liang "A theory on flat Histogram MC Algorithms" J. Stat. Phys., 2006]

▷ What distinguishes the algorithms is their specific update law

$$\underline{\theta}_n \longrightarrow \underline{\theta}_{n+1} \quad \text{from cycle } n \to n+1$$

The first of these algorithms was MMC, which we will now describe.

MMC

For MMC, we will use the symbol $\hat{P}_Y^{(n)}(y_i)$ to designate the PMF estimate $\theta_n(y_i)$ at step n, for consistence of notation with the previous section on IS.

The update law of MMC from step n to $n+1$ is the following:

▷ cycle $(n+1)$:

$$f_X^{(n+1)}(x) = \frac{f_X(x)}{C_n \hat{P}_Y^{(n)}(y_i)} \tag{34a}$$

Run N "simulations"

$X \rightarrow \boxed{g(\cdot)} \rightarrow Y$

Estimate $\forall$ bin $i = 1, \dots, M$

$$\hat{H}_Y^{(n+1)}(y_i) = \frac{N_i^{(n+1)}}{N} \tag{34b}$$

← COUNTS in bin i at step $(n+1)$

Form new estimate $\hat{P}_Y^{(n+1)}(y_i)$ as per (17):

$$\hat{P}_Y^{(n+1)}(y_\ell) = \hat{H}_Y^{(n+1)}(y_i) \cdot \left[\frac{1}{N_i^{(n+1)}} \sum_{j=1}^{N_i^{(n+1)}} W(X_j) \right]$$

unif. weight

$$\frac{f_X}{f_X^{(n+1)}} \overset{(34a)}{=} C_n \hat{P}_Y^{(n)}(y_i)$$

Hence

UPDATE LAW for MMC:

$$\boxed{\hat{P}_Y^{(n+1)}(y_i) = C_n \hat{P}_Y^{(n)}(y_i) \hat{H}^{(n+1)}(y_i)} \tag{34c}$$

with initial guess

$$\hat{P}_Y^{(0)}(y_i) = \frac{1}{M} \text{ (uniform)}, \quad C_0 = M \tag{34d}$$

so that step 1 is in fact pure MC: $f_X^{(1)}(x) \equiv f_X(x)$.

Note: while $\left\{\hat{H}_Y^{(n+1)}(y_i)\right\}_{i=1}^{M}$ is a proper PMF by construction,

$\underline{\hat{P}}_{n+1} \triangleq \left\{\hat{P}_Y^{(n+1)}(y_i)\right\}_{i=1}^{M}$ whose average is the desired PMF: $E\left[\hat{P}_Y^{(n+1)}(y_i)\right] = P_Y(y_i)$

since IS estimators are unbiased (cfr (15)), is instead

a normalized PMF only on average:

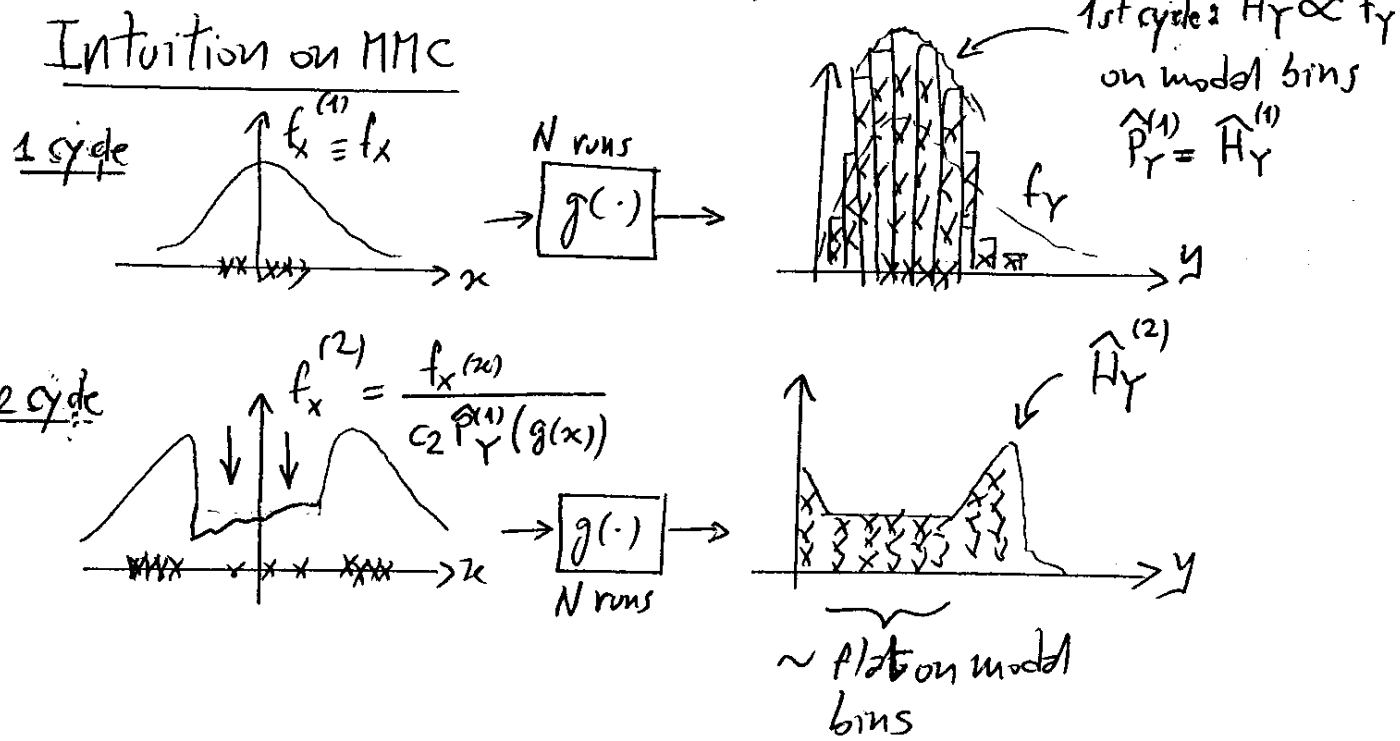
$$E\left[\sum_{i=1}^{M}\hat{P}_Y^{(n+1)}(y_i)\,\middle|\,\underline{\hat{P}}_n\right] = E\left[\sum_{i=1}^{M}\frac{N_i^{(n+1)}}{N}C_n\hat{P}_Y^{(n)}(y_i)\,\middle|\,\underline{\hat{P}}_n\right]$$

$$= C_n\sum_{i=1}^{M}E\left[\hat{H}^{(n+1)}(y_i)\,|\,\underline{\hat{P}}_n\right]\cdot\hat{P}_Y^{(n)}(y_i)$$

$$= H^{n}(y_i) \overset{(33m)}{=} \frac{P_Y(y_i)}{C_n\hat{P}_Y^{(n)}(y_i)}$$

$$= C_n\sum_{i=1}^{M}\frac{P_Y(y_i)}{C_n\hat{P}_Y^{(n)}(y_i)}\hat{P}_Y^{(n)}(y_i) = 1$$

Intuition on MMC

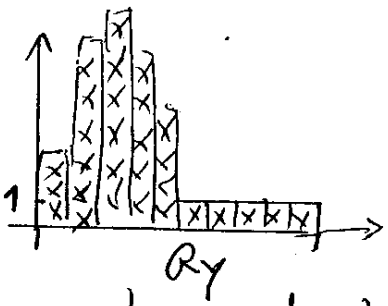


Problem: unexplored bins where $\hat{H}_Y^{(n)}(y_i) = 0$.

We would divide by 0 in (34.a)

In STANDARD MMC solution is to use

(35) $$\widehat{H}^{(n+1)}(y) = \frac{\max\left(N_i^{(n+1)}, 1\right)}{N}$$



i.e. add 1 in bins not visited. $\{\widehat{H}_i^{(n+1)}\}_{i=1}^{M}$ is not normalized.
Normalization goes into constant C_n in (34c).

I will next show an example:

Input space $\underline{X}$ has dimension $d=10$ and transformation is

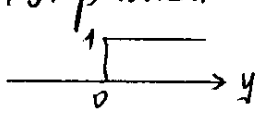
$$Y = g(\underline{X}) = \sum_{i=1}^{10} X_i^2$$

where $\{X_i\}$ are IID RV's $\sim N(0,1)$ (standard Gaussian)

Exact PDF Y is Chi-Squared

$$f_Y(y) = \frac{e^{-\frac{y}{2}}\, y^{\left(\frac{d}{2}-1\right)}}{2^{\frac{d}{2}}\left(\frac{d}{2}-1\right)!}\, U(y)$$

unit step function



At each cycle, I will use N runs. I will show plots of $\frac{\widehat{P}_Y^{(n)}(y_i)}{\Delta y} \simeq \widehat{f}_Y^{(n)}(y_i)$ for the first cycles $n=1,..,6$.

At each cycle, I will also show MC estimation with $N_{MC} = n \cdot N$ runs for comparison.

MMC PDF estimation of $Y = \Sigma_{i=1}^{10} X_i^2$, $X_i \sim \mathcal{N}(0,1)$

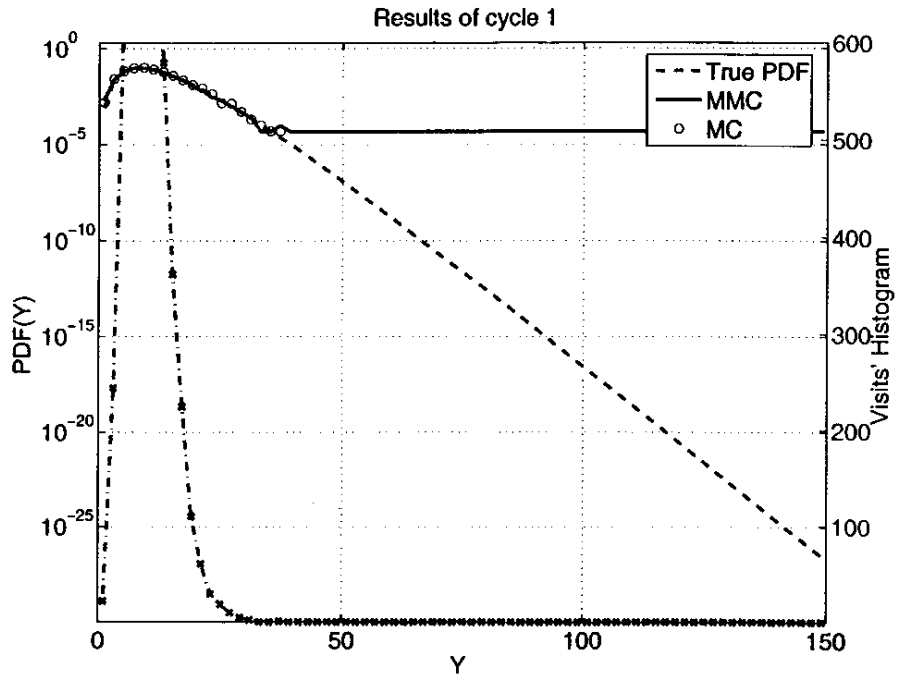


Figure 1: Standard MMC, N=10000, M=75 bin.

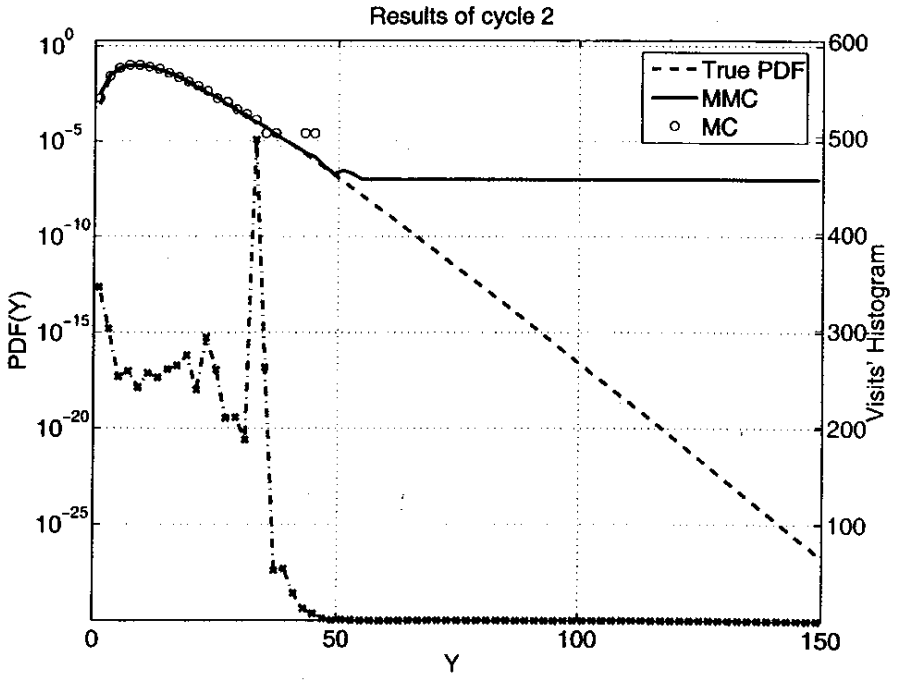


Figure 2: Standard MMC, N=10000.

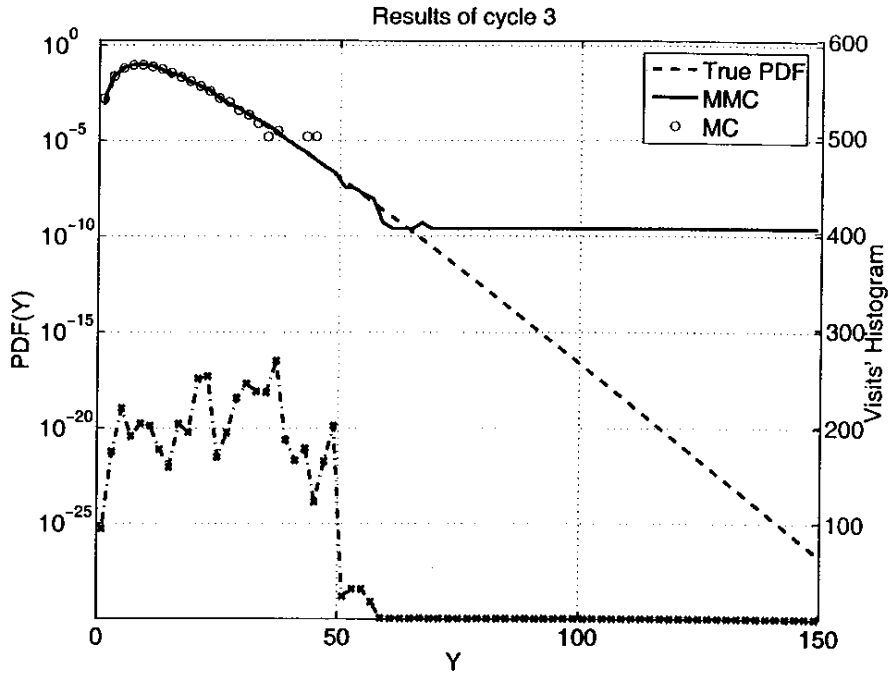


Figure 3: Standard MMC, N=10000.

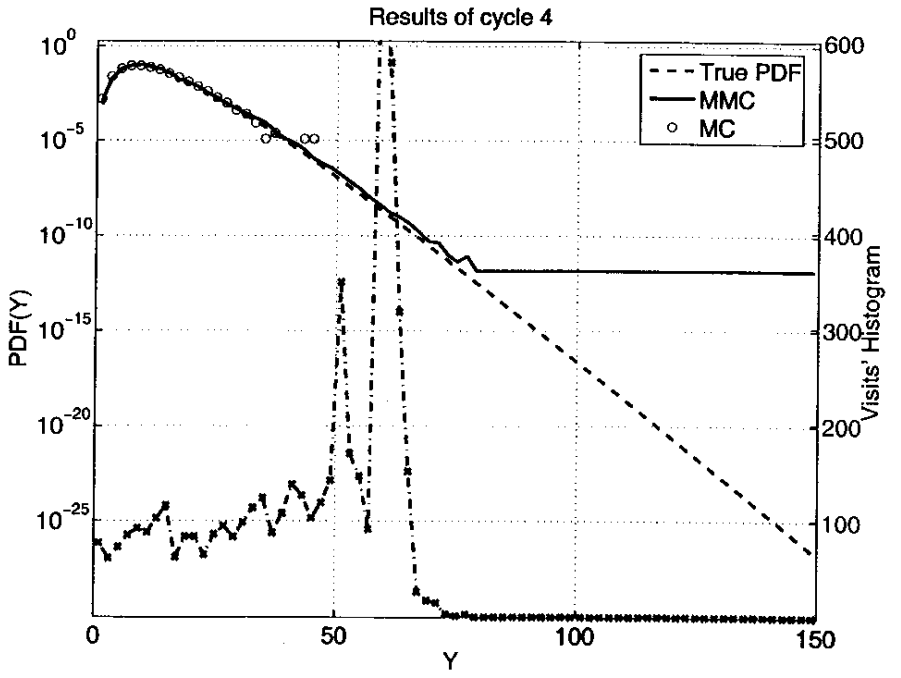


Figure 4: Standard MMC, N=10000.

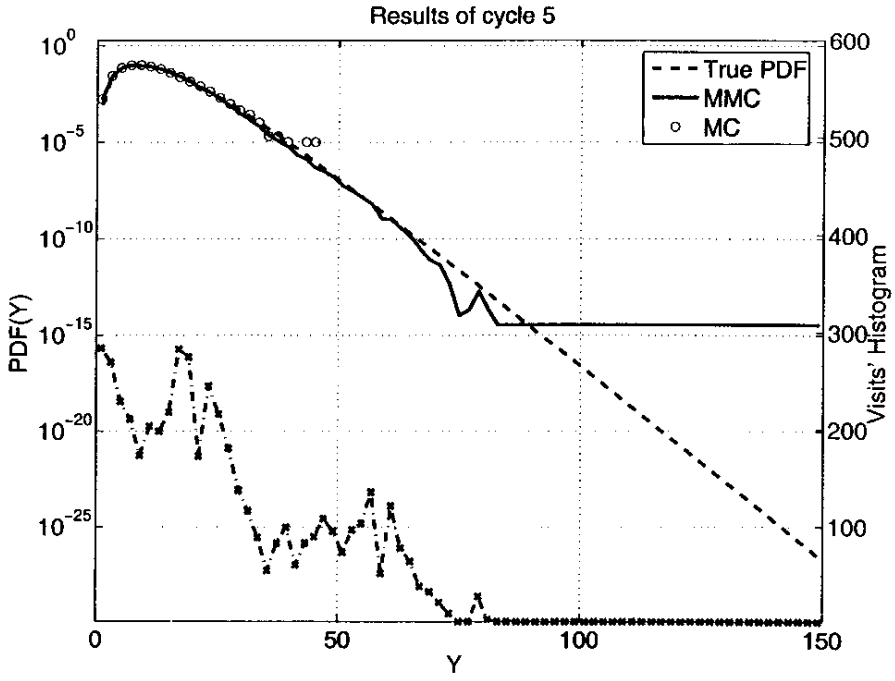


Figure 5: Standard MMC, N=10000.

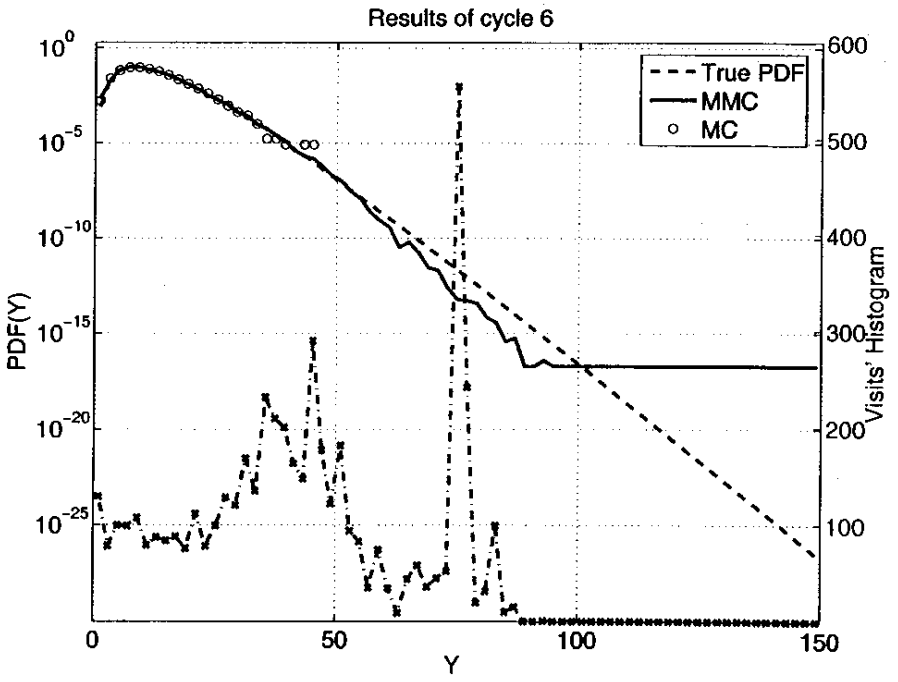
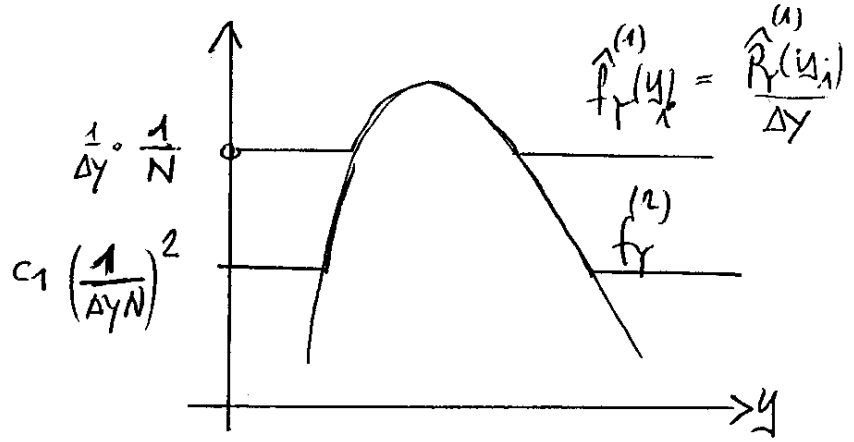


Figure 6: Standard MMC, N=10000.



$\Delta y = \frac{Q_Y}{M} = \frac{150}{75} = 2$ in example.

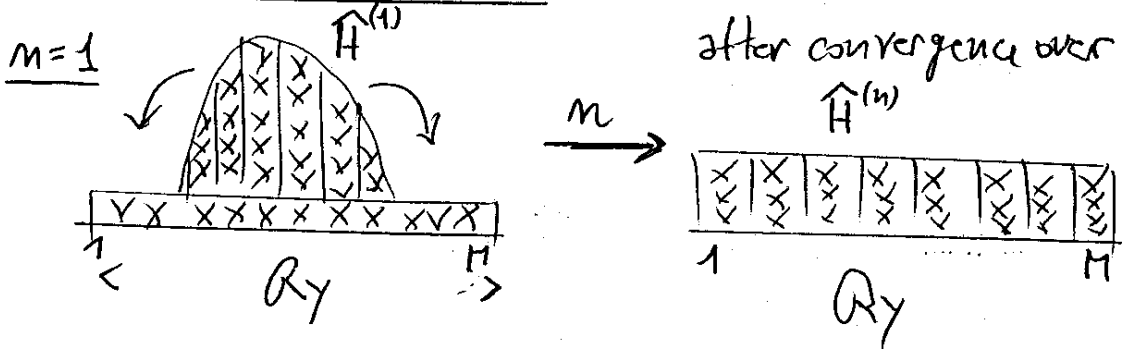
First MC cycle reaches "floor" when on avg have 1 count, i.e. at a level $\left(\frac{1}{N}\right)$ for $\hat{R}_Y^{(1)}(y_i) \equiv \hat{H}_Y^{(1)}$.

Then from update (34c) floor at 2nd cycle goes to a level $\frac{1}{N}$ for $\hat{H}_Y^{(2)}$:

$$\hat{f}_Y^{(2)}\Big|_{floor} = \frac{\hat{P}_Y^{(2)}\Big|_{floor}}{\Delta y} \overset{(34c)}{=} c_1 \underbrace{\left(\frac{1}{\Delta y} \frac{1}{N}\right)}_{\hat{P}_Y^{(1)}\Big|_{Floor}} \cdot \left(\frac{1}{N} \cdot \frac{1}{\Delta y}\right) \quad (36)$$

and so forth: that's how much MMC "digs" at each cycle.

Empirical Choice of N:



At convergence, N samples are uniformly spread over M bins.

Because of (33g), since $E[N_j^*] = \frac{N}{M}$, if desired relative error is C, then $E[N_1^*] \overset{(33g)}{=} \frac{1}{C^2}$

and thus we select

$$(37) \quad \boxed{N = \frac{M}{C^2}}$$

(Typically $E[N_j^*] = 100 \div 1000$ for satisfactory accuracy, cfr. page 12)

Effect of cycle size N

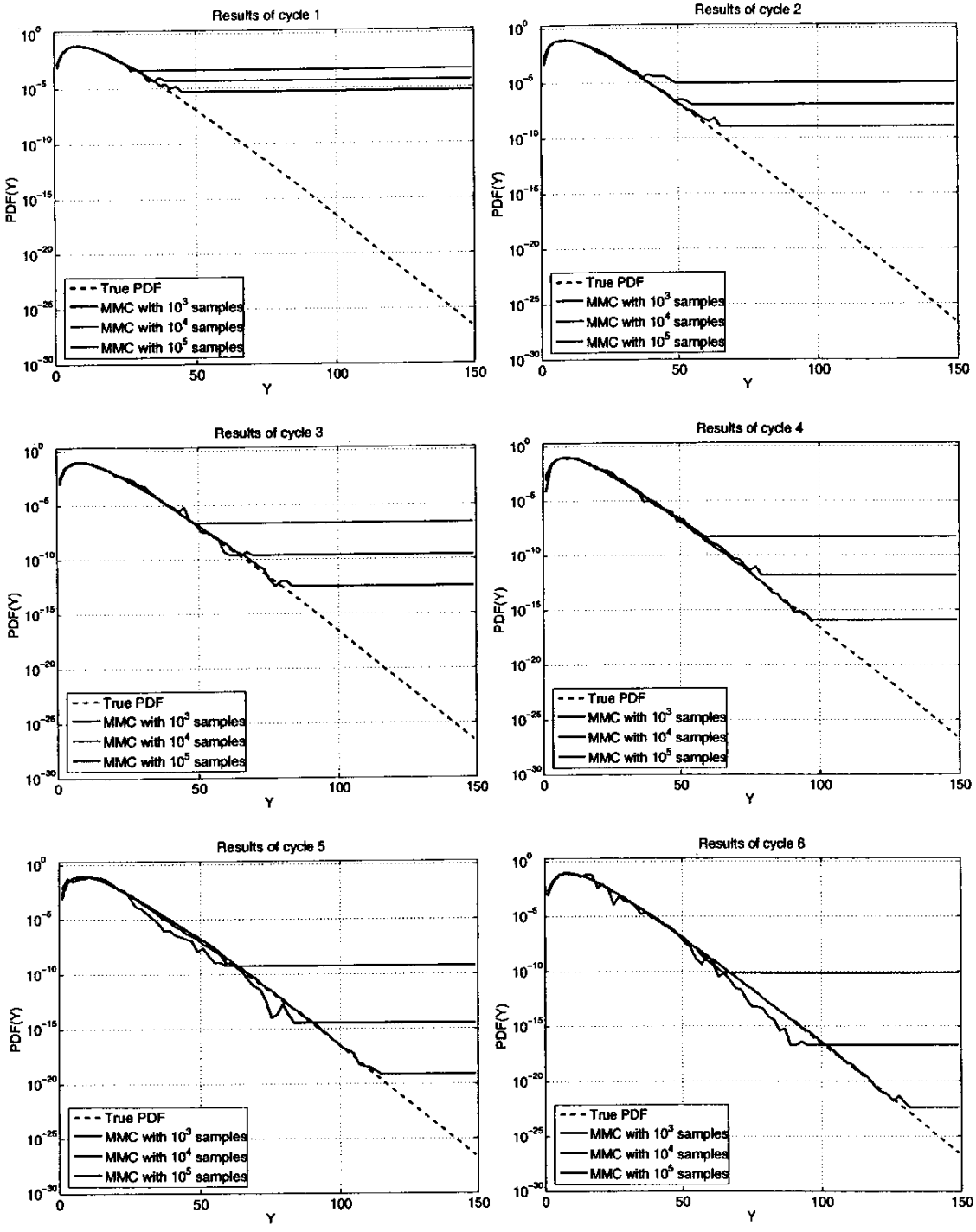
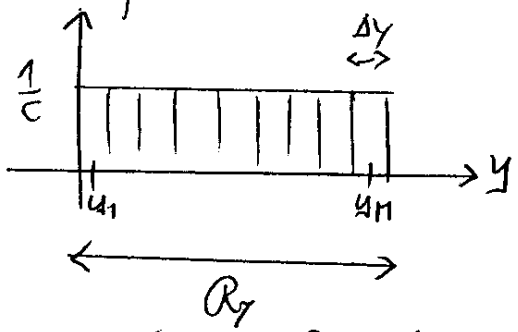


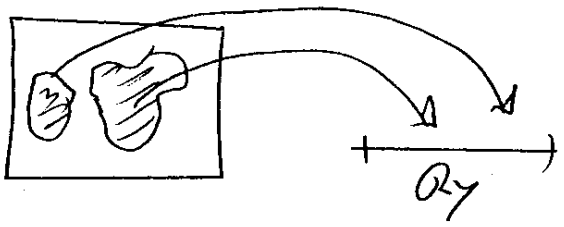
Figure 19: Standard MMC: Check effect of block size $N = 10^3, 10^4, 10^5$.

Caveat

In order to use the FH concept, must restrict the range of Y that want to explore before starting procedure



If, during simulation $g(X_j) \notin R_Y$, discard sample, increase counter N_0 (outliers) and continue.



we will then get an estimate of

$$f_Y(y \mid Y \in R_Y)$$

To get un-conditional estimate, use $\forall y \in R_Y$:

$$f_Y(y) = f_Y(y \mid Y \in R_Y) \cdot \underbrace{P\{Y \in R_Y\}}_{\text{estimate thus as: } \cong \frac{(N - N_0)}{N}} \quad (38)$$

Elements of Convergence Analysis

For simplicity, ignore the bias introduced on visits histogram (34e), by and assume $\widehat{H}^{(n)}(y_i) = N_i^{(n)}/N$ as in (34b).
Then estimator $\widehat{P}_Y^{(n+1)}(y_i)$ in update equation (34c) is unbiased at all cycles, since it is an IS estimator (cfr (15)).

$$E\left[\widehat{P}_Y^{(n+1)}(y_i)\right] \equiv P_Y(y_i) \quad \forall \text{cycles}, \forall \text{bins } i$$

This is a major difference with other flat histogram algorithms, e.g. Wang-Landau, which are biased.

Variance can be obtained from general result for UWIS, eq. (33o):

$$\mathrm{Var}\left[\widehat{P}_Y^{(n+1)}(y_i) \,\middle|\, \underline{\widehat{P}}_n\right] = \frac{P_Y(y_i)}{N} c_n \widehat{P}_Y^{(n)}(y_i)\left(1 - \frac{H_Y^{(n+1)}(y_i)}{c_n \widehat{P}_Y^{(n)}(y_i)}\right)$$

$\frac{P_Y(y_i)}{c_n \widehat{P}_Y^{(n)}(y_i)}$ → neglect for well biased bins

Hence

$$\sigma^2_{\mathrm{MC}}(y_i; n+1) = \mathrm{Var}\left[\widehat{P}_Y^{(n+1)}(y_i)\right] = \mathop{E}_{\underline{\widehat{P}}_n}\left[\mathrm{Var}\left[\widehat{P}_Y^{(n+1)}(y_i) \,\middle|\, \underline{\widehat{P}}_n\right]\right]$$

iterated expectation formula

$$\simeq \frac{P_Y(y_i)}{N} \mathop{E}_{\underline{\widehat{P}}_n}\left[c_n \widehat{P}_Y^{(n)}(y_i)\right] \quad \text{with } c_n = \sum_{i=1}^{M} \frac{P_Y(y_i)}{\widehat{P}_Y^{(n)}(y_i)}$$

If # bins M is large, some incorrelation so that

$$\sigma^2_{\mathrm{MC}}(y_i; n+1) \simeq \frac{P_Y^2(y_i)}{N} \cdot E[c_n] \longrightarrow \frac{P_Y^2(y_i)}{N} \cdot M \quad \text{(FHIS)}$$

when $E[c_n] \longrightarrow M$. Note that relative error is indep. of bin i on explored bins.

SMOOTHED MMC

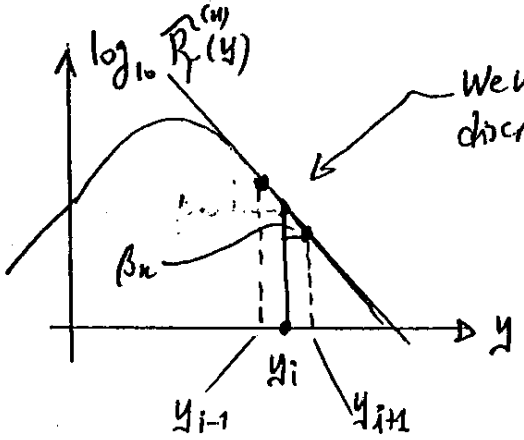
[Berg "Intro to MMC simulations" Fields Inst. Commun. V.26, 2000]

The basic (MMC) algorithm is prone to significant stochastic fluctuations from cycle n to cycle $(n+1)$, which can be duly "smoothed out".

The update equation (34) of MMC is recalled below:

$$\widehat{P}_Y^{(n)}(y) = C_{n-1}\widehat{P}_Y^{(n-1)}(y)\,\widehat{H}_Y^{(n)}(y) \qquad (39)$$

where C_{n-1} is a suitable constant, and the discrete variable y_i has been replaced by a continuous y. When $\widehat{P}_Y^{(n)}(y)$ is plotted versus y in a log scale:



We usually linearly interpolate the discrete points when drawing plot

slope of such a curve at point y_i

Let

$$(40)\quad \boxed{\beta_n(y_i) \triangleq \frac{\log \widehat{P}_Y^{(n)}(y_{i+1}) - \log \widehat{P}_Y^{(n)}(y_i)}{\Delta y}}$$

chosen arbitrarily to the right of y_i. Could choose to the left, or a mix....

$$\overset{(39)}{=} \frac{\log \widehat{P}_Y^{(n-1)}(y_{i+1}) - \log \widehat{P}_Y^{(n-1)}(y_i)}{\Delta y} + \frac{\log \widehat{H}_Y^{(n)}(y_{i+1}) - \log \widehat{H}_Y^{(n)}(y_i)}{\Delta y}$$

i.e.

$$\boxed{\beta_n(y_i) = \beta_{n-1}(y_i) + \underbrace{\frac{\log \widehat{H}_Y^{(n)}(y_{i+1}) - \log \widehat{H}_Y^{(n)}(y_i)}{\Delta y}}_{\triangleq\, \delta_n(y_i)\ \equiv\ \delta_{n,i}}} \qquad (41)$$

slope update equation

$\delta_n(y_i)$ ← WOULD LIKE TO AVOID WILD FLUCTUATIONS OF THIS TERM from MMC to n...

So, instead of update (41), the smoothed MMC uses [Yevick, PTL Feb 03], [N. Rossi, Laurea Thesis, 2005, p. 28]:

$$\beta_n(y_i) = \hat{g}_n(y_i)\left[\beta_{n-1}(y_i) + \delta_n(y_i)\right] + \left[1 - \hat{g}_n(y_i)\right]\beta_{n-1}(y_i) \qquad (42)$$

where $0 < \hat{g}_n(y_i) < 1$ is a suitable "reliability factor" of the slope update (41). In essence: if "proposed" update is reliable ($\hat{g}_n \to 1$) keep it, else keep previous slope. (42) simplifies to:

$$\beta_n(y_i) = \beta_{n-1}(y_i) + \hat{g}_n(y_i)\,\delta_n(y_i) \qquad (43)$$

$$= \beta_{n-1}(y_i) + \frac{\log\left[\hat{H}_\Upsilon^{(n)}(y_{i+1})\right]^{\hat{g}_n(y_i)} - \log\left[\hat{H}_\Upsilon^{(n)}(y_i)\right]^{\hat{g}_n(y_i)}}{\Delta\gamma}$$

and using definition (40) we see that smoothed update (43) implies the following smoothed update of $\hat{P}_\Upsilon^{(n)}$, alternative to (39):

SMOOTHED MMC ⇒

$$\frac{\hat{P}^{(n)}(y_{i+1})}{\hat{P}^{(n)}(y_i)} = \frac{\hat{P}_\Upsilon^{(n-1)}(y_{i+1})}{\hat{P}_\Upsilon^{(n-1)}(y_i)} \cdot \left[\frac{\hat{H}_\Upsilon^{(n)}(y_{i+1})}{\hat{H}_\Upsilon^{(n)}(y_i)}\right]^{\hat{g}_n(y_i)} \qquad (44)$$

Thus is in fact a smoothing both in "time" n, and in bin "space" i.

② Choice of $\hat{g}_n$

Recall that $\hat{H}_\Upsilon^{(n)}(y_i) = \frac{N_i^{(n)}}{N} \leftarrow \text{Binomial}\left(N, H_\Upsilon^{(n)}(y_i)\right)$

So
$$\begin{cases} E\left[\hat{H}_\Upsilon^{(n)}(y_i)\right] = H_\Upsilon^{(n)}(y_i) \triangleq \overline{H} \\ \mathrm{Var}\left[\hat{H}_\Upsilon^{(n)}(y_i)\right] = \frac{H_\Upsilon^{(n)}(y_i)\left(1 - H_\Upsilon^{(n)}(y_i)\right)}{N} \simeq \frac{H_\Upsilon^{(n)}(y_i)}{N} = \frac{\overline{H}}{N} \end{cases}$$

write $\hat{H}_\Upsilon^{(n)}(y_i) = \overline{H} + \Delta H$ and assume fluctuations "small" $\frac{\Delta H}{\overline{H}} \ll 1$

Hence

$$\log\left(\hat{H}_Y^{(n)}(y_i)\right) = \log\left(\bar{H} + \Delta H\right) = \log \bar{H} + \log_{10}\left(1 + \frac{\Delta H}{\bar{H}}\right)$$

$$= \left(\log_{10} e\right) \ln\left(1 + \frac{\Delta H}{\bar{H}}\right) \cong \left(\log_{10} e\right) \frac{\Delta H}{\bar{H}}$$

hence

$$\mathrm{Var}\left[\log\left(\hat{H}_Y^{(n)}(y_i)\right)\right] \cong \frac{\left(\log_{10} e\right)^2}{\bar{H}^2} \cdot \underbrace{\mathrm{Var}[\Delta H]}_{\uparrow \frac{\bar{H}}{N}} = \frac{\left(\log_{10} e\right)^2}{N} \cdot \frac{1}{\bar{H}} \qquad (45)$$

uncorrelated RV's in the assumption of many bins M

From (41) would get:

$$\mathrm{Var}\left[\beta_n(y_i) \mid \beta_{n-1}(y_i)\right] = \mathrm{Var}\left\{\frac{1}{\Delta y} \cdot \left(\log \hat{H}_Y^{n}(y_{i+1}) - \log \hat{H}_Y^{n}(y_i)\right)\right\}$$

if we approximate $\bar{H}$ with actual value $\hat{H}_Y^{(n)}(y_i)$ in (45) get

$$\cong \frac{1}{\Delta y^2} \cdot \frac{\left(\log_{10} e\right)^2}{N} \left\{\frac{1}{\hat{H}_Y^{(n)}(y_{i+1})} + \frac{1}{\hat{H}_Y^{(n)}(y_i)}\right\}$$

So update (41) is reliable when this • is small!

Hence Berg defines:

$$g_n(y_i) \triangleq \left(\frac{1}{H_Y^{(n)}(y_{i+1})} + \frac{1}{\hat{H}_Y^{(n)}(y_i)}\right)^{-1} = \frac{\hat{H}^{(n)}(y_i) \cdot \hat{H}^{(n)}(y_{i+1})}{\hat{H}_Y^{(n)}(y_{i+1}) + \hat{H}^{(n)}(y_i)} \qquad (46)$$

as the (un-normalized) reliability factor of update (41), and normalization is done with respect to past values

$$[0,1] \ni \hat{g}_n(y_i) \triangleq \frac{g_n(y_i)}{\sum_{p=1}^{n} g_p(y_i)} \qquad (47)$$

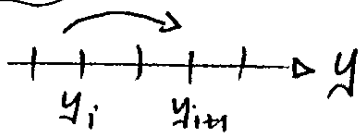
Here we use $\hat{H}^{(n)}(y_i) = \frac{N_i^{(n)}}{N}$ and allow zero values.

▷ When either $\hat{H}^{(n)}(y_i)$ or $\hat{H}^{(n)}(y_{i+1})$ are zero, $g_n(y_i) = \hat{g}_n(y_i) = 0$

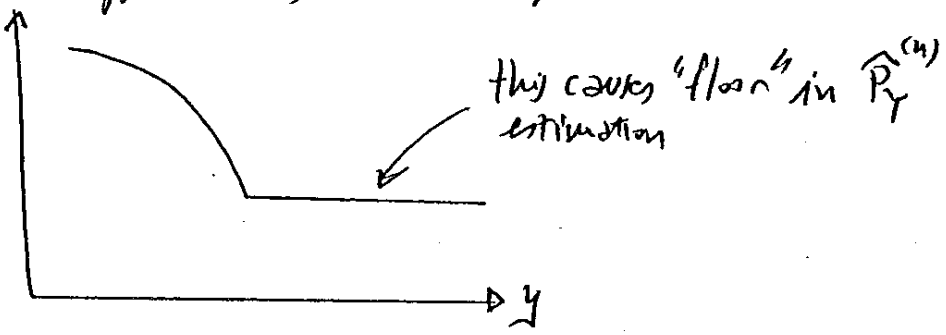
Hence bins in which reliability is zero are updated from (44) as

$$\hat{P}^{(n)}(y_{i+1}) = \hat{P}^{(n)}(y_i) \cdot \frac{\hat{P}_Y^{(n-1)}(y_{i+1})}{\hat{P}_Y^{(n-1)}(y_i)}$$

ie by propagating value of bin on LEFT



(if at previous step that happened also, then $\frac{\cdot}{\cdot} = 1$)



▷ For a generic bin, initially not visited, $g_n(y_i) = 0$ until it is visited (and also it's neighbour $i+1$). If this happens at $\hat{n}$, then $\forall n > \hat{n}$

$$\hat{g}_n^{(y_i)} = \frac{g_n(y_i)}{\sum_{p=\hat{n}}^{n} g_p(y_i)} \longrightarrow 0 \quad \text{as } n \text{ increases.}$$

Hence smoothing (43) converges to a fixed slope β_∞: that is, the updates that count are those in a time window starting from first visit $\hat{n}$.

Aside — Composition of Estimators

Suppose we want to estimate the probability p of a certain event, and we have at our disposal M uncorrelated estimators [RV's] $\hat{P}_1, \dots, \hat{P}_M$, each unbiased ($E[\hat{P}_i] = p$) with variance σ_i^2.

We form an unbiased estimator by forming a linear combination

$$\hat{P} = a_1 \hat{P}_1 + \dots + a_M \hat{P}_M \quad , \quad (C) \quad \boxed{\sum_{i=1}^{M} a_i = 1 \quad , \quad a_i \geq 0}$$

What are the coefficients $\{a_i\}$ that minimize $\mathrm{Var}[\hat{P}] \equiv \sigma_P^2$?

Sol | We look for

$$\min_{\{a_i\}} \sigma_P^2 = \min_{\{a_i\}} \sum_{i=1}^{M} a_i^2 \sigma_i^2$$

(VARIANCE IS SUM OF VARIANCES FOR UNCORRELATED RV'S)

subject to constraint (C)

Form Lagrangian

$$L = \sum_{i=1}^{M} a_i^2 \sigma_i^2 - \lambda \left(\sum_{i=1}^{M} a_i - 1 \right)$$

Impose

$$\frac{\partial L}{\partial a_i} = 0 \quad \Rightarrow \quad 2 a_i \sigma_i^2 - \lambda = 0 \Rightarrow a_i = \frac{\lambda/2}{\sigma_i^2}$$

From constraint (C):

$$1 = \sum_{i=1}^{M} a_i = \frac{\lambda}{2} \sum_{k=1}^{M} \frac{1}{\sigma_k^2} \Rightarrow \frac{\lambda}{2} = \frac{1}{\sum_{J=1}^{M} \frac{1}{\sigma_J^2}}$$

hence optimal coeff. are

(let $g_i \triangleq \frac{1}{\sigma_i^2}$)

$$a_{i\,(opt)} = \frac{\frac{1}{\sigma_i^2}}{\sum_{J=1}^{M} \frac{1}{\sigma_J^2}} = \left[\frac{g_i}{\sum_{J=1}^{M} g_J} \right] \triangleq \tilde{g}_i$$

And min variance is

$$\left(\sigma_P^2\right)_{\min} = \frac{\sum_{i=1}^{M} g_i^2 \sigma_i^2}{\left(\sum_{J=1}^{M} g_J\right)^2} = \frac{1}{\sum_{J=1}^{M} \frac{1}{\sigma_J^2}}$$

($\frac{1}{\sigma_i^2} \equiv g_i$)

hence

$$\boxed{\frac{1}{\left(\sigma_P^2\right)_{\min}} = \sum_{J=1}^{M} \frac{1}{\sigma_J^2} = \sum_{J=1}^{M} g_J}$$

and optimum estimator is

$$\boxed{\hat{P} = \sum_{i=1}^{M} \tilde{g}_i \hat{P}_i}$$

end

Berg's Argument to get (47)

Berg reasoned as follows: Eq. (42) is in fact saying that estimator $\beta_n(y_i)$ is a linear combination of estimators

$$\begin{cases} 1)\ \beta_{n-1}(y_i) \quad \text{with variance } \sigma_{n-1}^2 \triangleq \dfrac{1}{g_{n-1}} \\ 2)\ \beta_{n,0}(y_i) \triangleq \big(\beta_{n-1}(y_i) + \delta_n(y_i)\big) \quad \text{with variance } \sigma_{n,0}^2 \triangleq \dfrac{1}{g_{n,0}} \end{cases}$$

If the two estimators were unbiased and uncorrelated, from previous article the best linear combination would be

$$\beta_n(y_i) = \underbrace{\left[\frac{g_{n,0}}{g_{n,0}+g_{n-1}}\right]}_{\hat{g}_n}\beta_{n,0} + \underbrace{\frac{g_{n-1}}{g_{n,0}+g_{n-1}}}_{1-\hat{g}_n}\beta_{n-1}$$

and its (variance)$^{-1}$

$$\frac{1}{\sigma_n^2} \triangleq g_n = g_{n,0} + g_{n-1}$$

Thus iteration, started at $n=1$ with $g_0 = 0$, gives $g_n = \sum_{j=1}^{n} g_{j,0}$

Hence

$$\hat{g}_n = \frac{g_{n,0}}{\sum_{j=1}^{n} g_{j,0}} \quad \text{which is eq. (47)}$$

□

▷ Critique: Let $X = \beta_{n-1}$

$Y = \delta_n$

Hence $\mathrm{Cov}[\beta_{n,0}\,\beta_{n-1}] = \mathrm{Cov}[(X+Y)X] = \mathrm{Var}[X] + \underbrace{\mathrm{Cov}(X,Y)}_{\approx 0}$

so estimators are (obviously) correlated.

Since (42) can be written as (43), it is clear that

$$\sigma_n^2 = \sigma_{n-1}^2 + \hat{g}_n^2\,\mathrm{Var}[\delta_n]$$

is minimized by $\hat{g}_n = 0$! (do not update...)

Berg's argument should be interpreted as follows:
From (41), we get a recursion

$$\beta_{n,i} = \beta_{n-1,i} + \delta_{n,i} \qquad n = 1, 2, \ldots$$

From which we conclude

(41b) $$\beta_{n,i} = \sum_{J=1}^{n} \delta_{J,i}$$ where each δ_J is an unbiased estimator of β (the slope). These are uncorrelated.

Hence which is the best way to combine them? Certainly not (41b)!
From what we learned in the Aside, we know best linear combination is

(41c) $$\hat{\beta}_{n,i} = \sum_{J=1}^{n} \hat{g}_{J,i}\, \delta_{J,i}$$ where $$\hat{g}_{n,i} \triangleq \frac{g_{n,i}}{\sum_{J=1}^{n} g_{J,i}}$$ (41d)

and $$g_{J,i} \triangleq \frac{1}{\sigma^2_{J,i}} = \frac{1}{\mathrm{Var}[\delta_{J,i}]}$$

Such a scheme would require rescaling all weights $\hat{g}_{Ji}$, $J = 1, \ldots, n$ at each nth iteration. Berg's scheme instead does that incrementally.

Example

We next compare the iterations for the standard MMC and the smoothed MMC, for the same problem already discussed at page 32.

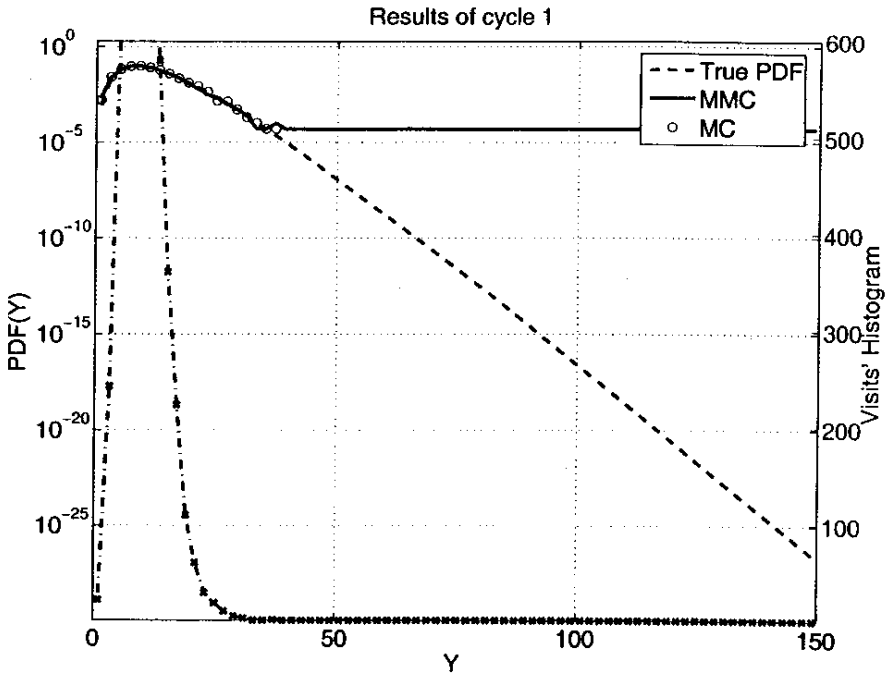


Figure 7: Standard MMC, N=10000.

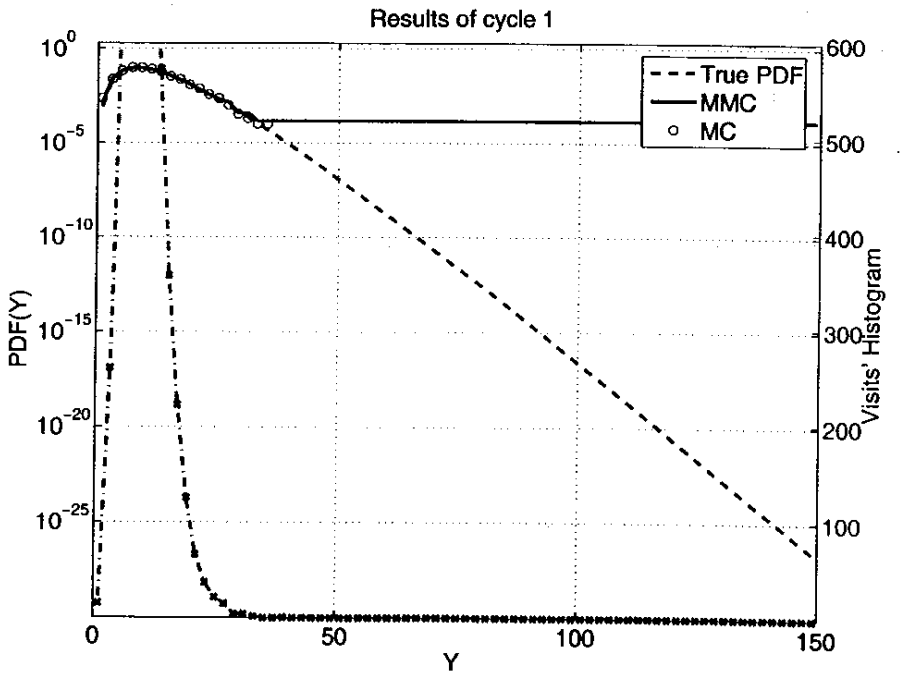


Figure 8: Smoothed MMC, N=10000.

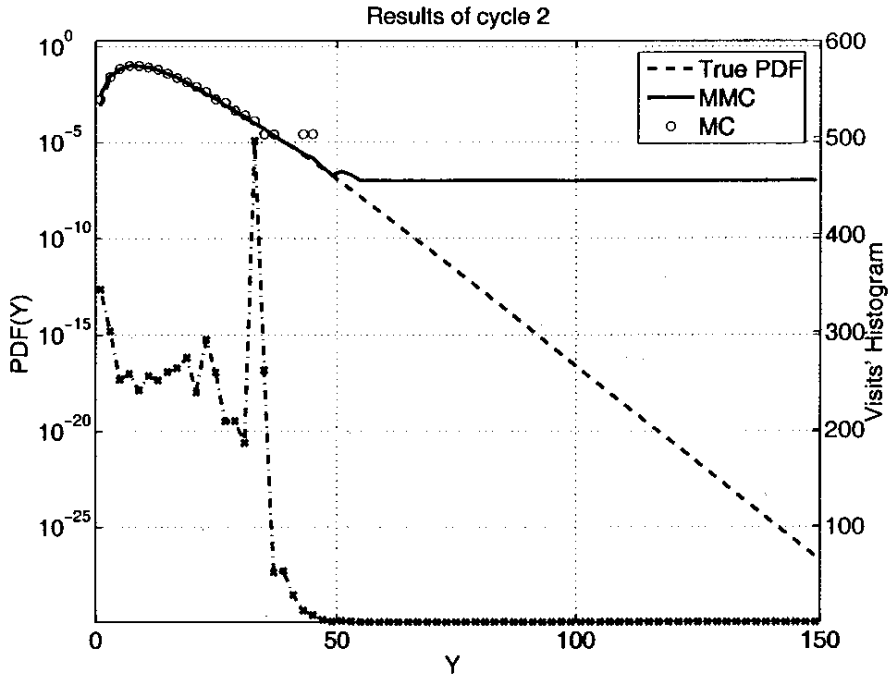


Figure 9: Standard MMC, N=10000.

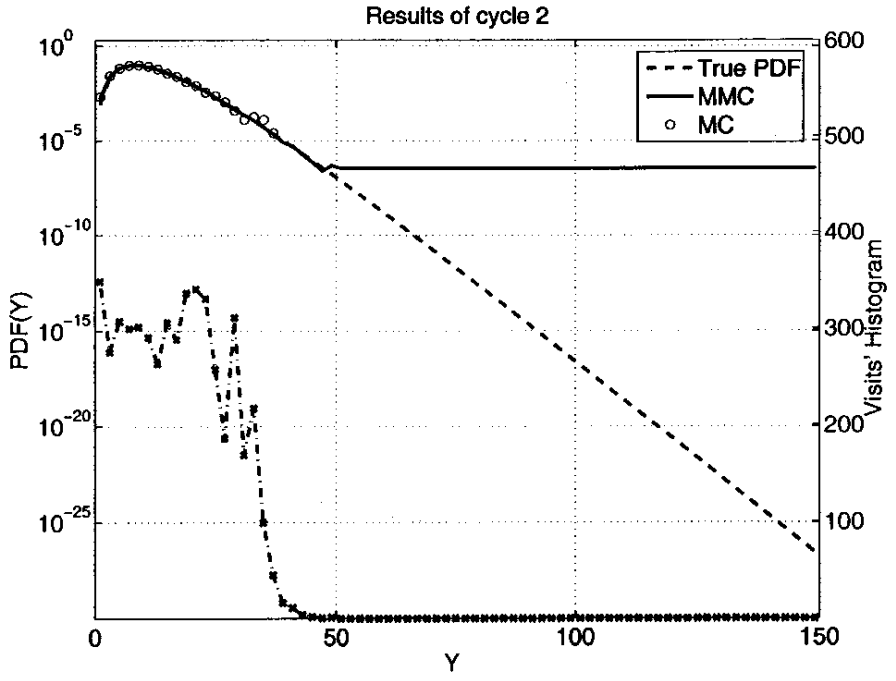


Figure 10: Smoothed MMC, N=10000.

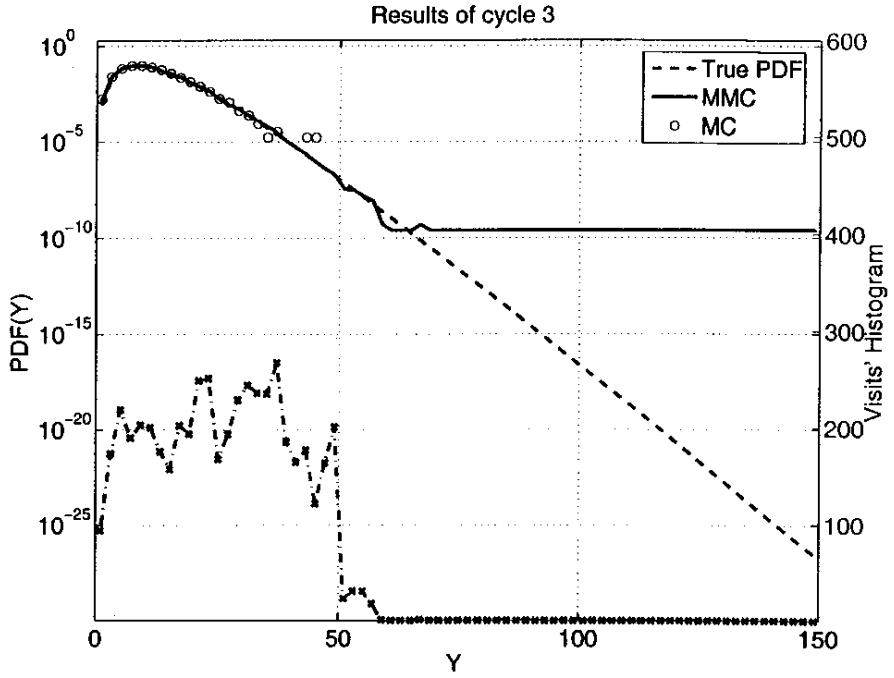


Figure 11: Standard MMC, N=10000.

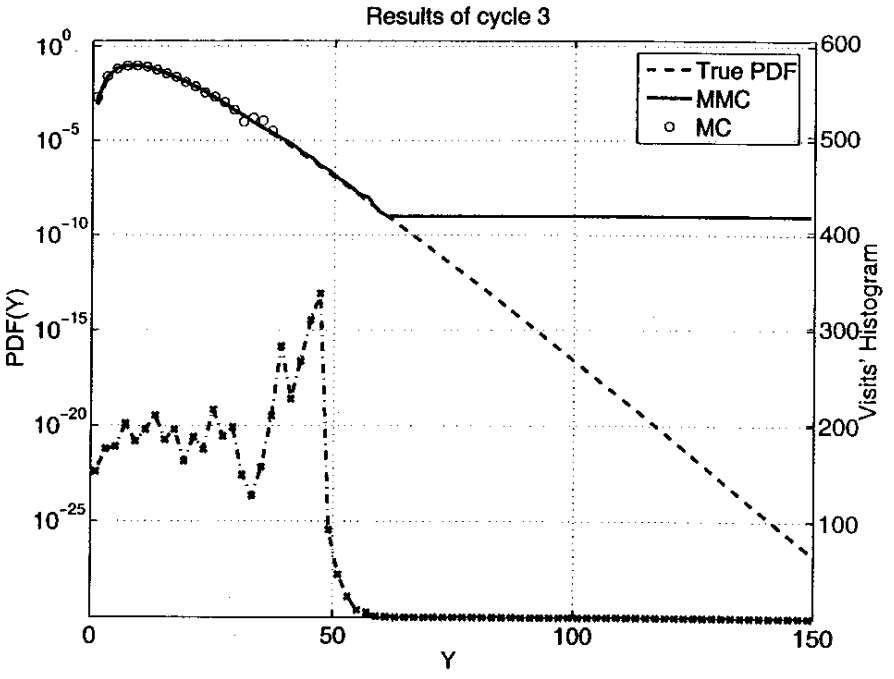


Figure 12: Smoothed MMC, N=10000.

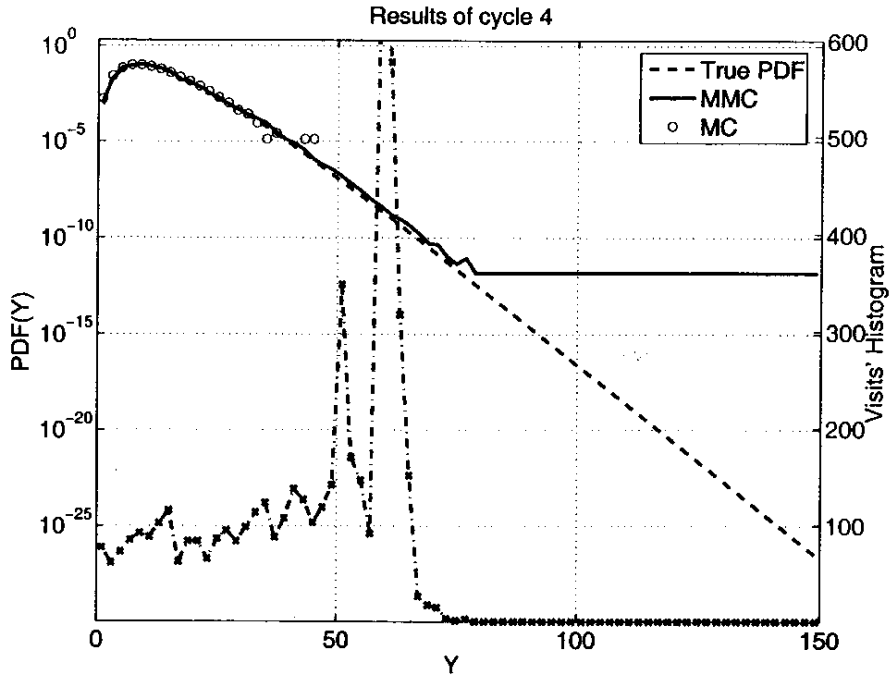


Figure 13: Standard MMC, N=10000.

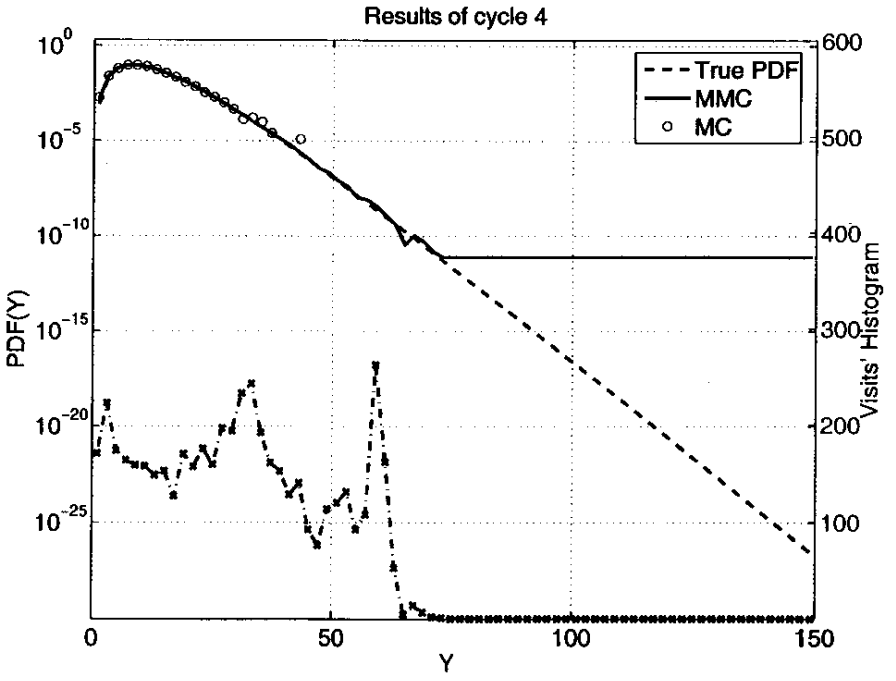


Figure 14: Smoothed MMC, N=10000.

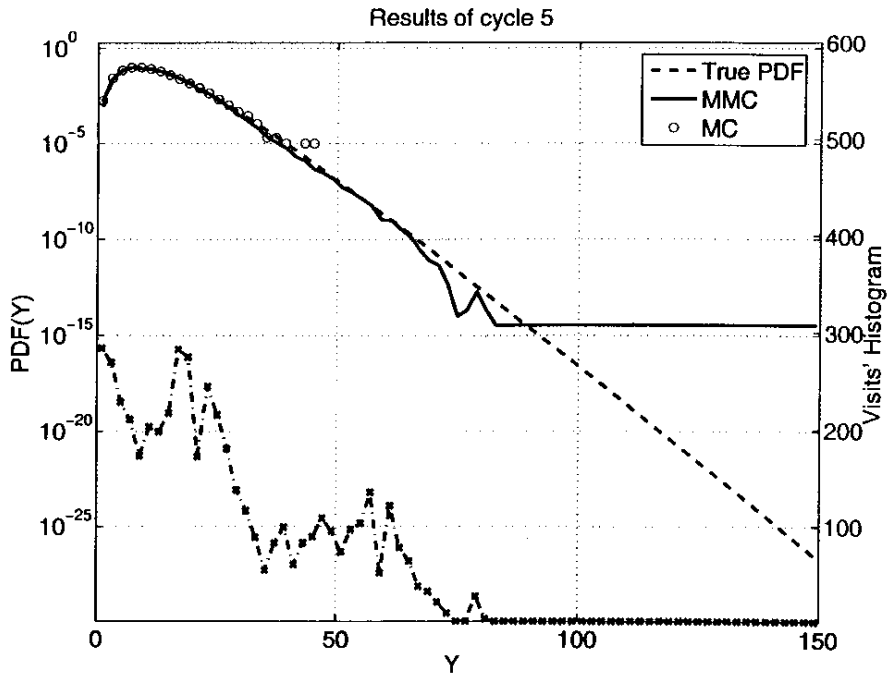


Figure 15: Standard MMC, N=10000.

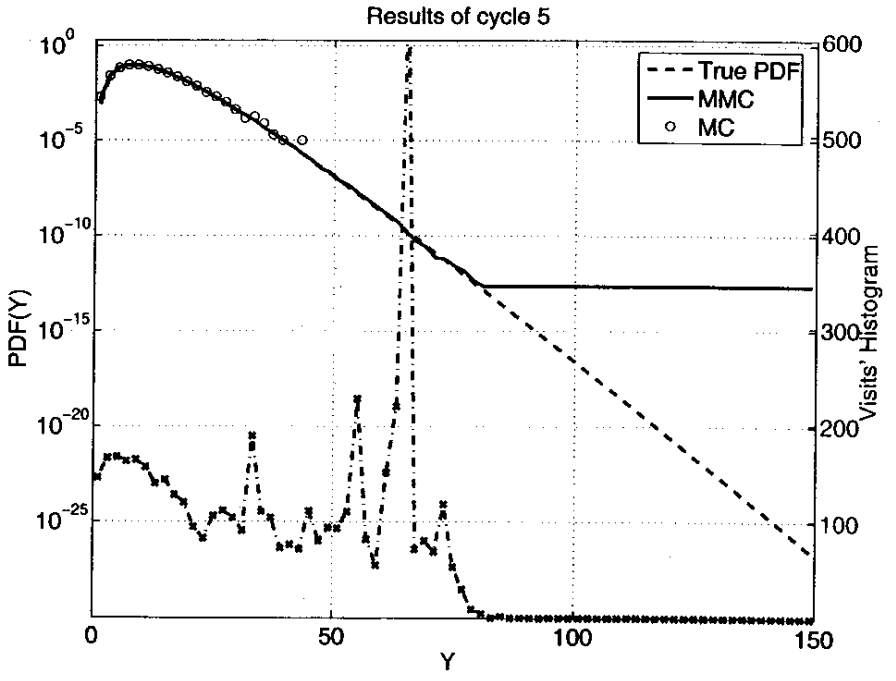


Figure 16: Smoothed MMC, N=10000.

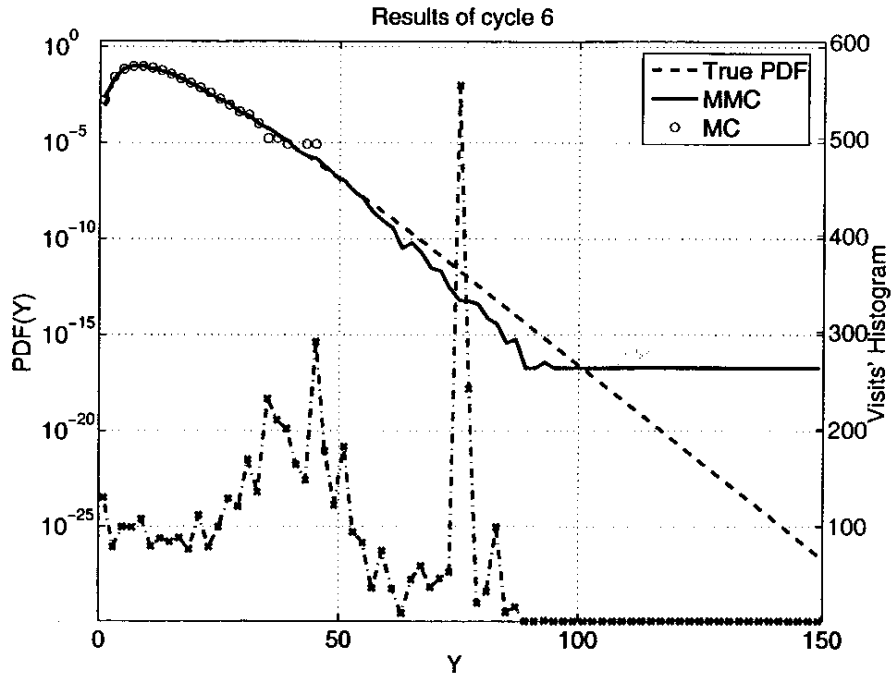


Figure 17: Standard MMC, N=10000.

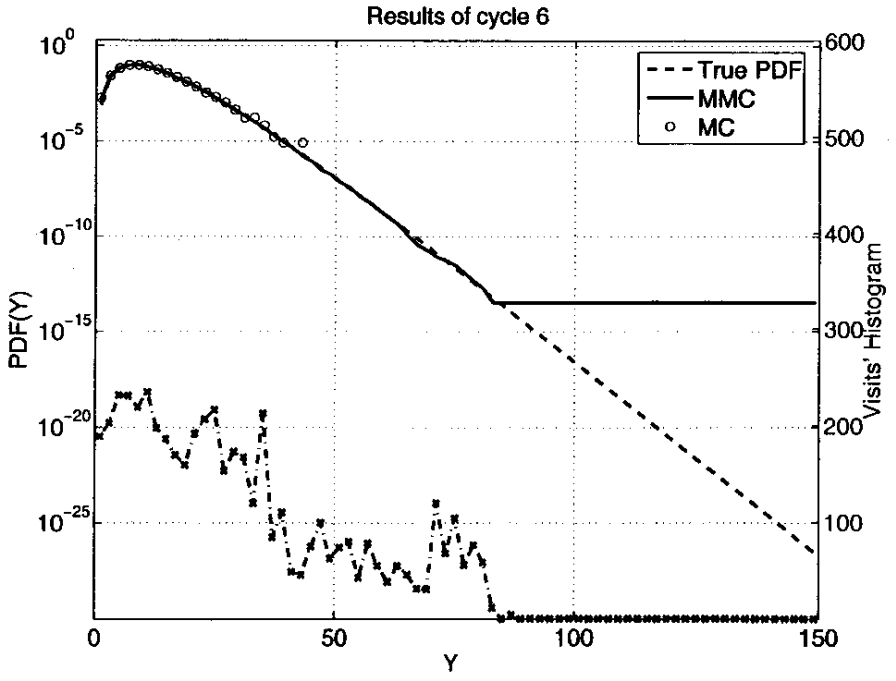


Figure 18: Smoothed MMC, N=10000.

Effect of cycle size N

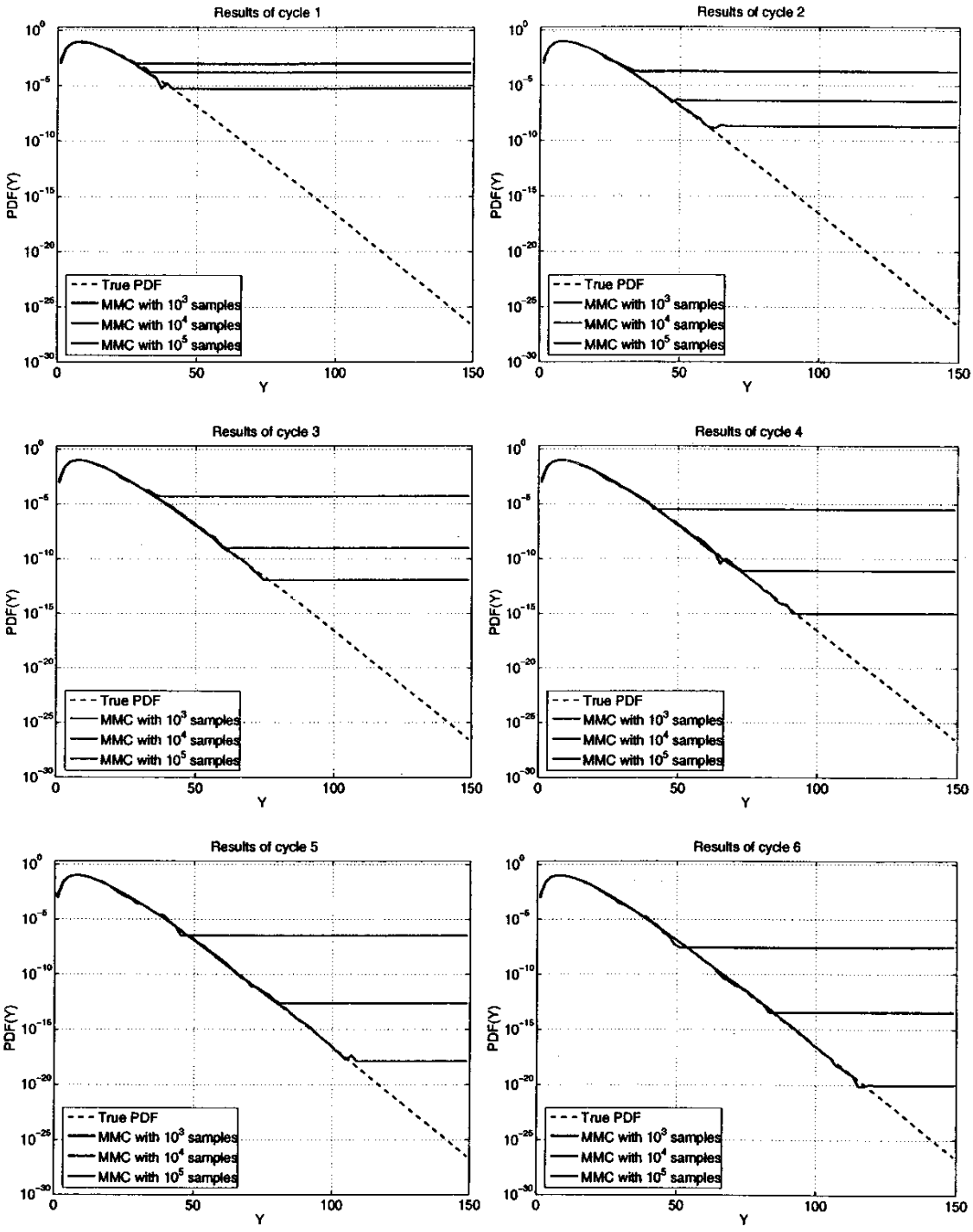


Figure 20: Smoothed MMC: Check effect of block size $N = 10^3, 10^4, 10^5$.

It is clear that smoothed MMC "dips" less than standard MMC, but with greater smoothness.

Exercise: prove that best linear combination $\hat{\beta}_n$ in (41c) with best weights $\hat{g}_{n,i}$ in (41d) has the following update:

$$\hat{\beta}_{n,i} = \hat{\beta}_{n-1,i}\,(1-\hat{g}_{n,i}) + \delta_{n,i}\,\hat{g}_{n,i} \qquad (41e)$$

which, as in (44), translates into the update

$$(41f) \qquad \frac{\hat{P}_Y^{(n)}(y_{i+1})}{\hat{P}_Y^{(n)}(y_i)} = \left[\frac{\hat{P}_Y^{(n-1)}(y_{i+1})}{\hat{P}_Y^{(n-1)}(y_i)}\right]^{(1-\hat{g}_{n,i})} \left[\frac{\hat{H}_Y^{(n)}(y_{i+1})}{\hat{H}_Y^{(n)}(y_i)}\right]^{\hat{g}_{n,i}}$$

Note: Trouble is that the estimate

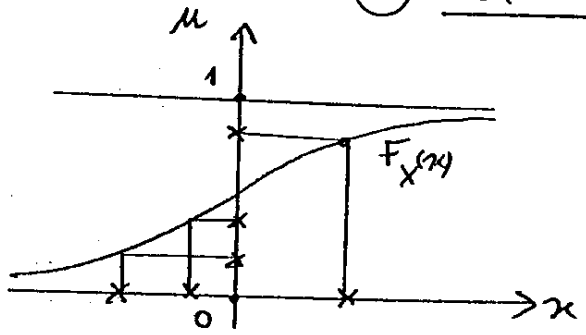
$$\widehat{\mathrm{Var}}\left[\log\left(\hat{H}_Y^{(n)}(y_i)\right)\right] \overset{(45)}{=} \frac{(\log_{10} e)^2}{N}\,\frac{1}{\hat{H}_Y^{(n)}(y_i)}$$

is very bad when $\hat{H}_Y^{(n)}(y_i) \gg E[\hat{H}_Y^{(n)}(y_i)] \equiv \overline{H}$ (over-estimate of histogram of visits), so that (41f) gives much worse estimation than Berg's algorithm (44), when $g_{n,i} = \frac{1}{\mathrm{Var}[\delta_{n,i}]}$ is computed as per (46).

∘ Generating $\{X_i\}$ with given $f_X^*(x)$

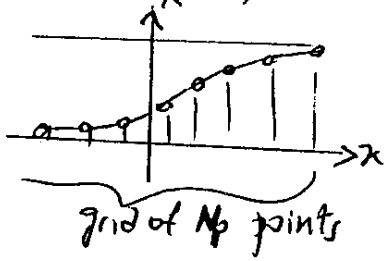
So far we have assumed we know how to generate samples $\{X_1, \dots, X_N\}$ from a known PDF $f_X^*(x)$. Let's review pro's and con's of available generation methods

(1) CDF method



If take $U \sim$ uniform$[0,1]$, and compute $Z = F_X^{-1}(U)$ then Z has CDF $F_X(\cdot)$!!

So trick is to build a table with $F_X(x)$ values (and interpolate in-between).



In our case

$$F_X^{(n)}(x) = \int_{-\infty}^{x} f_X^{(n)}(x')\,dx' = \int_{-\infty}^{x} \frac{f_X(x')}{\frac{C^{(n)}}{\Delta y}\hat{P}_Y^{(n-1)}(g(x'))}\,dx' \qquad (48)$$

At each time step n, need "off-line" evaluation of $y = g(x')$ for many points x' to get integral. When $\underline{x}' \in \mathbb{R}^d$, then need too many "simulations" $\boxed{y = g(\underline{x}')}$ to build multidimensional table of $F_{\underline{X}}^{(n)}(\underline{x})$: $(N_p)^d$!!

(2) Rejection method [Papoulis p. 229]

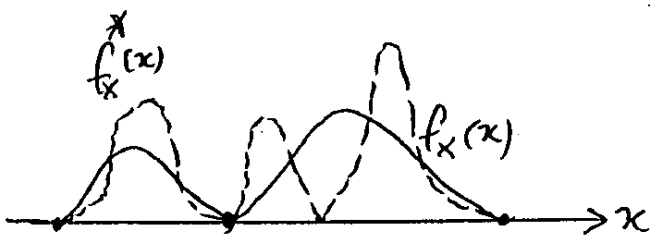
Want to generate an RV X^* with desired PDF $f_X^*(x)$, starting from a machine that generates RV X with PDF $f_X(x)$.

The idea is to choose a suitable event M, such that

$$f_X^*(x) = f_X(x|M) \tag{49}$$

From meaning of conditioning, this is obtained by generating a sample from $f_X(\cdot)$, and keeping it if M occurs, and rejecting it otherwise.

[Clearly from (49) can only generate PDFs f_X^* which are zero at specific $\hat{x}$ whenever $f_X(\hat{x})$ is zero, since in that case also $f_X(\hat{x}|M) = 0$.



so we can "warp", but cannot touch the zeros of $f_X(x)$!

Suppose there exists a $d < \infty$, such that $\forall x$

$$\frac{1}{W(x)} \triangleq \frac{f_X^*(x)}{f_X(x)} \leq d < \infty \tag{50}$$

(warping cannot be larger than d for all x). Then

Rejection Theorem

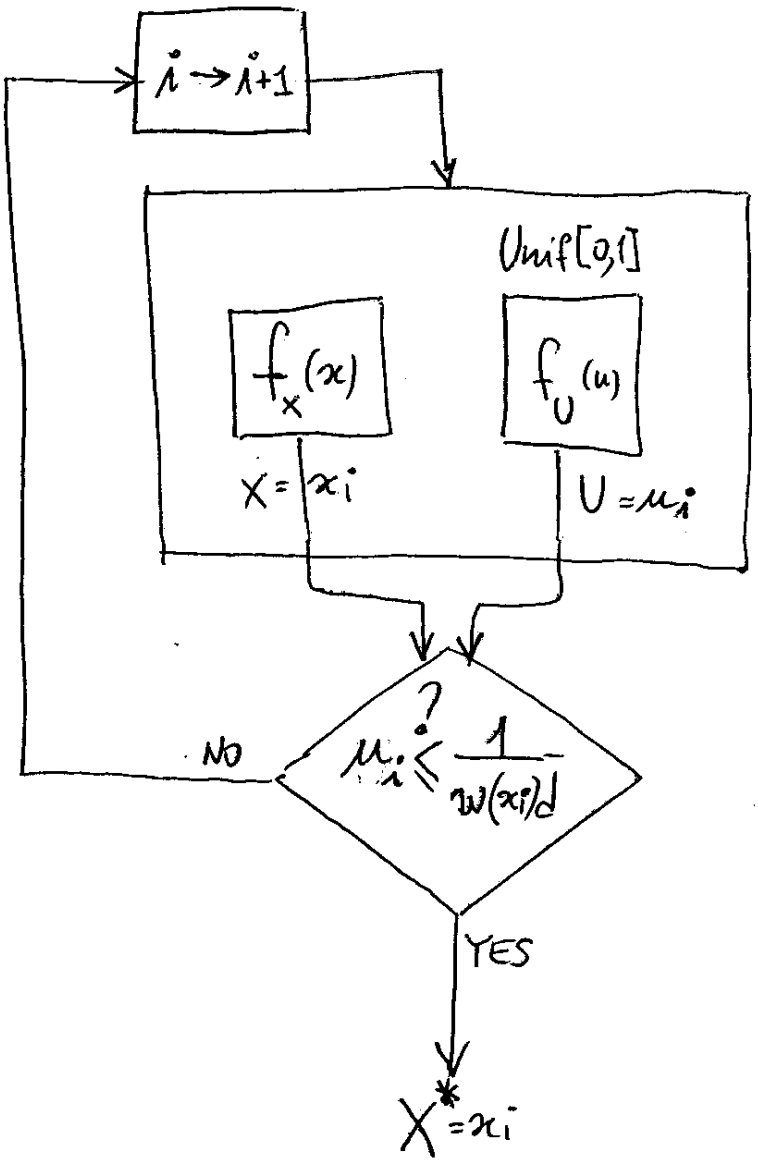
Let $U \sim$ uniform$[0,1]$ RV, independent of X having assigned $f_X(x)$.

Then event

$$M = \left\{ U \leq \frac{1}{W(x) \cdot d} \right\} \tag{51}$$

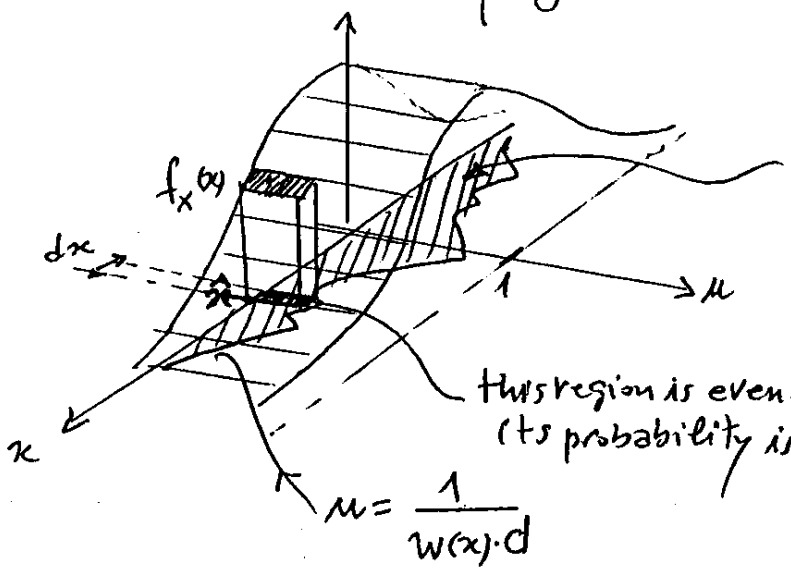
is such that $f_X^*(x) = f_X(x|M)$

Rejection Machine



Proof By indep.

$$f_{XU}(x,u) = \begin{cases} 1 \cdot f_X(x) & \text{for } 0 \le u \le 1 \\ 0 & \text{else} \end{cases}$$



Region $\left\{ u \le \frac{1}{w(x) \cdot d} \right\}$ corresponds to event $\mathcal{H}$

this region is event $\{\hat{x} < X < \hat{x} + dx, \mathcal{H}\}$: its probability is volume under f_{XU}.

From figure, understand that

$$P\{\hat{x} < X < \hat{x} + dx, \mathcal{H}\} = \underbrace{dx \cdot \left[\frac{1}{w(\hat{x}) \cdot d}\right]}_{\text{basis}} \cdot \underbrace{f_X(\hat{x})}_{\text{height}} \quad (52)$$

Adding up all these infinitesimal volumes $\forall \hat{x} \in \mathbb{R}$ gives

$$\underbrace{P\{-\infty < X < \infty, \mathcal{H}\}}_{P(\mathcal{H})} = \iint_{\left\{u \le \frac{1}{c \cdot w(x)}\right\}} f_{XU}\, dx\, du = \int_{-\infty}^{\infty} \frac{f_X(x)}{d \cdot w(x)}\, dx$$

i.e.

$$\boxed{P(\mathcal{H}) = \frac{1}{d} \int_{-\infty}^{\infty} f_X^*(x)\, dx = \frac{1}{d}} \quad (53)$$

and also

$$f_X(\hat{x}|\mathcal{H})\, dx = \frac{P\{\hat{x} < X < \hat{x} + dx, \mathcal{H}\}}{P(\mathcal{H})} = \frac{P\{\hat{x} < X < \hat{x} + dx | \mathcal{H}) P(\mathcal{H})}{P(\mathcal{H})}$$

by definition of conditional density

$$\overset{(52)}{=} \frac{f_X(\hat{x})}{w(\hat{x}) \cdot d}\, dx = f_X^*(\hat{x})\, dx$$

$1/d$

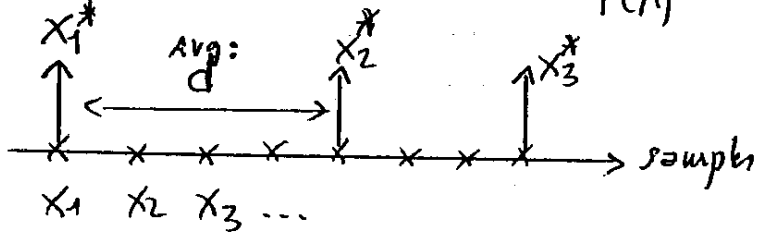
Hence $f_X(x|\mathcal{H}) = f_X^*(x) \quad \forall x$ □

Note From (50), $\forall x \quad f_x^*(x) \le d\, f_x(x)$

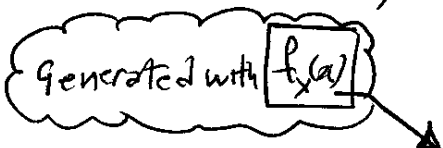
hence $\int \underbrace{f_x^*(x)\,dx}_{1} \le d \int \underbrace{f_x(x)\,dx}_{1} \Rightarrow \boxed{d \ge 1}$

Note: Samples from X are accepted with (success) probability $P(\pi)$ and rejected otherwise. Number of generated samples to get one success (ie one sample of X^*) is thus geometrically distributed, with mean

$$E[\text{Trials up to 1 success}] = \frac{1}{P(\pi)} = d$$



Note: in MMC,

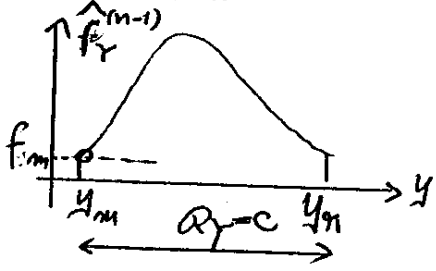


$$f_X^{(n)}(x) = \frac{f_X(x)}{\frac{C}{\Delta Y}\hat{P}_Y^{(n-1)}(g(x))} \Rightarrow W(x) = \frac{\overset{(n-1)}{C}\overset{(n-1)}{\hat{P}}_Y(g(x))}{\Delta Y}$$

So, for each proposal x_J, must evaluate $Y_J = g(x_J)$, then accept it if [from (51)] $U \le \frac{1}{W(x_J)\,d}$ and reject it otherwise.

d is the largest value of $\frac{f_x^*}{f_x} = \frac{1}{C \hat{f}_Y^{(n-1)}(g(x))}$

ie $d = \frac{1}{c \cdot f_m} = \frac{1}{R_Y f_m}$ $\quad c \approx C$



Ex we explore range $y \in [0, 10]$ and PDF goes down to $f_m = 10^{-20}$

Hence

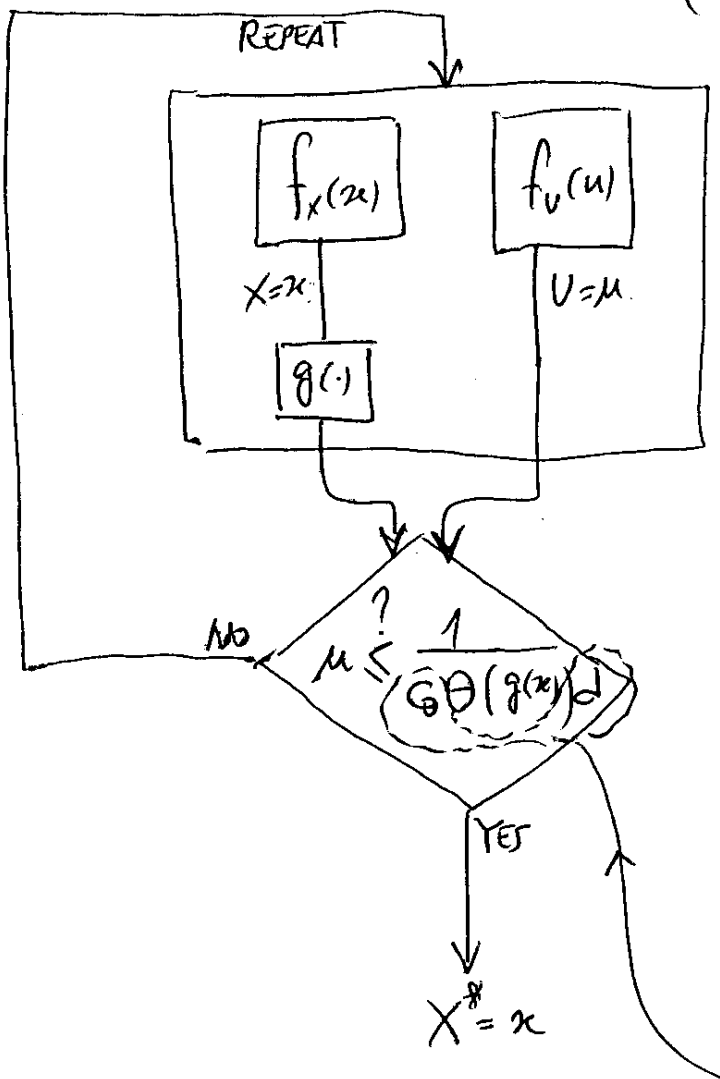
$$\text{Avg \# Trials for 1 sample} = d = \frac{1}{10 \cdot 10^{-20}} = 10^{19}$$

not really practical !

We need a more efficient generation method !

Rejection Machine for UWIS

Weight is $w(x) = C_\Theta \theta(y_i)$ if $x \in D_i$
(ie if $g(x) \approx y_i$)



▷ to sample from f^*, it is not necessary to know domains $\{D_j\}_{j=1}^M$ in state space: just need to verify, for each proposal $g(x)$, at which domain D_i it belongs, and use the appropriate weight for it: $w(x) = C_\Theta \theta(y_i)$

▷ Can absorb normalization constant inside constant $d' \equiv C_\Theta d$

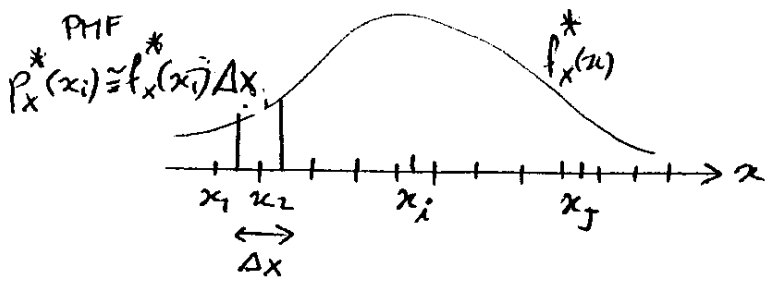
▷ How do we choose d'?

It should be the smallest number satisfying (50): $\forall x \in \Gamma$

$$\frac{1}{w(x)} = \frac{1}{C_\Theta \theta(y_i)} \leq d \quad \Rightarrow \quad \boxed{d' \geq \frac{1}{\theta(y_i)} \quad \forall i=1,\dots,M}$$

③ Markov chain Monte Carlo (MCMC) method [Metropolis 195 Hastings 197

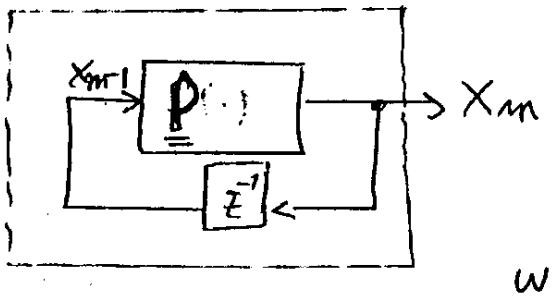
To simplify treatment, suppose we discretize the state space, and we thus know PMF $p_X^*(x_i)$. that we would like to get samples from.



The idea is to get our samples from a "stochastic machine" that outputs a "memoryless" sequence

$$\{X_m,\ m \geq 1\}$$

i.e. a discrete-time markov chain (DTMC),



whose STEADY STATE distribution $\underline{\pi}$ coincides with the desired $\{p_X^*(x_i)\}$

A DTMC is characterized by its transition matrix $\underline{P} = \{P_{iJ}\}$, with $P_{iJ} = P\{X_m = x_i \mid X_{m-1} = x_J\}$ for all states x_i, x_J

The steady-state distribution is eigenvector of $\underline{P}$ with eigenvalue 1 :

$$\underline{P}\,\underline{\pi} = \underline{\pi} \tag{54}$$

In our case we look for $\underline{P}$, such that $\underline{\pi} \equiv \underline{p}_X^* = \begin{bmatrix} p_X^*(x_1) \\ p_X^*(x_2) \\ \vdots \end{bmatrix}$

We clearly would like $\underline{P}$ to have only one $\underline{\pi}$, and that the PMF of X_m, call it $\underline{P}(m) = \begin{bmatrix} P\{X_m = x_1\} \\ P\{X_m = x_2\} \\ \vdots \end{bmatrix}$ converges to $\underline{\pi}$,

i.e. $$\lim_{m \to \infty} \underline{P}(m) = \underline{\pi} \equiv \underline{p}_X^* \tag{55}$$ i.e. we require DTMC be ergodic.

Clearly, there are infinitely many matrices $\underline{\underline{P}}$ that satisfy (54) for a desired $\underline{\pi} \equiv \underline{p}_X^*$.

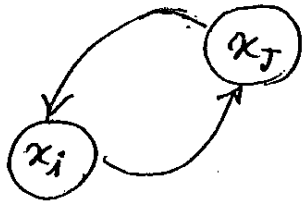
A simple solution is found by requiring that the DTMC be <u>time-reversible</u>.

In fact a <u>necessary and sufficient condition</u> for time reversibility is that, $\forall x_i, x_j \in S$ (state space)

$$\pi_i P_{ji} = \pi_j P_{ij} \tag{56}$$

$\pi_i P_{ji}$: $P\{X_m = x_j, X_{m-1} = x_i\}$

$\pi_j P_{ij}$: $P\{X_m = x_i, X_{m-1} = x_j\}$



This means that, at equilibrium, the prob. of being at x_i at time $m-1$, and moving to x_j at time m must equal the reverse transition.

Equations (56) are how we determine all our unknowns $\{P_{ij}\}$.

A time-reversible chain is Ergodic. [Leon Garcia "Probability..." Addison Wesley 1994]

▷ To be useful as a sample generator in our multicanonical iterative scheme, the MCMC should have a mechanism to find the proper matrix $\underline{\underline{P}}^{(m)}$ at each MMC iteration m, in order to synthesize samples from the desired $\underline{p}_X^{(m)}$, i.e. should determine $\underline{\underline{P}}^{(m)}$ such that

$$\boxed{\underline{\underline{P}}^{(m)} \cdot \underline{p}_X^{(m)} = \underline{p}_X^{(m)}}$$

This is achieved by the Hastings MCMC method (1970)

◦ Hastings MCMC

Given a desired PMF $\underline{\pi}$, Hastings' method finds the required transition probabilities p_{ij} (such that $\underline{\underline{P}}\,\underline{\pi} = \underline{\pi}$) of a reversible DTMC as follows.

▷ Start with any transition matrix $\underline{\underline{Q}} = \{q_{ij}\}$ ("candidate")

I call it the "explorer"

▷ Pick two states, (x_i, x_j):

It may happen that either

$$\text{a)} \quad \underbrace{\pi_i q_{ji}}_{i \to j} > \underbrace{\pi_j q_{ij}}_{j \to i}$$

or that

$$\text{b)} \quad \pi_i q_{ji} < \pi_j q_{ij}$$

In case a) should DECREASE number of transitions $i \to j$ (and leave untouched those $j \to i$) so as to get equality as in (56).

How?

Every time explorer chain "proposes" a move $i \to j$, we "accept" it with probability α_{ji}, and "refuse" it ($i \to i$) otherwise.

This is chosen so that (56) is satisfied:

$$\underbrace{\pi_i q_{ji} \alpha_{ji}}_{P\{X_{m-1} = x_i, X_m = x_j\}} = \pi_j q_{ij} \implies \alpha_{ji} = \frac{\pi_j q_{ij}}{\pi_i q_{ji}} < 1$$

To keep transitions $j \to i$ unchanged, use acceptance prob.

$$\alpha_{ij} = 1, \text{ so that } P\{X_{m+1} = x_j, X_m = x_i\} = \pi_j \cdot q_{ij} \cdot 1$$

In case b) use

$$\alpha_{ij} = \frac{\pi_i q_{ji}}{\pi_j q_{ij}} < 1 \quad \text{and} \quad \alpha_{jj} = 1$$

In Summary:

For every transition $j \to i$ "proposed" by the candidate (explorer) chain, use an acceptance probability

$$\alpha_{ij} = \min\left[\frac{\pi_i q_{ji}}{\pi_j q_{ij}}, 1\right]$$

so that have synthesized a reversible DTMC having transition probabilities

$$P_{ij} = q_{ij}\alpha_{ij} \quad \odot$$

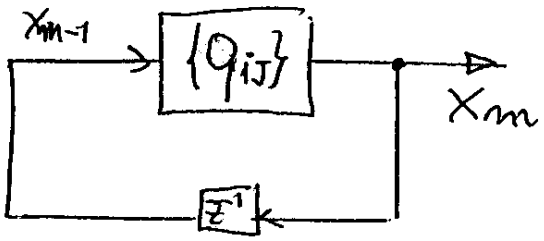
that satisfy (56) and thus produce samples x_m whose asymptotic ($m \to \infty$, steady state) distribution is $\underline{\pi}$, the desired one.

Implementation

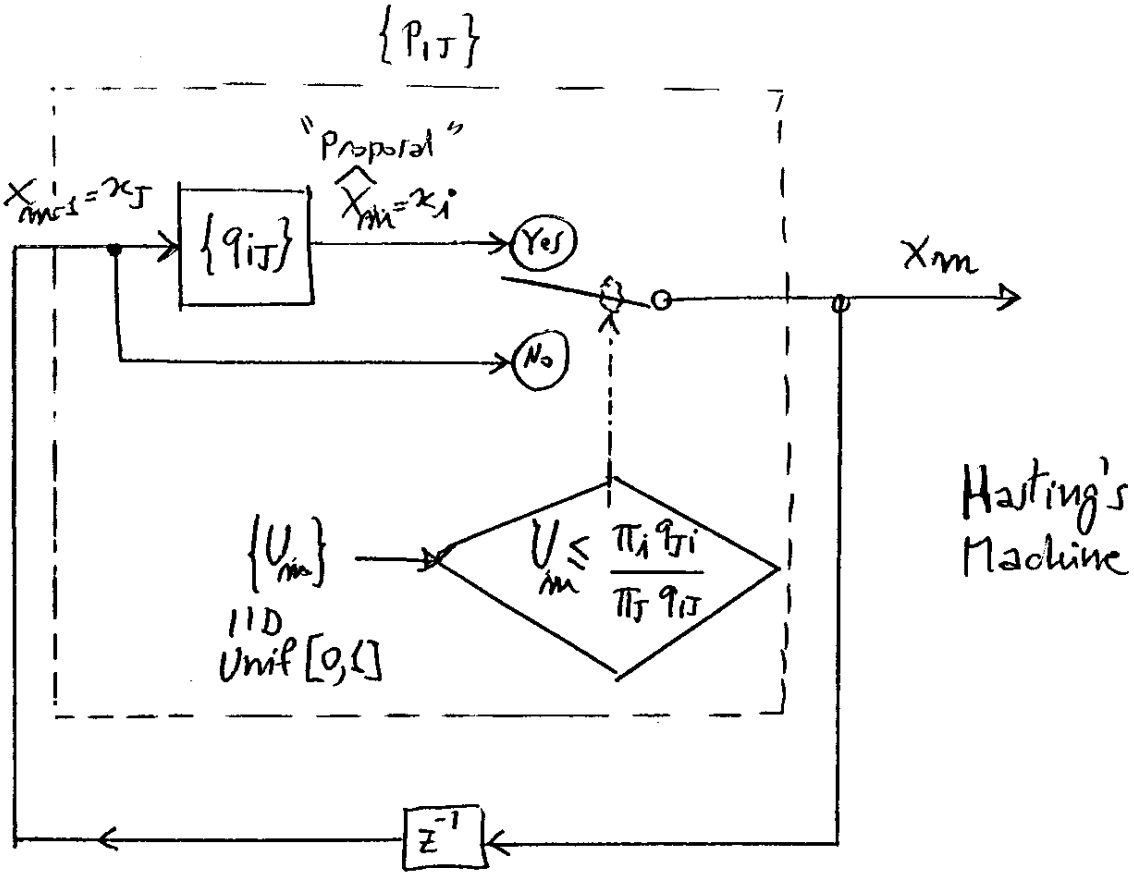
$\odot$ More correctly:

$$P_{ij} = \begin{cases} q_{ij}\alpha_{ij} & \text{if } i \neq j \\ q_{jj} + \sum_{k \neq j} q_{jk}(1-\alpha_{kj}) & \text{if } i = j \end{cases}$$

We start with a given DTMC machine $Q = \{q_{ij}\}$



and modify it as follows:



▷ Metropolis

is a special case of Hasting MCMC, in which $q_{iJ} = q_{Ji}$ so that

$$\alpha_{iJ} = \min\left[\frac{\pi_i}{\pi_J}, 1\right]$$

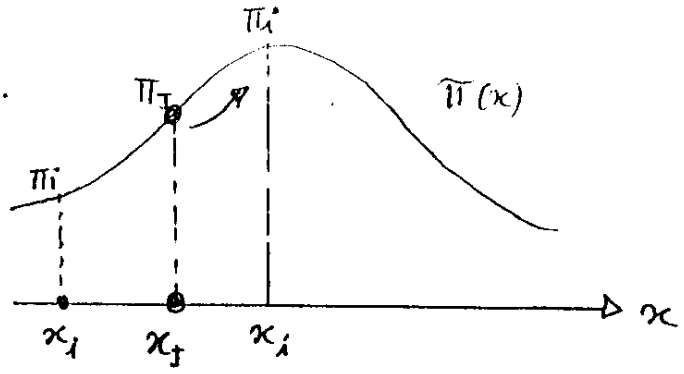
Apply to (MMC):

$$\alpha_{iJ} = \min\left[\frac{f_X(x_i)}{\cancel{\hat{P}_Y^{(m-1)}(g(x_i))}} \cdot \frac{\cancel{\hat{P}^{(m-1)}(g(x_J))}}{f_X(x_J)}, 1\right]$$

※ The biggest advantage of Metropolis (and Hastings) is that normalizing constants cancel out; it is enough to know $\pi \propto f_X^*(\cdot)$ up to a normalizing constant.

Intuition on Metropolis

We are at time m at x_J: $X_m = x_J$.
The explorer makes a proposal x_i.



- If $\pi_i > \pi_J$ proposal always accepted ($\alpha_{ij}=1$)
- Else proposal accepted with prob. $\alpha_{ij} = \frac{\pi_i}{\pi_J}$ (odds ratio)

▷ So upward moves are encouraged ⇒ explore more frequently modal range of $\pi(x)$!

▷ A big jump to a lower probability $\pi_i \ll \pi_J$ is most of the times rejected.

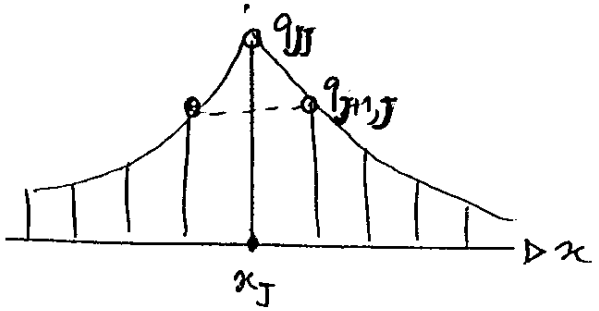
This is how the correct $\pi(x)$ is actually sampled.

More on the "explorer"

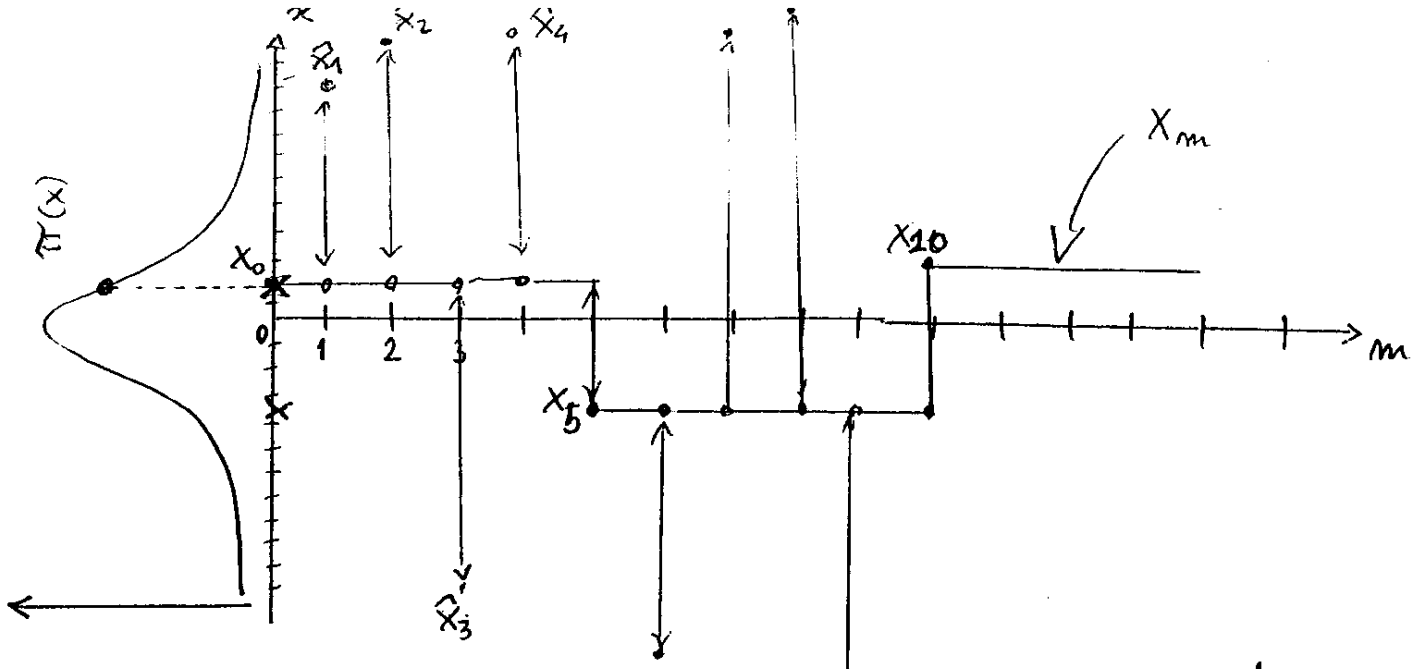
A possible way to implement $q_{iJ} = q_{Ji} \quad \forall ij$ is to choose

$$q_{iJ} = P_Z(|i-J|)$$

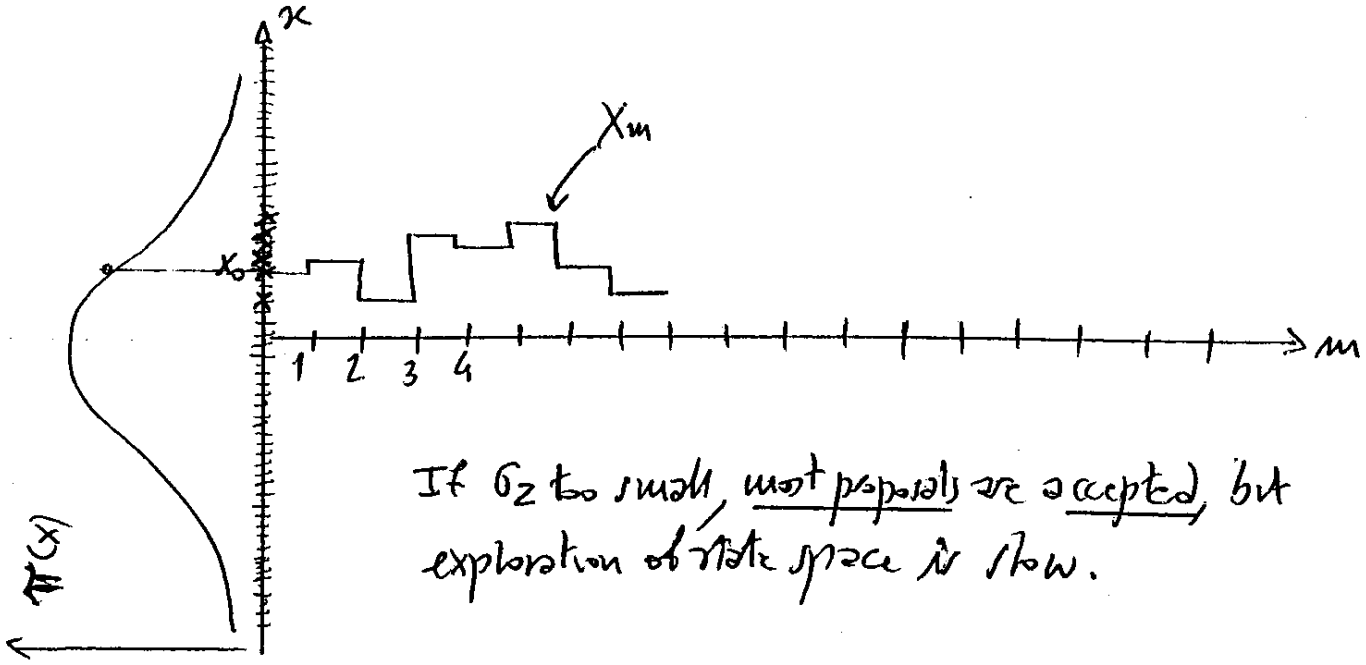
being Z a non negative discrete RV with integer values and PMF $\{P_Z(k), k=0,1,2\ldots\}$ [e.g. $Z \sim$ Poisson(λ)]



Hence from x_J, explorer proposes symmetric movements around x_J.
Let σ_Z^2 be the variance of Z.
Q: what is impact of σ_Z^2 on movement of X_m?



If σ_z too large, most proposals are rejected and the chain barely moves.

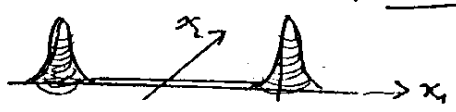


If σ_z too small, most proposals are accepted, but exploration of state space is slow.

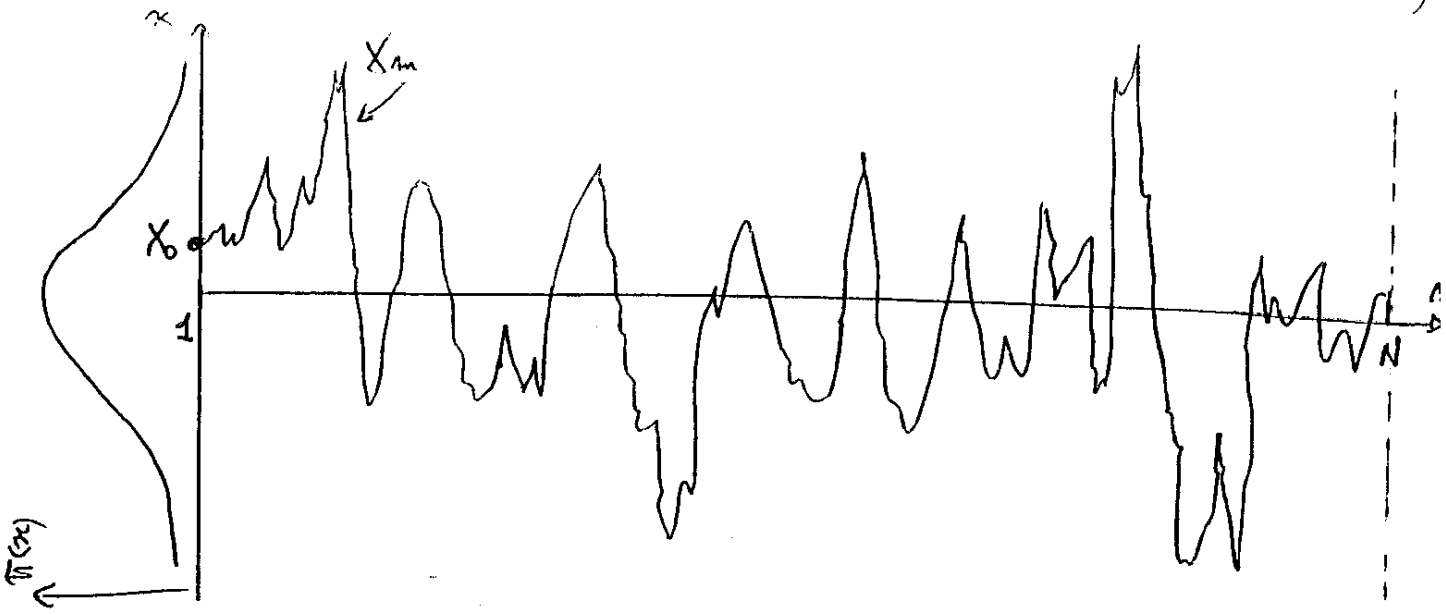
Q: what is a good σ_z? For uni-modal distribution, practice says σ_z should be such that acceptance rate $\simeq 25\%$ (in $\underline{x}$ multivariate) and $\simeq 50\%$ if x scalar.

But for multi-modal $\pi(x)$, σ_z must be increased to "jump" between modes, hence acceptance rate can be much lower

BEST



Q| When is a simulation long enough? (for sampling from $\pi(x)$)



A| When the time-series $\{X_m, m=1,..,N\}$ displays most features of the DTMC $\{X_m\}$, and from time averages one can reliably estimate statistical properties (ERGODICITY). Should "see" several cycles of $\{X_m\}$ up and down from modal area.

▷ When X_0 is up in the tails of $\pi(x)$, it takes longer to $\{X_m\}$ to come back to "modal stripe" ($X_m \in [\eta_X - 2\sigma_X, \eta_X + 2\sigma_X]$). We mathematically say that state vector $\underline{p}_n$ converges more slowly to $\underline{\pi}$. For a time-series such as that in the figure, we can be sure that $\underline{p}_N \cong \underline{\pi}$

▷ DTMC has built-in correlations. Samples are not independent. Hence estimators such as

$$\bar{Y}_m = \frac{1}{m}\sum_{i=1}^{m} Y_i = \frac{1}{m}\sum_{i=1}^{m} g(X_i)$$

have

$$Var[\bar{Y}_m] = \frac{1}{m}\left\{ Var[g(X)] + 2\sum_{k=1}^{m-1} \frac{m-k}{m} Cov[g(X_i)g(X_{i+k})] \right\}$$

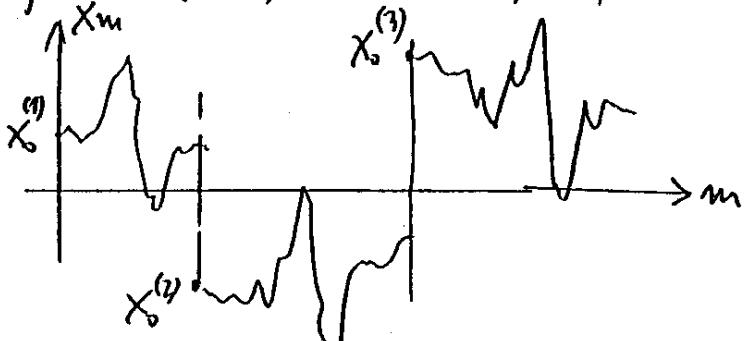
(FOR IID samples) (generally >0, variance increased!)

Metropolis-rejected Restarts [Geyer's book p.152]

With multi-modal distributions $\pi(x)$, the MCMC sampler can get stuck in one MODE (or "hump" of the PDF $\pi(x)$) and thus incorrectly sample from $\pi(x)$.

The idea then is to restart the MCMC after M iterations from a new starting point, in order to explore all the "modes" of $\pi(x)$

But if restarting points $\{x_0^{(k)}\}$ are not properly chosen, the resulting composite chain



does not have $\pi(x)$ as its limiting distribution.

To explain how to choose $\{x_0^{(1)}, x_0^{(2)}, \dots\}$, I must go back to Hastings MCMC.

▷ Independence chains

It is a particular Hastings MCMC, in which the explorer

$(J \to i)$ $\quad q_{iJ} = P_W(i)$ depends solely on proposed state i and not on starting state

where W is some discrete RV with PMF $P_W(\cdot)$ and variance σ_W^2

Hence, prop-sals $\{\hat{X}_m\}$ are independent (not $\{X_m\}$ because of rejection mechanism)

□

Hence implement the Metropolis-rejected restart as follows:

After running a block of M time steps with a low variance state-dependent explorer $\{q_{ij}\}$, and having reached state $X_m = x_j$, at time $m+1$ do the following:

▷ Select the next proposal $\overset{x_i^*}{i}$ using a state-independent explorer $\{\tilde{q}_{ij}\}$, with $\tilde{q}_{ij} = p_W(i)$, having large variance σ_W^2, and do Metropolis-Hastings rejection based on the odds ratio

$$\odot \qquad R = \frac{\pi_i \, \tilde{q}_{ji}}{\pi_j \, \tilde{q}_{ij}} = \frac{\pi_i \, p_W(j)}{\pi_j \, p_W(i)}$$

$\alpha_{ij} = \min(1, R)$ If accepted, use x_i as the new starting point: $X_{m+1} \equiv X_0^{(2)} = x_i$ and start a new block of M runs with low-variance explorer $\{q_{ij}\}$.

Else keep the old sample $X_{m+1} = x_j$, and continue sampling from large variance sampler $\{\tilde{q}_{ij}\}$ for some time (e.g. until you get a large jump accepted). Then start a new block of M runs of $\{q_{ij}\}$.

In all cases, by composing at subsequent times MCMC algorithms that all have the same $\underline{\pi}$, we are not altering the steady-state statistical properties of the sampled process, as we show next.

Composition and Mixing [Geyer's book, p. 49]

The rationale behind the Metropolis-rejected restarts is the following.

In basic MCMC, using the explorer $\{q_{ij}\}$ and the rejection rule based on the odds ratio $R = \frac{\pi_i q_{ji}}{\pi_j q_{ij}}$ we are implementing a reversible DTMC having transition matrix $\underline{\underline{P}}$ such that

$$\underline{\underline{P}}\,\underline{\pi} = \underline{\pi} \quad .$$

Now the composition at subsequent times of d different update mechanisms which individually would produce transition matrices $\underline{\underline{P}}_1, \ldots, \underline{\underline{P}}_d$, gives the compound transition matrix

$$\underline{\underline{P}} = \left(\underline{\underline{P}}_d \cdots \underline{\underline{P}}_2 \underline{\underline{P}}_1\right)$$

It is easy to see that if $\underline{\underline{P}}_i \underline{\pi} = \underline{\pi}$ for all i, then $\underline{\underline{P}}\,\underline{\pi} = \underline{\pi}$.

Composition produces a non-homogeneous DTMC that converges to the desired $\underline{\pi}$.

▷ Another update mechanism that preserves $\underline{\pi}$ is convex mixing, which corresponds to forming

$$\underline{\underline{P}} = \sum_{i=1}^{d} a_i \cdot \underline{\underline{P}}_i \qquad \text{with } \sum_{i=1}^{d} a_i = 1,\ a_i \geq 0$$

It corresponds to selecting the i-th update with probability a_i at each time step.

Small-world MCMC [Y. Guan et al., Stat. and Computing Vol 16, n. 2, June 06, p 193]

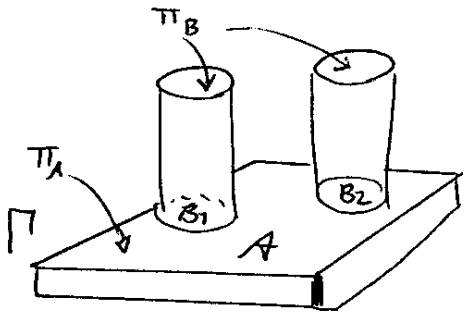
A simpler implementation that has the capability of exploring all modes is the following:

Suppose the "local" explorer is $\tilde{Q} = \{\tilde{q}_{ij}\}$ which gives proposals that slowly move around. Then as a "global" explorer use $Q = \{q_{ij}\}$ with

$$q_{ij} = (1-p)\tilde{q}_{ij} + \underbrace{\frac{1}{|\Gamma|}}_{\text{uniform distrib. over } \Gamma} \cdot p$$

where the parameter p is the probability that the explorer, at each time, chooses not from the local explorer, but from a uniform distribution (hence can make large jumps).

Empirical choice of p



Assume our distribution $\pi(x)$ on state space Γ has two important regions B_1, B_2 with equal level $\pi(x) = \pi_B \ \forall x \in B_1, x \in B_2$ and an unimportant region A, with $\pi(x) = \pi_A \lll \pi_B \ \forall x \in A$.

When in A, local explorer remains in A.
When in B, " " " " B.

Jumps $A \to B$, $B_1 \leftrightarrow B_2$ are achieved with uniform explorer.
Jumps $B \to A$ are "always" discarded, since $\pi_A/\pi_B \cong 0$ (Metropolis).

> Assume we know "size" of important regions $|B_1|, |B_2|$, $|B| \triangleq |B_1| + |B_2|$.

Then, when starting from A (initial state X_0 chosen uniformly) at each run the probability of jumping to B is:

$$P_{succ} = p \cdot \frac{|B|}{|\Gamma|}$$

Since trials are independent, expected number of trials to jump into B when starting in A is

$$E[T_{A\to B}] = \frac{1}{P_{succ}} = \frac{|\Gamma|}{p|B|}$$

Assume that, once in B, we run simulation for M time steps. The probability that explorer proposes a jump $B \to A$ (which is next rejected with probability $1 - \frac{\pi_A}{\pi_B} \cong 1$) is $p \cdot \frac{|A|}{|\Gamma|}$, hence average number of rejected $B \to A$ transitions is

$$E[T_{B\to A}] = M \cdot p \frac{|A|}{|\Gamma|}$$

If initial X_0 is chosen uniformly and $|B| \ll |A|$, then expected number of "useless" steps is

$$h(p) \triangleq E[T_{A\to B}] + E[T_{B\to A}] = \frac{|\Gamma|}{p|B|} + Mp\left(1 - \frac{|B|}{|\Gamma|}\right)$$

This is minimized by $\frac{\partial h}{\partial p} = 0 \Rightarrow$ Letting $r \triangleq |B|/|\Gamma|$, optimum p is:

$$\boxed{p = \frac{1}{\sqrt{Mr(1-r)}} \cong \frac{1}{\sqrt{Mr}}}$$

which is our rule for choosing p when we plan to run M times and guess that important regions have "area" r.

NOTE: SW MCMC is not convex mixing of update mechanisms since there is a single update rule based on the odds ratio that uses the mixed explorer $\{q_{ij}\}$.

One-Variable-at-a-Time Metropolis-Hastings [Geyer's book p. 78]

Consider a d-dimensional state space, so that $\underline{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} \in \mathbb{R}^d$.

Reversibility of the DTMC (56) is achieved by using $\forall$ pair of states $\underline{x}_i, \underline{x}_j$:

$$(R) \qquad \pi_i \cdot (q_{ji}\,\alpha_{ji}) = \pi_j (q_{ij}\,\alpha_{ij}) \quad , \quad \alpha_{mp} = \min\left[1, \frac{\pi_m q_{pm}}{\pi_p q_{mp}}\right]$$

i.e by proposing a new state $\underline{x}_m$ given the present state $\underline{x}_p$ with (explorer) probability q_{mp}, and accepting the proposal with probability α_{mp}.

However, the update $\underline{x}_p \to \underline{x}_m$ can be done, instead of in a single step, by breaking the transition component-wise:

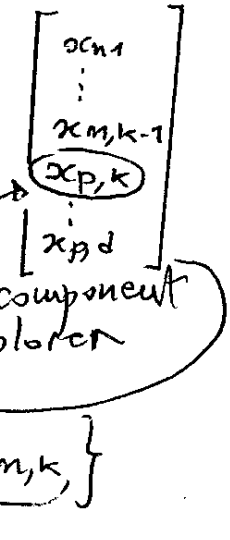
$$\underline{x}^{(0)} \equiv \underline{x}_p = \begin{bmatrix} x_{p,1} \\ x_{p,2} \\ \vdots \\ x_{p,d} \end{bmatrix} \to \underline{x}^{(1)} = \begin{bmatrix} \textcircled{x_{m,1}} \\ x_{p,2} \\ \vdots \\ x_{p,d} \end{bmatrix} \to \underline{x}^{(2)} = \begin{bmatrix} \textcircled{x_{m,1}} \\ \textcircled{x_{m,2}} \\ x_{p,3} \\ \vdots \\ x_{p,d} \end{bmatrix} \to \cdots \underline{x}^{(n)} \equiv \underline{x}_m = \begin{bmatrix} x_{m,1} \\ \vdots \\ x_{m,d} \end{bmatrix}$$

i.e. by using

$$q_{mp}\,\alpha_{mp} = \left[q_{mp}^{(1)}\alpha_{mp}^{(1)}\right] \cdot \left[q_{mp}^{(2)}\alpha_{mp}^{(2)}\right] \cdot \ldots \cdot \left[q_{mp}^{(d)}\alpha_{mp}^{(d)}\right]$$

where $$\alpha_{mp}^{(k)} = \min\left[1, \frac{\pi(\underline{x}^{(k)})}{\pi(\underline{x}^{(k-1)})} \frac{q_{mp}^{(k)}}{q_{pm}^{(k)}}\right]$$

and
$$\begin{cases} q_{mp}^{(k)} = P\left\{ x_k^{(k)} = x_{m,k} \mid \underline{x}^{(k-1)} \right\} \\ q_{pm}^{(k)} = P\left\{ x_k^{(k)} = x_{p,k} \mid \underbrace{\underline{x}^{(k-1)} \text{ with } k\text{-th component } x_{m,k}}_{\underline{x}^{(k)}} \right\} \end{cases}$$



In words:

for all components $k = 1, \ldots, d$ propose the move $\underline{x}^{(k-1)} \to \underline{x}^{(k)}$ with probability

$$q_{mp}^{(k)} = P\left\{ \underline{x}^{(k)} \,\middle|\, \underline{x}^{(k-1)} \right\}$$

(which is the probability of proposing the move $x_{p,k} \to x_{m,k}$ in the k-th component, given the remaining components of $\underline{x}^{(k-1)}$ and accept the move with probability $\alpha_{mp}^{(k)}$.

By construction then this component-wise procedure implements (R), i.e., a reversible DTMC having $\Pi(\underline{x})$ as its steady-state distribution (PMF).

MMC with MCMC

$$f_X^{(n)}(x) = \frac{f_x(x)}{c_n \hat{P}_Y^{(n-1)}(g(x))}$$

$x^* \rightarrow [\,g(\cdot)\,] \rightarrow y^*$

Draw N samples

$$\hat{H}_Y^{(n)}(y_i) = \frac{N_i^{(n)}}{N}$$

$$\hat{P}_Y^{(n)}(y_i) = \hat{H}_Y^{(n)}(y_i) \cdot c_n \hat{P}_Y^{(n-1)}(y_i)$$

update n

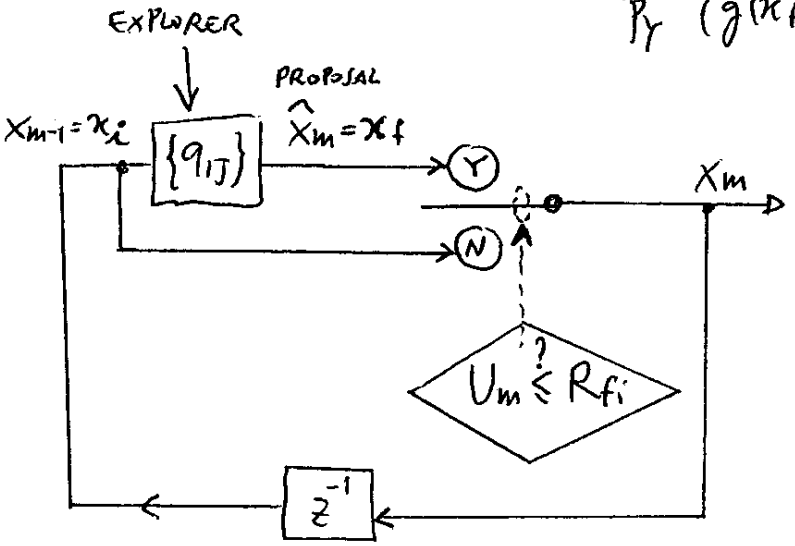
need an MCMC machine to implement this, with

$\pi(x) = f_x^{(n)}(x)\,\Delta x$. Recalling odds ratio for move $x_i \to x_f$:

(Cfr p. 63)

$$R_{fi} = \frac{\pi(x_f)\, q_{if}}{\pi(x_i)\, q_{fi}} = \frac{\dfrac{f_x(x_f)\,\Delta x}{c_n \hat{P}_Y^{(n-1)}(g(x_f))}}{\dfrac{f_x(x_i)\,\Delta x}{c_n \hat{P}_Y^{(n-1)}(g(x_i))}} \cdot \frac{q_{if}}{q_{fi}}$$

$$R_{fi} = \frac{\hat{P}_Y^{(n-1)}(g(x_i))}{\hat{P}_Y^{(n-1)}(g(x_f))} \cdot \frac{f_x(x_f)}{f_x(x_i)} \cdot \frac{q_{if}}{q_{fi}}$$



If we choose explorer as

$$q_{fi} = f_x(x_f)\cdot \Delta x$$

(independence chain cfr p. 67)

then

$$R_{fi} = \frac{\hat{P}_Y^{(n-1)}(y_i)}{\hat{P}_Y^{(n-1)}(y_f)} \quad \ldots$$

... depends solely on "control variable" $y = g(x)$

MMC using MCMC

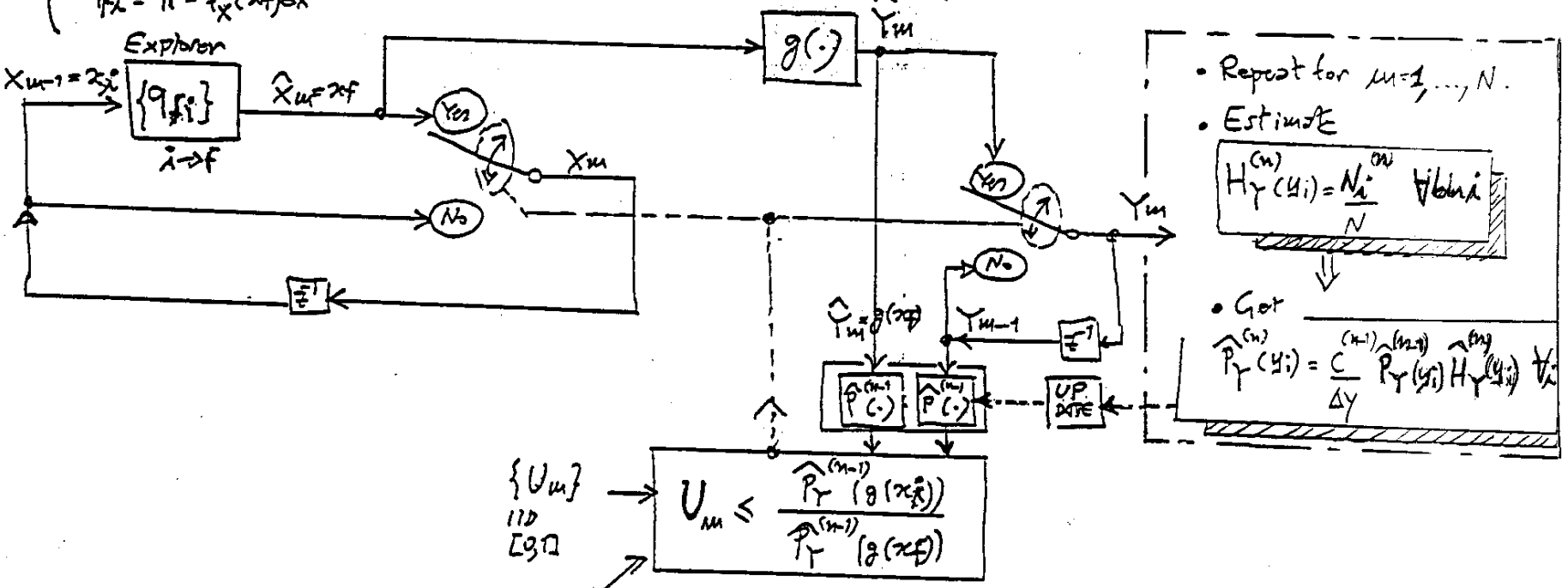
[Holzlohner et al, Opt. Lett. Oct. 2005]

We now show 1 complete step of Multicanonical:

This is an INDEPENDENCE CHAIN

Actual distribution of X, which we know how to generate

$q_{fi} \equiv \tilde{q}_f = f_x(x_f)dx$



$\{U_m\}$ IID $[0,1]$

$$U_m \le \frac{\hat{P}_Y^{(n-1)}(g(x_i^*))}{\hat{P}_Y^{(n-1)}(g(x_f))}$$

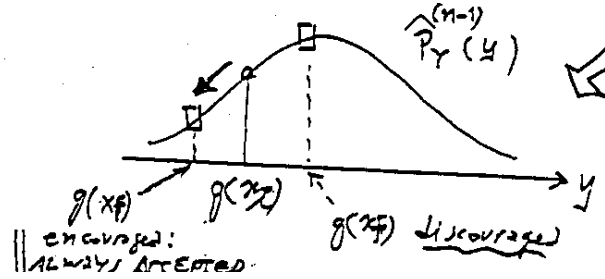
Explanation:

From Hastings

(56)

$$U_m \le \frac{\pi_f q_{if}}{\pi_i q_{fi}} = \frac{\pi_f \tilde{q}_i}{\pi_i \tilde{q}_f} = \frac{\frac{C}{\Delta y}^{(n-1)} \hat{P}_Y^{(n-1)}(g(x_f))}{\frac{C}{\Delta y}^{(n-1)} \hat{P}_Y^{(n-1)}(g(x_i))} \cdot \frac{f_x(x_i)dx}{f_x(x_f)dx}$$

desired PMF of X $(= f_x^{(n)}(x)\,dx)$



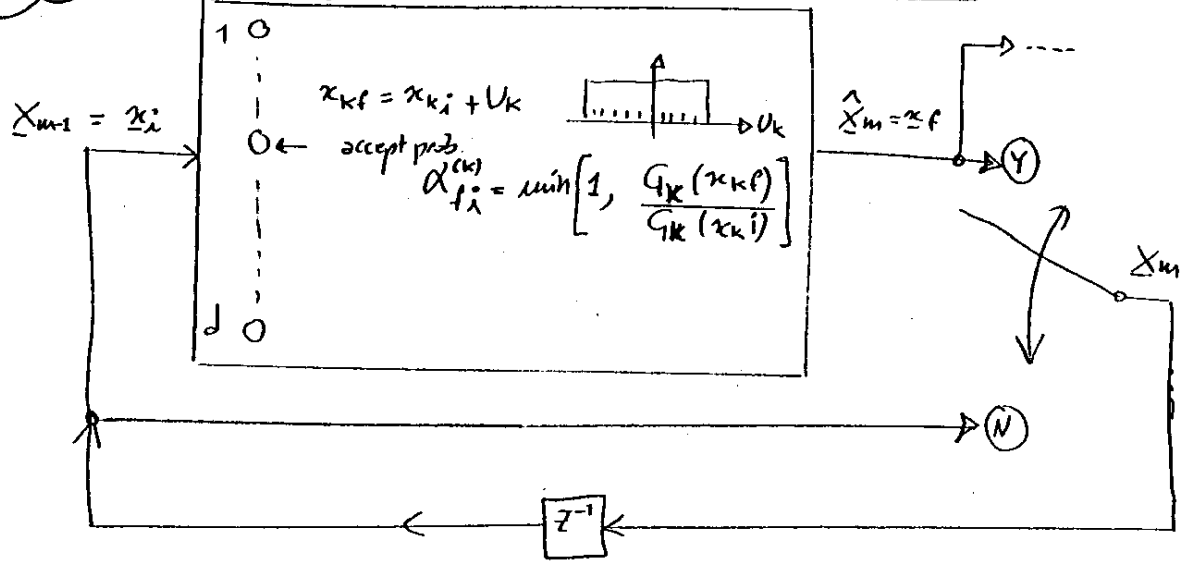
PUSHES TOWARDS TAILS OF $\hat{P}_Y(y)$!

Trouble: Algo pushes towards tails of $\hat{P}_Y^{(n-1)}(y)$. Once we are in tail, since next proposals will be in mode of $\hat{P}_Y^{(n-1)}(y)$, they will almost all be rejected ⇒ chain will hardly move!

We next describe a method that works for $\underline{X}$ with indep. components, i.e when $f_{\underline{X}}(\underline{x}) = \prod_{k=1}^{\bar{n}} G_k(x_k)$.

Menyuk's Trick [Opt. Lett. Oct. '03]

The IDEA ⇒ Explorer $\{q_{fi}\}$ implements $f_X(x)\Delta x$ with an MC MC machine!!

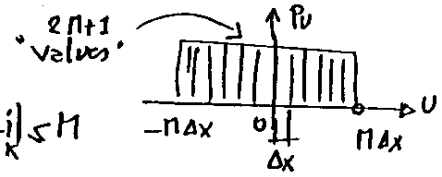


MC MC machine with Explorer implemented with sequential "scan" component by component (cfr page 72) with independent reject/accept on each component, with accept probability on k-th component:

$$\alpha^{(k)}_{fi} = \min\left[1, \frac{G_k(x_{kf})}{G_k(x_{ki})}\right],$$

k-th component explorer

Prop-sal on k-th component is $x_{kf} = x_{ki} + U_k$, where $U_k \sim \text{Uniform}[-\Delta x_M, \Delta x_M]$ with $\Delta x_M \triangleq M \cdot \Delta x$, hence



$$P\{U_k = x_{kf} - x_{ki} = (f_k - i_k)\Delta x\} = \begin{cases} \frac{1}{2M+1} & \text{if } |f_k - i_k| \le M \\ 0 & \text{else} \end{cases} \quad (56c)$$

which is symmetric in $(f_k - i_k)$, as in Metropolis.

Now Q: What is $\{q_{fi}\}$ of such a scheme?

I will show that

$$\frac{q_{if}}{q_{fi}} = \frac{f_x(x_i)}{f_x(x_f)} = \prod_{k=1}^{d} \frac{g_k(x_{ki})}{g_k(x_{kf})}$$

exactly as if explorer was an independence chain.

Proof

$$q_{fi} \triangleq P\{\hat{X}_m = x_f \mid X_{m-1} = x_i\} = \quad \text{by independence of componentwise rejections}$$

$$= \prod_{k=1}^{d} P\{\hat{X}_{k,m} = x_{kf} \mid X_{k,m-1} = x_{ki}\}$$

$$= \prod_{k=1}^{d} P\{U_k = \underbrace{x_{kf} - x_{ki}}_{\triangleq (f_k - i_k)\Delta x}\} \cdot \alpha_{fi}^{(k)} \qquad \textcircled{o}$$

Thus form ratio:

$$\frac{q_{if}}{q_{fi}} = \frac{\prod_{k=1}^{d} P\{U_k = (f_k - i_k)\Delta x\} \min\left[1, \frac{g_k(x_{ki})}{g_k(x_{kf})}\right]}{\prod_{k=1}^{d} P\{U_k = (f_k - i_k)\Delta x\} \min\left[1, \frac{g_k(x_{kf})}{g_k(x_{ki})}\right]}$$

These cancel by Metropolis symmetry

$$= \begin{cases} \dfrac{1}{\frac{g_k(x_{kf})}{g_k(x_{ki})}} & \text{if } g_k(x_{ki}) > g_k(x_{kf}) \\ \dfrac{g_k(x_{ki})}{g_k(x_{kf})} & \text{else!} \end{cases}$$

hence

$$= \frac{g_k(x_{ki})}{g_k(x_{kf})} \quad \text{always!}$$

Hence

$$\frac{q_{if}}{q_{fi}} = \prod_{k=1}^{d} \frac{g_k(x_{ki})}{g_k(x_{kf})} \qquad \square \text{ QED}$$

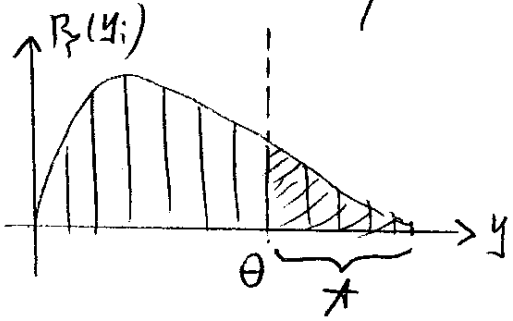
Advantage of this scheme is that rejects are less likely, since chain X_m is corrected, and "slowly" moves around if proposal Δx_m is not "too" large!

$\textcircled{o}$ More correctly

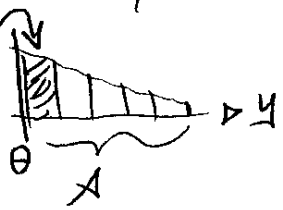
$$\begin{cases} P\{U_k = (f_k - i_k)\Delta x\} \cdot \alpha_{fi}^{(k)} & \text{if } f_k \neq i_k \\ P\{U_k = 0\} + \sum_{f_k \neq i_k} P\{U_k = (f_k - i_k)\Delta x\} \cdot (1 - \alpha_{fi}^{(k)}) & \text{else} \end{cases}$$

Output warping IS (OWIS)

▷ The previous paragraph has concluded our technical explanation of MMC. We now understand that MMC is the best estimator (in the sense of properties at p. (25) and result of exercise 3 at p. 286) of the PMF $P_Y(y_i)$, $\forall i = 1, \dots, M$: Given N "runs", it allocates the N samples in the "best way", so as to give same accuracy for estimation of every bin.



If, however, we just need the probability P_A of a rare event $\{Y \in A\}$ (say $A = \{Y > \text{threshold } \theta\}$) it is plausible that we do not need estimating $P_Y(y_i)$ for all bins $i \in A$ with the same accuracy: we can allocate our N samples in a more efficient way. (e.g. put more samples in the dominant bin



)

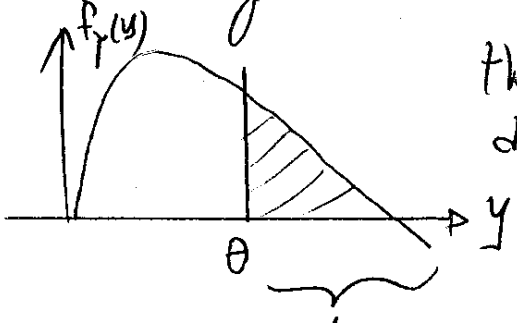
That's why I would like to go back to the interesting theorem on output warping proved at page 26. We now recall that result:

Given any strictly positive $d(y) > 0 \quad \forall y \in R_Y$,
we have that

$$f_X^*(x) = \frac{f_X(x)}{c\,d(g(x))} \xrightarrow{X} \boxed{g(\cdot)} \xrightarrow{Y} f_Y^X(y) = \frac{f_Y(y)}{C \cdot d(y)} \triangleq e(y) \cdot f_Y(y)$$

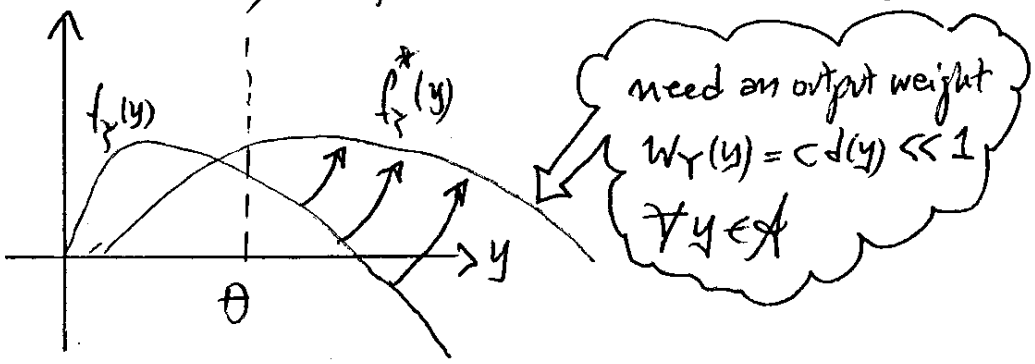
where we may interpret $e(y) \triangleq \frac{1}{C\,d(y)}$ as an "emphasis function" of $f_Y(y)$.

The idea now is the following:

If we know that 

the true $f_Y(y)$ has a decaying tail over A...

... then we may "emphasize" with a suitable $f_Y^*(y) = f_Y(y) \cdot e(y)$ so as to increase the probability in the tail, and thus obtain more (warped) samples in A, as per Importance Sampling:



▷ The magic is that now we are imposing the right warping ON THE OUTPUT PDF, instead of classical IS which (blindly) imposes some fancy "given" $f_X^*(x)$ obtained by some physical reasoning on the specific problem at hand!

So, we now have:

$$P_A = \int_A f_Y(y)\,dy = \int_{-\infty}^{\infty} I_A(y) f_Y(y)\,dy$$

$$= \int_{-\infty}^{\infty} I_A(y) \underbrace{\left[\frac{f_Y(y)}{f_Y^*(y)}\right]}_{w(y) = c\,d(y) = \frac{1}{e(y)} > 0} f_Y^*(y)\,dy = E^*\left[I_A(Y) w(Y)\right] \quad (57)$$

and thus we use the estimate of the expectation, i.e., the output-warped importance sampling (OWIS) estimator:

$$\hat{P}_A = \frac{1}{N} \sum_{j=1}^{N} I_A(Y_j) w(Y_j) = \underbrace{\left[\frac{1}{N} \sum_{j=1}^{N} I_A(Y_j)\, d(Y_j)\right]}_{\triangleq \hat{\theta}} \cdot C \quad (58)$$

Trouble is that C is unknown, since $f_Y(y)$ is unknown. However,

$$1 = \int_{-\infty}^{\infty} f_Y(y)\,dy = c \int_{-\infty}^{\infty} f_Y^*(y)\, d(y)\,dy \overset{LLN}{=} c \cdot E^*\left[d(Y)\right] \quad (59)$$

so that we can estimate C from the same N samples (warped) as: $\hat{C} = \dfrac{1}{\frac{1}{N}\sum_{j=1}^{N} d(Y_j)}$, and finally the OWIS estimator of event A is

$$\boxed{\hat{P}_A = \frac{\sum_{j=1}^{N} I_A(Y_j) d(Y_j)}{\sum_{j=1}^{N} d(Y_j)}} \quad (60)$$

when drawing $\{Y_j\}$ samples using $f_Y^*(y)$. How can we do that?

Of course with an MCMC machine!
one that has steady state PMF on states $\pi(x)$:

$$\pi(x) = f_x^{\star}(x)\Delta x = \frac{f_x(x)\Delta x}{c\, d(g(x))}$$

and with a Hastings odds ratio

$$R_{fi} = \frac{\pi(x_f) q_{if}}{\pi(x_i) q_{fi}} = \frac{f_x(x_f)}{f_x(x_i)} \frac{d(g(x_i))}{d(g(x_f))} \frac{q_{if}}{q_{fi}}$$

For the explorer as at p. 74, we may use an independence chain: $q_{fi} = f_x(x_f)\Delta x$

so that odds ratio simplifies to

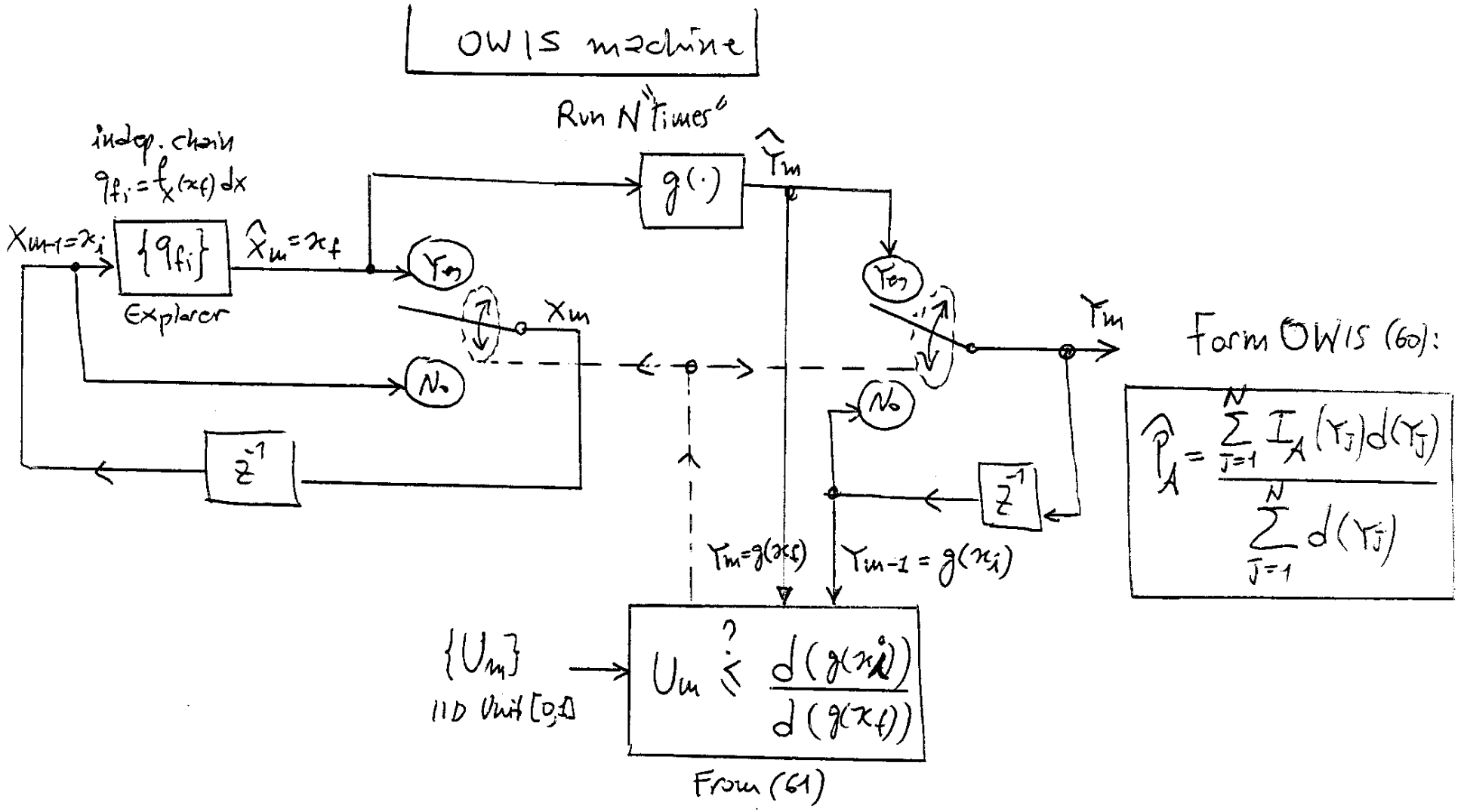
$$R_{fi} = \frac{d(g(x_i))}{d(g(x_f))} \qquad (61)$$

and explorer can be implemented with the one-variable-at-a time trick explained at pp. 75–76 in order to lower the rejection ratio.

I learned these ideas (in disguised form!!) from a genial paper by Marco Fecondini and co-workers at S. Anna School in Pisa [M. Fecondini et al., in Proc. TIWDC '04], where

$$d(y) = e^{-\beta y} \qquad (62)$$

was used, and β had to be optimized (i.e. the usual "tuning" of IS)


OWIS machine
Run N "Times"
indep. chain
$q_{fi} = f_X(x_f)\,dx$
$X_{m-1} = x_i$
$\{q_{fi}\}$
$\hat{X}_m = x_f$
Explorer
$g(\cdot)$
$\hat{Y}_m$
Y_m
X_m
N_0
z^{-1}
$Y_m = g(x_f)$
$Y_{m-1} = g(x_i)$
$\{U_m\}$
IID Unif [0,1]
$U_m \overset{?}{<} \frac{d(g(x_i))}{d(g(x_f))}$
From (61)
From OWIS (60):
$\hat{P}_A = \frac{\sum_{J=1}^{N} I_A(Y_J)\,d(Y_J)}{\sum_{J=1}^{N} d(Y_J)}$

▷ Why is this amazingly new?

In all previous telecom papers on IS, authors were trying "creative" input warping (variance-increase of f_x, translation method, etc, see References).

But ignorance in our community of the MCMC machine has blocked the advancement of research in this area!

Now new frontiers are open.

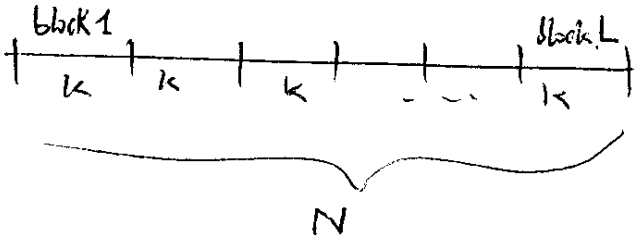
▷ Estimation of error in OW IS from samples [Hastings 1970, Secondini, 2004]

$$\hat{P}_A = \underbrace{\frac{1}{N}\sum_{J=1}^{N} I_A(Y_J)\, d(Y_J)}_{\triangleq \hat{\theta}} \Big/ \underbrace{\frac{1}{N}\sum_{J=1}^{N} d(Y_J)}_{\triangleq \hat{C}} \qquad (63)$$

the estimated variance of estimator is computed from the Delta method as

$$\hat{\mathrm{Var}}(\hat{P}_A) = \frac{\hat{\mathrm{Var}}(\hat{\theta}) - 2\hat{P}_A \hat{\mathrm{Cov}}[\hat{\theta},\hat{C}] + \hat{P}_A^2 \hat{\mathrm{Var}}(\hat{C})}{\hat{C}^2} \qquad (64)$$

where to remove correlations in MCMC we divide sample size N into L blocks of K runs each



so that $K >$ correlation time of MCMC

and evaluate

$$\hat{\theta} = \frac{1}{L}\sum_{i=1}^{L}\bar{\theta}_i \quad ; \quad \hat{c} = \frac{1}{L}\sum_{i=1}^{L}\bar{c}_i \qquad (65)(a)$$

($\bar{\theta}_i$, $\bar{c}_i$ on block i^o)

Then

$$(65b)\begin{cases} \hat{Var}[\bar{\theta}_i] = \frac{1}{L-1}\sum_{i=1}^{L}(\bar{\theta}_i - \hat{\theta})^2 \\ \hat{Var}[\bar{c}_i] = \frac{1}{L-1}\sum_{i=1}^{L}(\bar{\theta}_i - \hat{\theta})^2 \\ \hat{Cov}[\bar{\theta}_i, \bar{c}_i] = \frac{1}{L-1}\sum_{i=1}^{L}(\bar{c}_i - \hat{c})\cdot(\bar{\theta}_i - \hat{\theta}) \end{cases}$$

and finally, because of incorrelation among blocks:

$$(65c)\begin{cases} \hat{Var}[\hat{\theta}] = \frac{1}{L}\hat{Var}[\bar{\theta}_i] \\ \hat{Var}[\hat{c}] = \frac{1}{L}\hat{Var}[\bar{c}_i] \\ \hat{Cov}[\hat{\theta}, \hat{c}] = \frac{1}{L}\hat{Cov}[\bar{\theta}_i, \bar{c}_i] \end{cases}$$

<u>Note</u> Delta method states that given $Z = g(X,Y) = \frac{Y}{X}$, then

$$Z \overset{\text{Taylor}}{\simeq} g(\mu_X,\mu_Y) + \frac{\partial g}{\partial y}\Big|_{(\mu_X,\mu_Y)}\cdot(Y-\mu_Y) + \frac{\partial g}{\partial x}\Big|_{(\mu_X,\mu_Y)}(X-\mu_X)$$

$$= \frac{\mu_Y}{\mu_X} + \frac{1}{\mu_X}(Y-\mu_Y) - \frac{\mu_Y}{(\mu_X)^2}\cdot(X-\mu_X) = \frac{Y}{\mu_X} - \frac{\mu_Y}{\mu_X^2}(X-\mu_X)$$

hence

$$Var[Z] \simeq \frac{1}{\mu_X^2}\cdot Var\left[Y - \frac{\mu_Y}{\mu_X}X + \frac{\mu_Y}{\mu_X}\right]$$

$$= \frac{1}{\mu_X^2}\left\{Var[Y] + \left(\frac{\mu_Y}{\mu_X}\right)^2 Var[X] - 2\left(\frac{\mu_Y}{\mu_X}\right)Cov[Y,X]\right\} \qquad (66)$$

Exercise

i) The bias of the OWIS estimator is

$$b \triangleq E^*[\hat{P}_A] - P_A . \qquad (67)$$

Can you prove, using Jensen's inequality, that $b \geq 0$?

ii) Prove that an approximation to the relative bias error is

$$\left(\frac{b}{P_A}\right) \cong \frac{Var^*[\hat{C}] - \frac{1}{P_A} Cov^*[\hat{C},\hat{\theta}]}{E^*[\hat{C}]^2} \qquad (68)$$

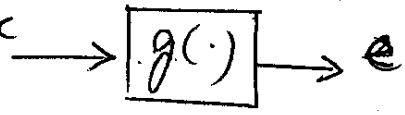
Berg's Problem

It is instructive now to take a look at the physical problem that Berg was facing when he invented the MMC.

Consider a system of d particles, each of which can be in ℓ different "spins". A string $[l_1, l_2, \ldots, l_d]$ (where l_i is the "spin" of the i-th particle) is a state x of our system (a "phase" in Berg's jargon). The state space Γ (called phase space) is thus discrete, with a number of states equal to $|\Gamma| = \ell^d$, extremely large.

The energy corresponding to a state is a given function of the state:

$$e = g(x)$$



where $g(\cdot)$ depends on the specific physical problem and is known or can be computed.

We know that, at equilibrium, the PMF of the R.V. X is given by the CANONICAL (Boltzmann) input PMF

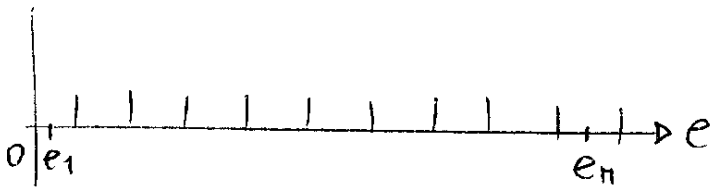
$$P_X(x) = \frac{e^{-\beta \cdot g(x)}}{Z(\beta)} \qquad (70)$$

where $\beta \equiv \frac{1}{T}$ is the inverse of the temperature, and

$$Z(\beta) = \sum_{x \in \Gamma} e^{-\beta g(x)} \qquad (71)$$

is the normalization constant (called "partition function") (which is extremely difficult to compute, given the enormous number of states $|\Gamma|$ in typical problems)

Now "slice" the energy axis into bins of equal width Δe centred at values $\{e_i\}$, $i = 1, \ldots, M$ ("small" Δe)



called CANONICAL ENERGY DISTRIBUTION

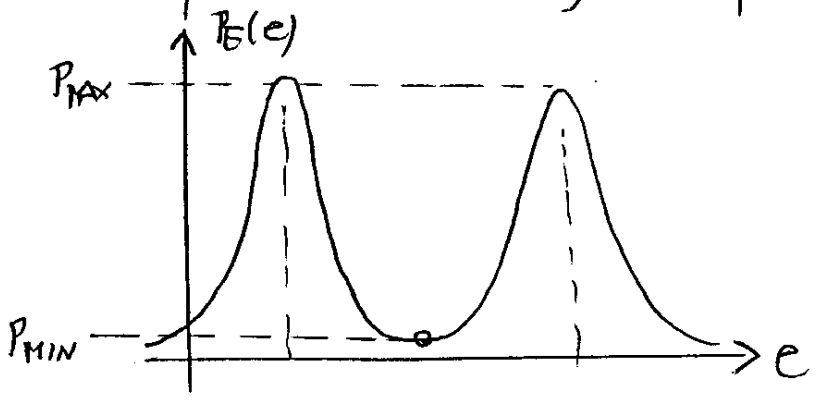
Let $\underline{P}_E = \{P_E(e_i)\}_{i=1}^M$ be the output PMF that we wish to estimate by simulation. Let $D_i = \{x \in \Gamma : g(x) \approx e_i\}$ be the domain in input space that maps thru $g(\cdot)$ into bin i, and let $|D_i| = \{\#$ of states x in $D_i\}$. Clearly

$$(72) \quad \boxed{P_E(e_i) = P\{E \approx e_i\} = \sum_{\{x \in D_i\}} P_X(x) = |D_i| \cdot \frac{e^{-\beta e_i}}{Z(\beta)}}$$

UW by physics $\longrightarrow \frac{e^{\beta e_i}}{Z(\beta)} \; \forall x \in D_i$

$\frac{|D_i|}{|\Gamma|} \triangleq n(e_i)$ is called the "density of states" (seen as a function of e) and depends on "temperature" β.

When the "lattice size" l is large enough, the canonical distribution $P_E(e)$ at a "critical temperature" β_c has two peaks (modes) of equal height

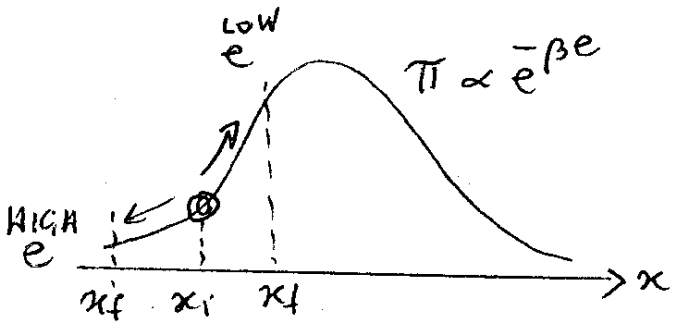


this situation is called a "first-order CRITICAL phase transition"

with Traditional MC

One can use an MCMC machine to sample from the input Boltzmann distribution $\pi(x) = p_X(x)$ using symmetric proposals ($q_{ij} = q_{ji}$) and odds ratio in move $x_i \to x_f$:

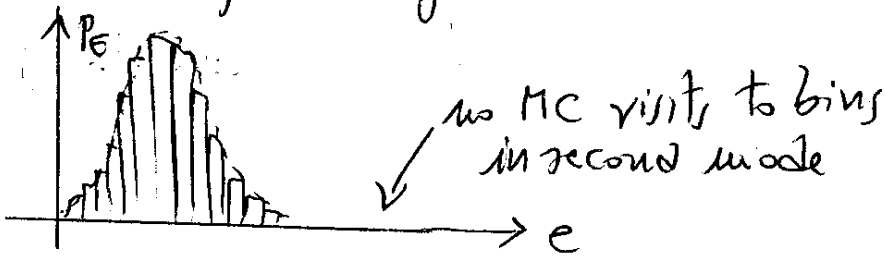
$$R_{fi} = \frac{\pi_f}{\pi_i} = \frac{e^{-\beta g(x_f)}}{Z(\beta)} \cdot \frac{Z(\beta)}{e^{-\beta g(x_i)}} = e^{-\beta(\underbrace{e_f}_{g(x_f)} - \underbrace{e_i}_{g(x_i)})}$$



as we know, in MCMC proposals x_f that increase π (lower energy e) are always accepted, those that decrease π are accepted with probability

$$\alpha_{fi} = R_{fi}$$

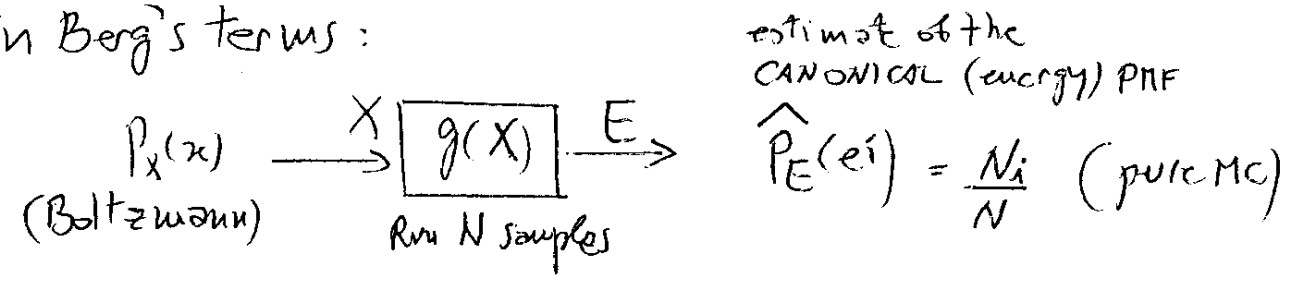
▷ Trouble is that MCMC sampler can get stuck in one (output) mode only

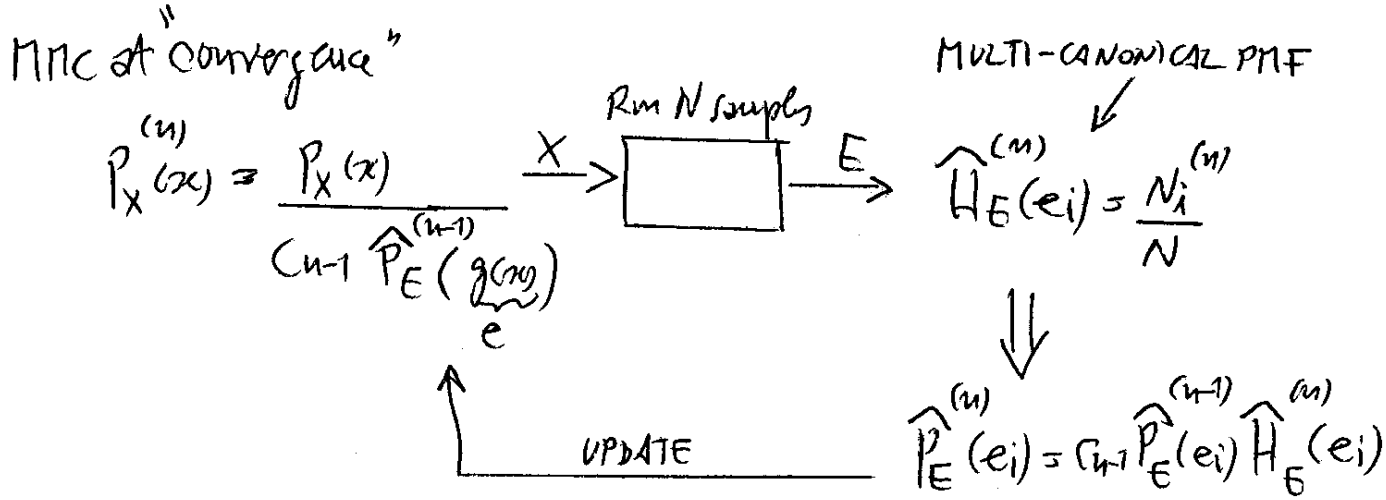


That's why Berg invented his Flat histogram MMC algorithm.

He also needed to estimate p_{MIN} correctly to derive some physical quantity of his interest.

In Berg's terms:





origin of name due to the fact that input warped distribution was thought to be approximated by

$$P_X^{(n)}(x) \cong C_M \cdot e^{-b_i^{(n)} e_i + a_i^{(n)}} \qquad \forall x \in D_i$$

$\triangleq$ micro-canonical (inverse) temperature

For a (very sloppy!) talk about MMC, read

J. Gubernatis, N. Hatano, Comput. Sim. Apr. 2000.

References

[IS] **Importance Sampling: references cited in the course**

[1] E. Veach, "Robust Monte Carlo methods for light transport simulation," Ph.D. thesis, Stanford University, 1997 (available at www.graphics.stanford.edu/papers/veach_thesis/thesis-bw.pdf. Material from Chapter 2 used for the course introduction. Chapter 9 is important for the invention of an advanced IS technique known as Multiple Importance sampling.)

[2] M. Jeruchim, "Techniques for estimating the bit error rate in the simulation of digital communication systems," J. Sel. Areas Commun., vol. SAC-2, pp. 153-170, Jan. 1994. (Cfr. course notes, pp. 9-12).

[3] K. S. Shanmugam, P. Balaban, "A modified Monte-Carlo simulation technique for the evaluation of error rate in digital communication systems," Trans. Commun, vol. COM-28, pp. 1916-1924, Nov. 1980. (Cfr. course notes, pp. 13-16).

[4] M. Secondini, E. Forestieri, and G. Prati, "Markov chain monte carlo technique for outage probability evaluation in PMD-compensated systems," in Proc. Tyrrhenian Int. Workshop Digital Communications (TIWDC'04), Pisa, Italy, 2004. Published in *Optical Communication theory and techniques,* E. Forestieri Ed., Springer 2005, pp. 121-128. (Cfr course notes, pp. 77-83).

[IS] **Importance Sampling: suggested further reading**

[5] P. M. Hahn, M. C. Jeruchim, "Developments in the theory and application of importance sampling," Trans. Commun., vol. COM-35, pp. 706-714, July 1987. (Useful discussion on the effect of dimesionality d of state space).

[6] G. Orsak, and B. Aazhang, "On the theory of importance sampling applied to the analysis of detection systems," Trans. Commun., vol. 37, pp. 332-339, Apr. 1989. (Elegant treatment of "Translation method" IS. Here one can find formulations on the average weight similar to course notes equations (19b), (27)).

[7] G. Biondini, W. L. Kath, and C. Menyuk, "Importance sampling for polarization mode dispersion: techniques and applications," J. Lightwave Technol., vol. 22, pp. 1201-1215, Apr. 2004. (Explains and uses the multiple IS technique. I found here the recursions on the evaluation of the sample mean and variance reported at page 8 of the course. Similar recursions can be found in standard textbooks on statistics, see e.g. S. Ross: *Introduction to probability and statistics for engineers and scientists*, 2nd Ed., Academic Press, 2000, Chapter 2, problems section.)

[8] A. Owen, T. Zhou, "Safe and effective importance sampling," J. Am. Stat. Assoc. vol. 95, no. 449, pp. 135-143, Mar. 2000. (Interesting examples on critical problems in the application of IS. Cites the multiple IS of Veach).

[9] W. Sandmann, "On optimal importance sampling for discrete time markov chains," in Proc. 2nd Int. Conf. on Quantitative Evaluation of Systems (QEST'05) (IEEE Computer Society Press. A powerful view of the zero-variance IS estimator for DTMC).

[10] J. S. Sadowsky, J. A. Bucklew, "On large deviations theory and asymptotically efficient monte carlo estimation," Trans. Inform. Theory, vol. 36, pp. 579-588, May 1990. (Rigorous treatment of efficiency in IS).

[11] J.-C. Chen, D. Lu, and J. S. Sadowsky, "On importance sampling in digital communications-Part I: fundamentals," J. Sel. Areas Commun., vol. 11, pp. 289-299, Apr. 1993.

[12] P. J. Smith, M. Shafi, and H. Gao, "Quick simulation: a review of importance sampling techniques in communications systems," J. Sel. areas Commun., vol. 15, pp. 597-613, May 1997. (Good review of literature, lots of references).

[13] P. Heidelberger, "Fast simulation of rare events in queueing and reliability models," Trans. Modeling and Computer Simulations, vol. 5, pp. 43-85, Jan. 1995.

[14] J. S. Stadler, and S. Roy, "Adaptive importance sampling," J. Sel. areas Commun., vol. 11, pp. 309-316, Apr. 1993. (one of first papers in communications that proposes an adaptive warping distribution whose functional form is built from estimated data, and not with an MCMC machine).

[FH] **Flat Histogram Algorithms: references cited in the course**

[15] F. Liang, "A theory on flat histogram monte carlo algorithms," J. Stat. Physics, vol. 122, no. 3, pp. 511-529, Feb. 2006. (Cfr course notes, pp. 28-29).

[16] Y. F. Atchade, J. S. Liu, "The Wang-Landau algorithm for MC computation in general state spaces", Technical report, University of Ottawa, 2004 (available at www.mathstat.uottawa.ca/~yatch436/gwl.pdf. Cfr course notes, p. 4)

[17] B. A. Berg, "Introduction to multicanonical monte carlo simulations," Fields Instr. Commun., vol. 26, no. 1, pp. 1-24, 2000. (also available at arXiv:cond-mat/9909236v1, 15 Sep. 1999. Cfr course notes, pp. 40-46).

[18] N. Rossi, "Algoritmi multicanonici innovativi e loro applicazione alla caratterizzazione statistica della dispersione modale di polarizzazione," Laurea Thesis in Telecommunications Engineering, Universita' di Parma, Feb. 2006. (Cfr. course notes, p. 41).

[19] D. Yevick, "The accuracy of multicanonical system models," Photon. Technol. Lett., vol. 15, pp. 224-226, Feb. 2003. (Cfr. course notes, p. 41).

[20] R. Holzlohner, and C. R. Menyuk, "Use of multicanonical monte carlo simulations to obtain accurate bit error rates in optical communications systems," Opt. Lett., vol. 28, no. 20, pp. 1894-1896, Oct. 2003. (Cfr. course notes, pp. 74-76).

[21] J. Gubernatis, and N. Hatano, "The multicanonical monte carlo method," Computer Simulations, vol. 2, no. 2, pp. 95-102, Mar./Apr. 2000. (Cfr. course notes, p. 88. Although this paper gives a very qualitative and often imprecise description of MMC, here I got my first idea on how IS was related to MMC. From there to get to your class notes, long hours of self-study followed.)

[FH] **Flat Histogram Algorithms: suggested further reading**

[22] A. O. Lima, I. T. Lima, and C. R. Menyuk, "Error estimation in multicanonical monte carlo simulations with applications to polarization mode dispersion emulators," J. Lightw. Technol., vol. 23, no. 11, Nov. 2005. (Contains methods to estimate error variance from correlated MMC samples)

[23] B. A. Berg, and T. Neuhaus, "Multicanonical ensemble: a new approach to simulate first-order phase transitions," Phys. Rev. Lett., vol. 68, no. 1, pp. 9-12, Jan. 1992. (trying to read it is a must. Good luck!)

[24] T. Lu, and D. Yevick, "Efficient multicanonical algorithms," Photon. Technol. Lett., vol. 17, no. 4, pp. 861-863, Apr. 2005. (Suggests improving Berg's smoothed MMC by interpolating the log of the output PMF with a spline over N points. Berg's interpolation with a straight line over 2 points (Cfr. course notes, p. 40) is a special case).

[25] A. O. Lima, C. R. Menyuk, and I. T. lima, "Comparing two biasing monte carlo methods for calculating outage probabilities in sytems with multisection PMD compensators," Photon. Technol. Lett., vol. 17, no. 12, pp. 2580-2582, Dec. 2005. (Contains a comparison of efficiency of Multiple IS versus MMC).

[MCMC] **Markov Chain Monte Carlo: references cited in the course**

[26] A. Papoulis: *Probability, random variables, and stochastic processes*, 3rd Ed., McGraw-Hill, 1991. (On Rejection method, Cfr. course notes. pp. 56-57).

[27] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equations of state calculations by fast computing machines," J. Chem. Phys., vol. 21, no.6, pp. 1087-1092, June 1953. (Cfr. course notes p. 59).

[28] W. K. Hastings, "Monte Carlo sampling methods using markov chains and their applications," Biometrika, vol. 57, no. 1, pp. 97-109, Apr. 1970. (Cfr. course notes p. 59. Very well written and very instructive).

[29] A. Leon-Garcia: *Probability and random processes for electrical engineering*, 2rd Ed., Addison-Wesley, 1994. (On Time reversible markov chains, Cfr. course notes. p. 60).

[30] C. J. Geyer, "Markov chain monte carlo lecture notes" Course notes originally used Spring Quarter 1998, University of Minnesota, last typeset November 21, 2005. Available from the web. (Cfr pp 67-73. This is the clearest synthetic book I read on MCMC. To follow it properly, you need a bit of knowledge of probability theory based on measure. However Geyer makes all efforts to let also those with basic probability theory read the book with understanding. A very, very well written text!)

[31] Y. Guan, R. Fleissner, P. Joyce, and S. M. Krone, "Markov chain monte carlo in small worlds," Stat. Comput., vol. 16, no. 2, pp. 193-202, 2006. (Cfr. course notes, pp. 70-71).

[MCMC] **Markov Chain Monte Carlo: suggested further reading**

[32] S. Chib, and E. Greenberg, "Understanding the Metropolis-Hastings algorithm," *The American Statistician*, vol. 49, no. 4, pp. 327-335, Nov. 1995. (Good tutorial reading on the subject).

[33] L. Tierney, "Markov chain for exploring posterior distributions," *The Annals of Statistics*, vol. 22, no. 4, pp. 1701-1728, Dec. 1994. (Good reading on MCMC).

[34] S. Caracciolo, "A general limitation on Monte-Carlo algorithms of Metropolis type," Phys. Rev. Lett., vol. 72, no. 2, pp. 179-182. (A very elegant analysis of convergence time of MCMC).

[35] C. J. Geyer, "Practical Markov chain Monte Carlo," *Statistical Science*, vol. 7, no. 4, pp. 473-483, Nov. 1992. (Contains good discussion of convergence time of MCMC, and deals with restarts).

[36] A. Gelman, and D. B. Rubin, "Inference from iterative simulation using multiple sequences," *Statistical Science*, vol. 7, no. 4, pp. 457-472, Nov. 1992. (Contains discussion of convergence time with restarts).

[37] J. S. Liu, F. Liang, and W. H. Wong, "A theory for dynamic weighting in Monte Carlo Computation," J. Am. Stat. Assoc., vol. 96, no. 454, pp. 561-573, 2001. (Technique to accelerate convergence when sampling from multimode distributions).

[38] C. J. Geyer, and E. Thompson, "Annealing Markov chain Monte Carlo with applications to ancestral inference," J. Am. Stat. Assoc., vol. 90, no. 431, pp. 909-920, Sep. 1995. (Technique to accelerate convergence when sampling from multimode distributions. Known as Simulated Tempering, Cfr. Parallel Tempering).

Solutions to Proposed Exercises