Multicanonical Simulation Technique and its Relation to Importance Sampling

Prof. Alberto Bononi

Dipartimento di Ingengeria dell'Informazione
Università di Parma

Acknowledgments: N. Rossi, A. Orlandini, F. Vacondio, M. Franceschini

Outline

- Historical Background
- Motivation
- Monte Carlo (MC)
- Importance Sampling (IS)
- Flat Histogram (FH) Methods
    - Multicanonical Monte Carlo (MMC)
    - Fast MMC
    - Wang Landau (WL)
- Generatig warped Random Variables
    - Rejection Method
    - Markov Chain Monte Carlo (MCMC)
- MMC with MCMC
- Conclusions

Historical Background

- Monte-Carlo (MC) methods originated during World War II for calculations in nuclear physics at Los Alamos National Labs, where the first computer (ENIAC) was physically located.
- **1949**: N. Metropolis, S. Ulam, “The Monte-Carlo method” J. Am. Stat. Assoc.

  Authors named the method after the famous casino in Monaco.
- The problem: evaluation of stiff multiple integrals

$$I = \int_{\Gamma} f(\underline{x})\, d\underline{x}$$

over some domain Γ in R^n

- The solution:

  use a prob. density function (PDF) $p(\underline{x})$ on Γ (s.t. $f(\underline{x})\neq 0 \Rightarrow p(\underline{x})>0$) to get

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

Historical Background

- The random sampling idea is to estimate I as the sample mean:

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

from a series of samples $\{\underline{X}_i$, i=1,..,N$\}$ generated from the density p($\underline{x}$) (Horvitz-Thompson, J. Am. Stat. Assoc, 1952).

- 1949 MC proposal used a uniform density p($\underline{x}$).
  With non-uniform p($\underline{x}$), method is now known as *Importance Sampling* (IS).

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

- **1953** : N. Metropolis *et al.*, J. Chem. Phys.: proposed use in (1) of a **Markov Chain** $\{\underline{X}_i$, i=1,..,N$\}$ whose *steady-state* density is p($\underline{x}$). This is the most general Random Variable (RV) generation method known to date. Can generate (asymptotically) samples from *any* p($\underline{x}$).

  Method and its variations known as *Markov-Chain Monte-Carlo* (MCMC).

Historical Background

- **1992** : B. A Berg, T. Neuhaus, Phys. Rev. Lett.

  proposed an adaptive IS technique known as Multicanonical Monte Carlo (MMC) , which uses MCMC as the "engine" to generate warped densities in IS. The breaking novelty is the IS adaptation algorithm, based on the "flat-histogram" concept.

- Berg's papers are hard to read for non-physicists, and the probability theory ideas are hidden by the physical details of their problem. The method took a long time to escape from the physics community.

  Recently, other flat-histogram methods have emerged [e.g. Wang-Landau Phys. Rev Lett, 2001]. My exposure to MMC ideas comes from the optical communications literature [D. Yevick, Photon. Technol. Lett. 2002, R. Holzlohner *et al*, Opt. Lett. 2003].

- In this lecture, I will explain the MMC, a general method to estimate the PDF of the scalar output $Y=g(\underline{X})$ of a system g(.) with random input $\underline{X}$, initially without mentioning the MCMC engine, so that the flat-histogram based adaptive IS aspect becomes clear. Later I will devote time to MCMC.

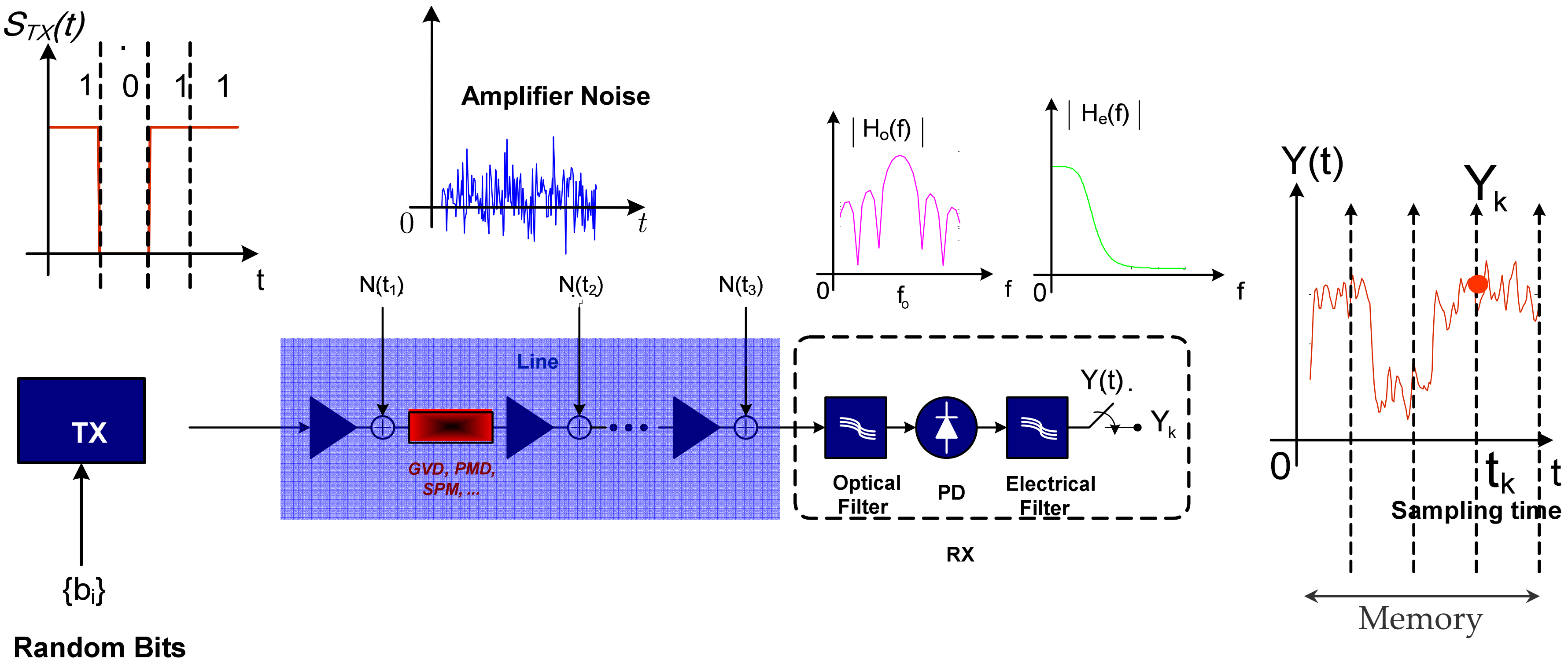
Outline

- Historical Background
- **Motivation**
- Monte Carlo (MC)
- Importance Sampling (IS)
- Flat Histogram (FH) Methods
  - Multicanonical Monte Carlo (MMC)
  - Fast MMC
  - Wang Landau (WL)
- Generatig warped Random Variables
  - Rejection Method
  - Markov Chain Monte Carlo (MCMC)
- MMC with MCMC
- Conclusions

Motivation

In telecommuincations, we often need to estimate the probability density function (PDF) of a random variable (RV) of interest, from which we derive the probability of rare events, such as errors, outages, buffer overflows,.......

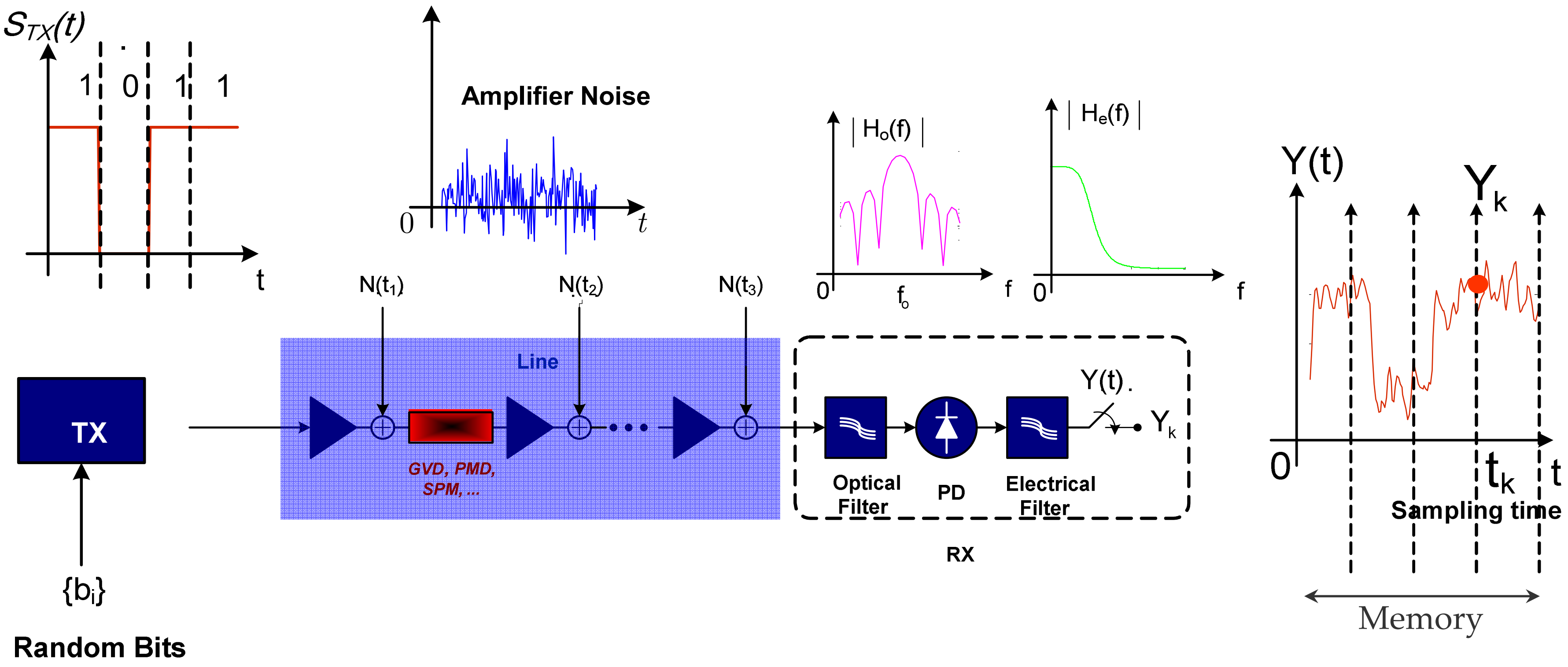
Motivation



Decision Variable $Y=g(\underline{X})$

"state" $\underline{X}$=(set of noise samples along the line
and of random bits adjacent to bit of interest
falling within system memory)

Motivation



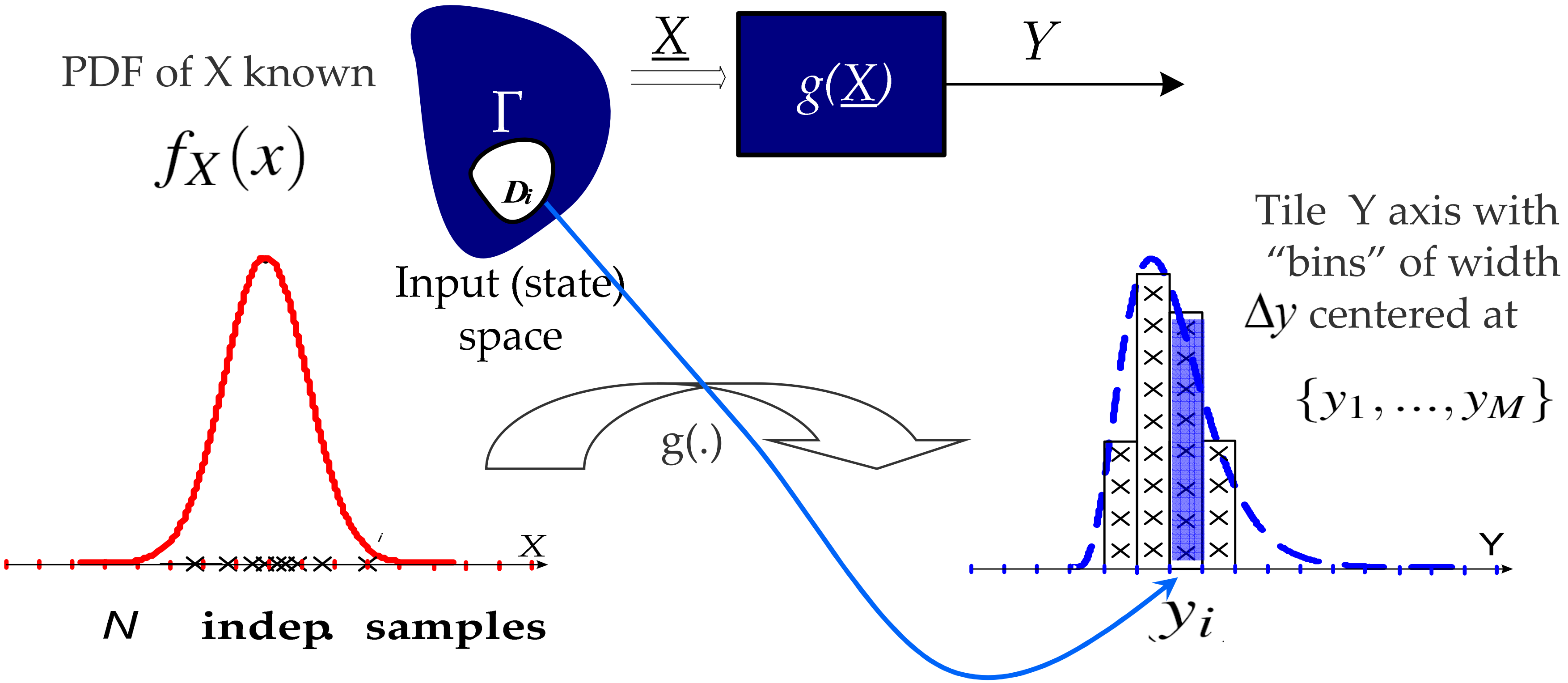
$$Y=g(\underline{X})$$

- g(.) is "costly" to compute (simulate)
- The longer the memory, the larger the dimensionality of the state

Outline

- Historical Background
- Motivation
- **Monte Carlo (MC)**
- Importance Sampling (IS)
- Flat Histogram (FH) Methods
    - Multicanonical Monte Carlo (MMC)
    - Fast MMC
    - Wang Landau (WL)
- Generatig warped Random Variables
    - Rejection Method
    - Markov Chain Monte Carlo (MCMC)
- MMC with MCMC
- Conclusions

Monte Carlo (MC)



PDF of X known

$f_X(x)$

Input (state) space

Tile Y axis with "bins" of width Δy centered at

$\{y_1, \ldots, y_M\}$

N indep. samples

Thereby estimate the probability mass function (PMF) of discretized Y:

$$P_i \equiv P\{Y \approx y_i\}$$

and evaluate PDF from PMF as $f_Y(y_i) \cong P_i / \Delta y$

Monte Carlo

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

Indicator of i-th bin visit (FLAG)

Probability that a sample falls in bin i:

$$P_i = \int_{D_i} f_X(x)dx = \int_{\Gamma} I_i(g(x)) f_X(x)dx \; = E[I_i(g(X))]$$

$$\Downarrow$$

Sample mean of RV $I_i(g(X))$

$$\boxed{\hat{P}_i^{MC} \triangleq \frac{1}{N}\sum_{j=1}^{N} I_i(g(X_j)) = \frac{N_i}{N}}$$

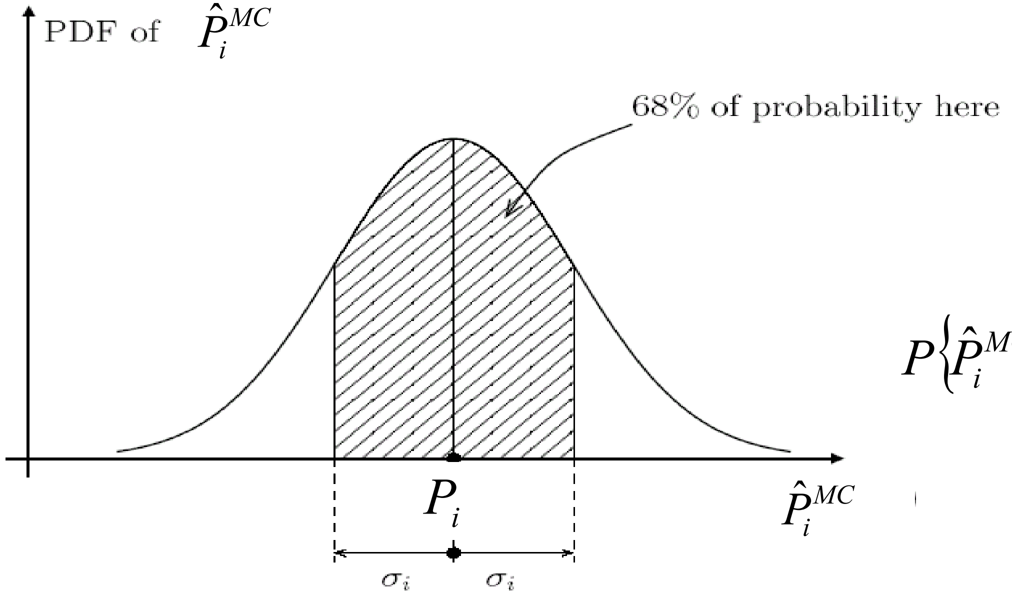
$N_i \sim Binomial(N,P_i)$ [# successes in N indep. trials, success prob. P_i]

$$(2) \quad \begin{cases} E[\hat{P}_i^{MC}] = P_i & \text{unbiased estimator} \\ Var[\hat{P}_i^{MC}] = \dfrac{P_i(1-P_i)}{N} \end{cases}$$

Monte Carlo: Accuracy

Define $\varepsilon_i \triangleq Var[\hat{P}_i]/P_i^2$ quadratic relative error of an unbiased estimator

As N→∞ , by Central Limit $\hat{P}_i^{MC} \rightarrow Normal\left(P_i, \sigma_i^{\,2} = \dfrac{P_i(1-P_i)}{N}\right)$



68%-Confidence interval

Confidence interval

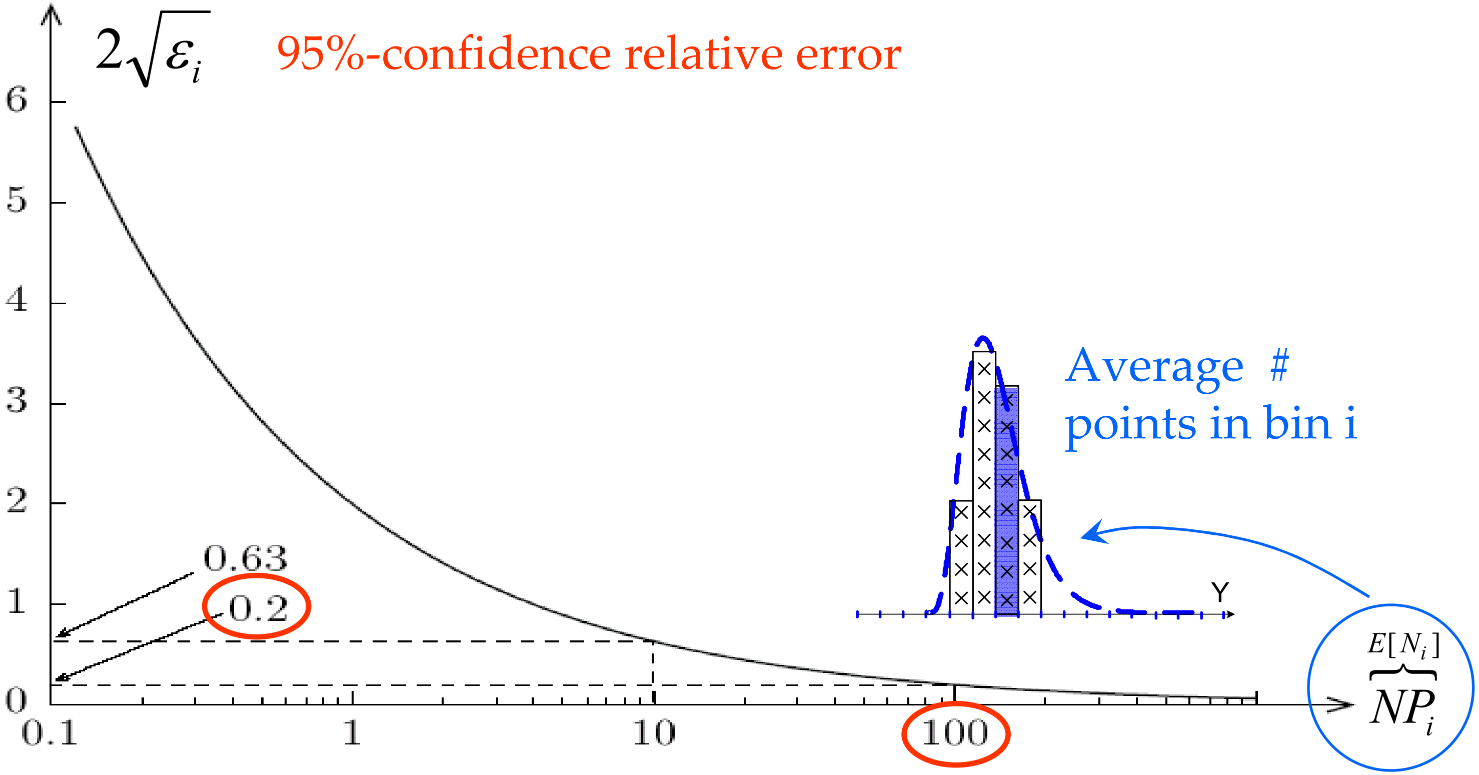
Confidence level

$$P\left\{\hat{P}_i^{MC} \in (P_i \pm \sigma_i) = P_i\left(1 \pm \sqrt{\varepsilon_i}\right)\right\} \cong 0.68$$

$$P_i\left(1 \pm 2\sqrt{\varepsilon_i}\right)\Big\} \cong 0.95$$

Monte Carlo: Accuracy

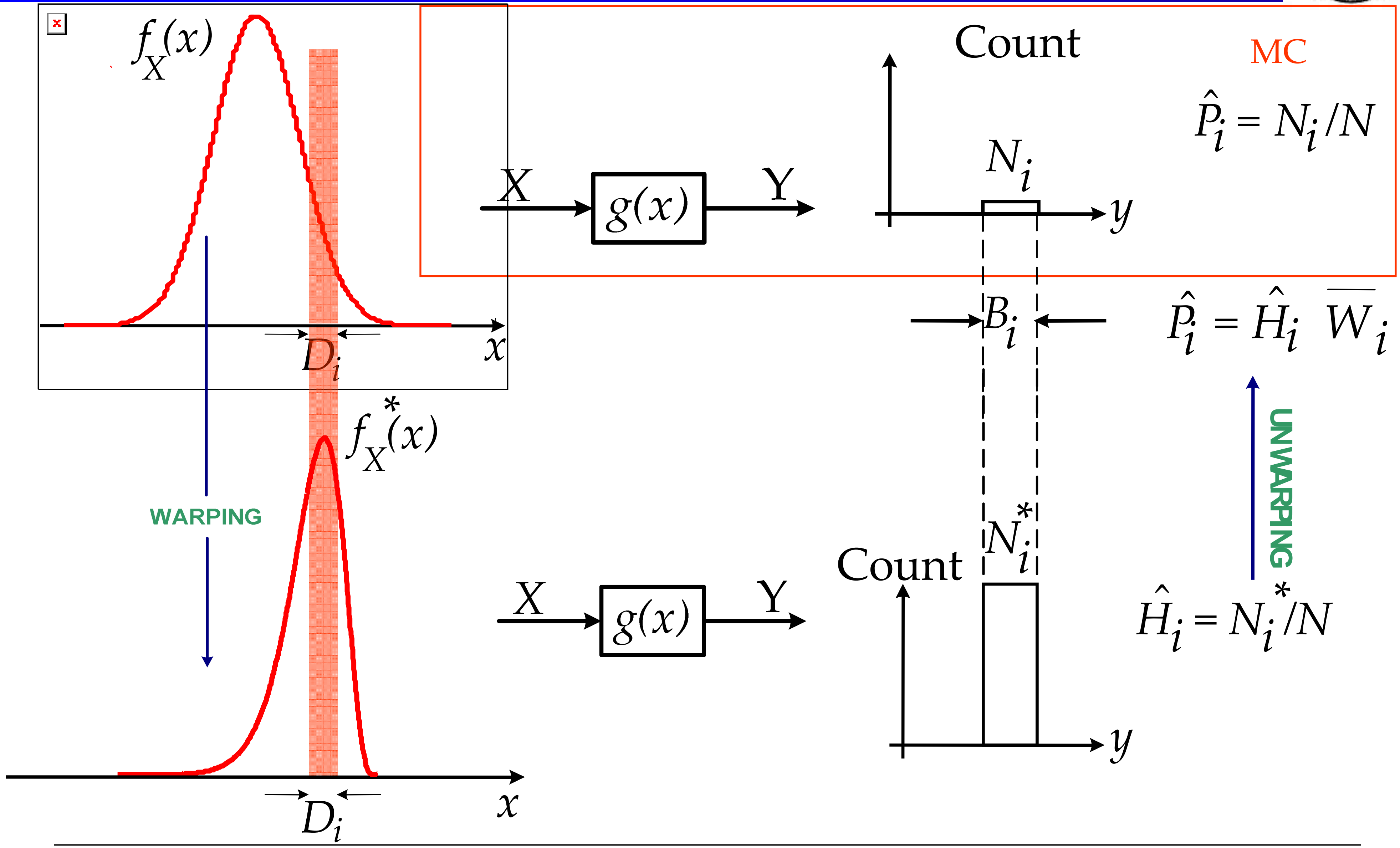
From (2): $\boxed{\varepsilon_i^{MC} = \frac{1-P_i}{NP_i}}$



Outline

- Historical Background
- Motivation
- Monte Carlo (MC)
- **Importance Sampling (IS)**
- Flat Histogram (FH) Methods
  - Multicanonical Monte Carlo (MMC)
  - Fast MMC
  - Wang Landau (WL)
- Generatig warped Random Variables
  - Rejection Method
  - Markov Chain Monte Carlo (MCMC)
- MMC with MCMC
- Conclusions

Not enough points in bin i ? ⇒ Importance Sampling !



Importance Sampling

Coefficient $\overline{w}_i$ to unwarp formally found as follows:

$$P_i = \int_\Gamma I_i(g(x)) \left[\frac{f_X(x)}{f_X^*(x)}\right] f_X^*(x)dx = E^*[I_i(g(X))\boxed{w(X)}]$$

IS weight (known function, hopefully << 1 on D$_i$)

Estimate P$_i$ as sample mean of RV I$_i$(g(X))w(X)

$$\boxed{\hat{P}_i^{IS} \triangleq \frac{1}{N}\sum_{j=1}^{N} I_i(g(X_j))w(X_j) = \underbrace{\left(\frac{N_i^*}{N}\right)}_{\hat{H}_i} \underbrace{\left[\frac{1}{N_i^*}\sum_{n=1}^{N_i^*} w(X_n)\right]}_{\overline{w}_i}}$$

IS Estimator

Histogram of visits

Average weight on bin i

Importance Sampling

Rationale behind IS estimator:

$$P_i = \int_{D_i} \left[\frac{f_X(x)}{f_X^*(x)} \right] f_X^*(x) dx = \overbrace{\left[\int_{D_i} f_X^*(x) dx \right]}^{H_i = E^*\left[\hat{H}_i\right]} \underbrace{\int_{D_i} [w(x)] \frac{f_X^*(x)}{\left[\int_{D_i} f_X^*(x) dx \right]} dx}_{E^*\left[w(X) \; \middle| X \in D_i\right]}$$

Similarly, it is easy to prove that

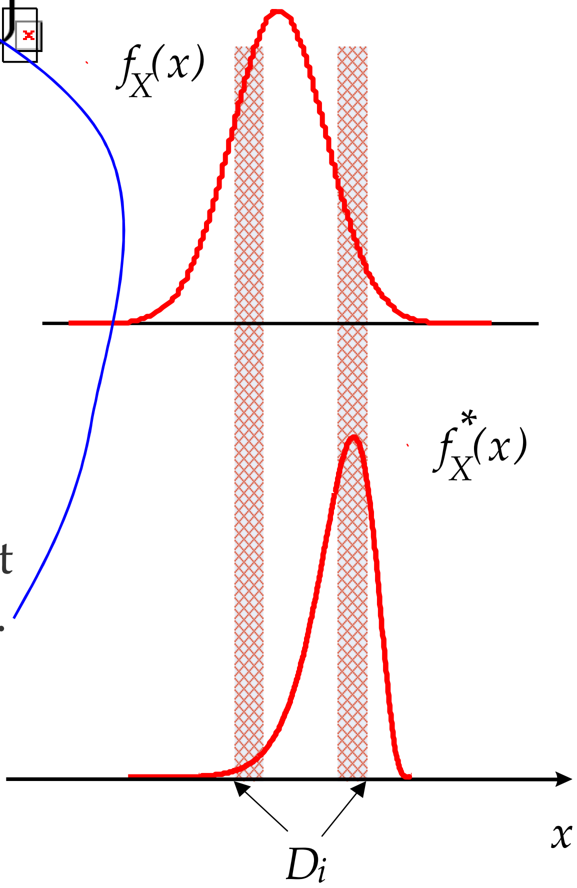
$$\begin{cases} E^*[\hat{P}_i^{IS}] = P_i & \text{unbiased estimator, like MC} \\ Var^*[\hat{P}_i^{IS}] = \dfrac{\left(H_i E^*\left[w^2(X) \; \middle| X \in D_i\right] - P_i^2\right)}{N} \end{cases}$$

and get

$$\varepsilon_i^{IS} = \frac{1}{N} \left\{ \frac{1}{H_i} \left(\frac{Var^*[w(X)|X \in D_i]}{(E^*[w(X)|X \in D_i])^2} + 1 \right) - 1 \right\}$$

Importance Sampling: Accuracy

$$\varepsilon_i^{IS} = \frac{1}{N}\left\{\frac{1}{H_i}\left(\frac{\boxed{Var^*[w(X)|X \in D_i]}}{(E^*[w(X)|X \in D_i])^2}+1\right)-1\right\}$$



True limit of IS is our a priori ignorance of domains D_i. Hence may get the wrong warping, assigning widely different weights over the same D_i, thus increasing the variance...........

Uniform Weight IS (UWIS)

$$\varepsilon_i^{IS} = \frac{1}{N}\left\{\frac{1}{H_i}\left(\frac{\boxed{Var^*[w(X)|X \in D_i]}}{(E^*[w(X)|X \in D_i])^2} + 1\right) - 1\right\} \qquad \longrightarrow 0$$

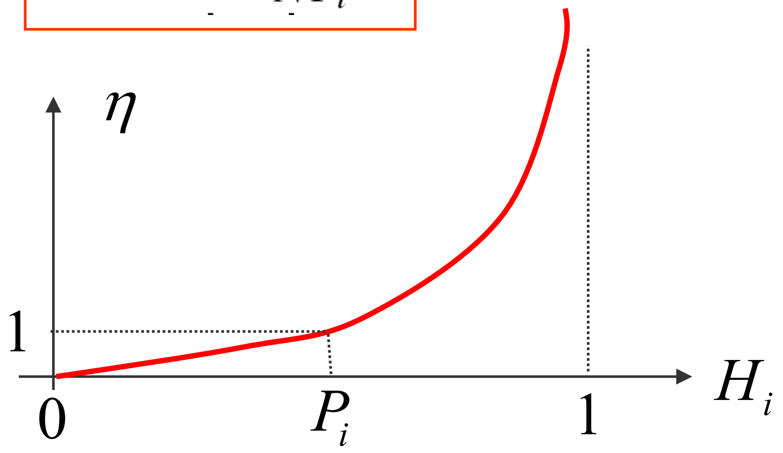
- Best warpings give uniform weight over whole D_i (UWIS = uniform weight IS)
- UWIS can be realized by using the MCMC engine, even without global knowledge of domains D_i, as we will see.

$$\varepsilon_i^{UWIS} = \frac{1}{N}\left\{\frac{1}{H_i} - 1\right\}$$

$$\boxed{\varepsilon_i^{MC} = \frac{1-P_i}{NP_i}}$$

Efficiency:

$$\eta = \frac{N_{MC}\,\varepsilon_i^{MC}}{N_{UWIS}\,\varepsilon_i^{UWIS}} = \frac{1-P_i}{1-H_i}\,\frac{H_i}{P_i}$$



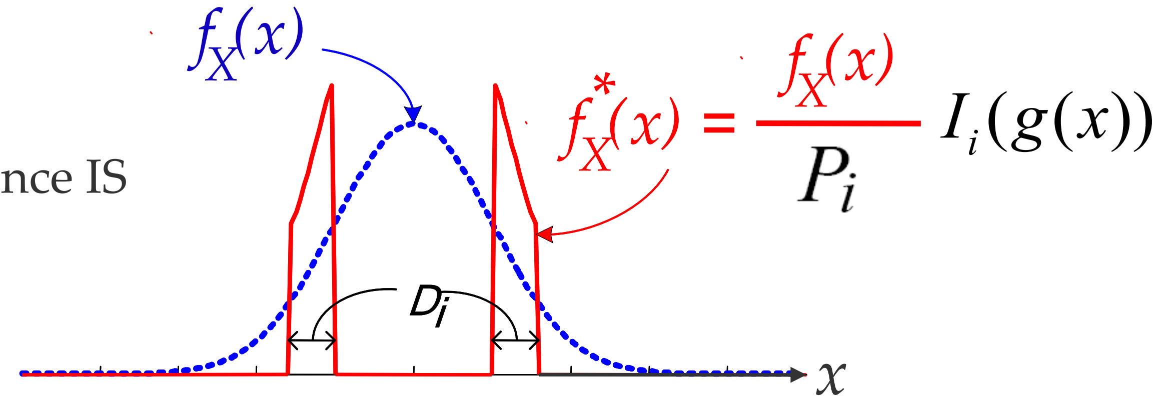
gain over MC can be impressive when $H_i >> P_i$

Zero Variance IS (ZVIS)

$$\varepsilon_i^{UWIS} = \frac{1}{N}\left\{\frac{1}{H_i} - 1\right\}$$

If $H_i = 1$

Zero Variance IS



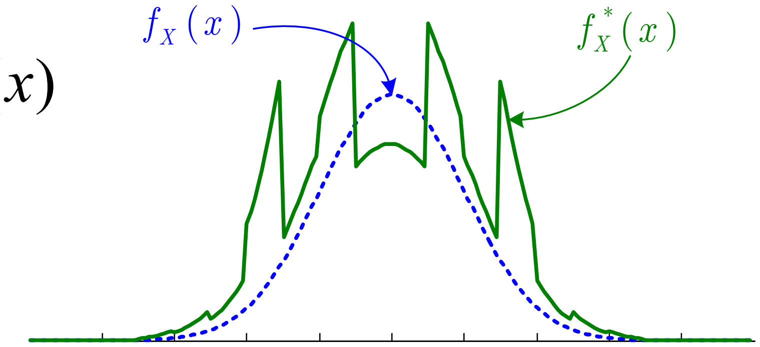
$$f_X^*(x) = \frac{f_X(x)}{P_i} I_i(g(x))$$

This is optimal (exclusively for estimate of bin i).
All samples fall within D_i !
Not realizable, as requires knowledge of P_i, ie, of what we wish to estimate...

Flat Histogram IS (FHIS)

An “optimal” UWIS exists to estimate of whole PMF $\{P_1, P_2, \ldots, P_M\}$ of Y. Obtained by adding up ZVIS of all bins and renormalizing:

$$f_X^*(x) = \frac{1}{M}\underbrace{\sum_{i=1}^{M}\frac{I_i(g(x))}{P_i}}_{} f_X(x)$$



Introducing the staircase function:

$$P(y) \equiv \sum_{i=1}^{M} P_i I_i(y)$$

...we see that $\frac{1}{P(g(x))}$

Hence

$$\boxed{f_X^*(x) = \frac{f_X(x)}{MP(g(x))}}$$

Not realizable, like ZVIS, but can be approximated, as we will see...

$P(g(x))$ is probability of the bin where $y = g(x)$ falls

It is UWIS since $w(x) = MP_i$ for every $x \in D_i$

Flat Histogram IS

Properties of FHIS:

1)
$$\overbrace{H_i}^{\boxed{= E^*\left[\hat{H}_i\right]}} = \int_{D_i} f_X^* dx = \int_{D_i} \frac{f_X(x)}{MP(g(x))} dx$$

$$= \frac{\int_{D_i} \cancel{f_X}(x)dx}{M\cancel{P_i}} = \frac{1}{M}$$

for all bins i :
flat vists histogram (FH)
on average

2)
$$\varepsilon^{FHIS} = \frac{1}{N}\left\{\frac{1}{H_i} - 1\right\} = \frac{M-1}{N}$$

for all bins i :
same relative precision!

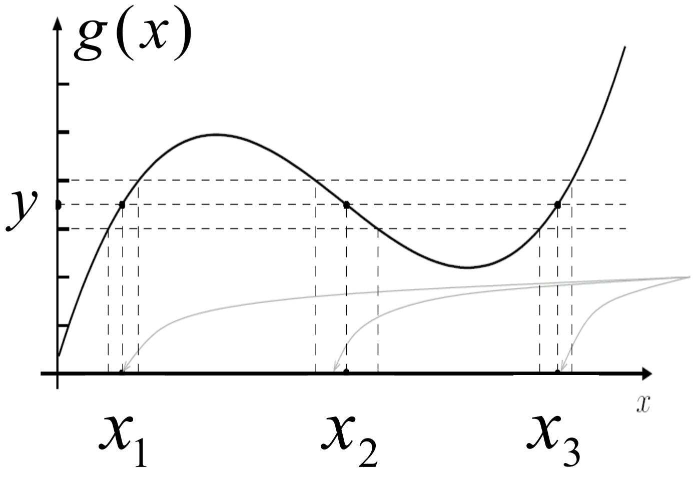
That’s the best that one can do with the N samples !!!

FHIS: Alternative Proof

The Flat Histogram result can be proven even without quantization of Y:

(3) $$f^*_X(x) = \frac{f_X(x)}{c\,\boxed{f_Y(g(x))}} \longrightarrow \boxed{g(.)} \longrightarrow f^*_Y(y) = ?$$

Use the Fundamental Theorem (in 1D for simplicity):



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

$$= \cancel{\left[\sum_{k=1}^{K} \frac{f_X(x_k)}{|g'(x_k)|}\right]} \cdot \frac{1}{c f_Y\cancel{(y)}} = \frac{1}{c} \quad \text{flat!}$$

And vice-versa: if $f^*_Y(y)$ is flat, then (3) must hold. Hence when output warped PDF flattens on a certain range, then one can read-off the correct $\boxed{f_Y(y) \text{ from the denominator}}$ in $f^*_X(x)$

Outline

- Historical Background
- Motivation
- Monte Carlo (MC)
- Importance Sampling (IS)
- **Flat Histogram (FH) Methods**
    - Multicanonical Monte Carlo (MMC)
    - Fast MMC
    - Wang Landau (WL)
- Generatig warped Random Variables
    - Rejection Method
    - Markov Chain Monte Carlo (MCMC)
- MMC with MCMC
- Conclusions

Flat Histogram Methods (FH)

A family of algorithms (MMC, Wang-Landau and others), which, starting from known PDF of X, $f_X(x)$ build a sequence of uniform-weight warped PDFs

$$f_X^{(n+1)}(x) = \boxed{\frac{f_X(x)}{c_n \Theta_n(g(x))}}, \; n = 0, 1, 2, \ldots \qquad \text{(UW)}$$

where $\underline{\Theta}_n \triangleq \{\Theta_n(y_i)\}_{i=1}^{M}$ is an estimate of PMF of Y at cycle n

and c_n a normalization constant, from which we draw samples to form a new estimate $\underline{\Theta}_{n+1}$ of PMF of Y, up to convergence to FH:

$$f_X^{*}(x) = \boxed{\frac{f_X(x)}{MP(g(x))}} \qquad \text{(FH)}$$

At convergence (empirically verified by a flat visits histogram on average) have:

$$c_n \longrightarrow M \qquad \underline{\Theta}_n \longrightarrow \underline{\mathsf{P}} \triangleq \{P_i\}_{i=1}^{M}$$

Algorithms differ in their update law $\quad \underline{\Theta}_n \longrightarrow \underline{\Theta}_{n+1}$

Flat Histogram Methods: Convergence

The "average drift" towards FH can be understood by calculating the probability that samples fall in bin i during cycle n+1

$$H_i^{(n+1)} = \int_{D_i} \left[\frac{f_X(x)}{c_n \Theta_n(g(x))} \right] dx = \frac{1}{c_n} \frac{\int_{D_i} f_X(x)dx}{\Theta_n(y_i)} = \frac{P_i}{c_n \Theta_n(y_i)} \qquad (*)$$

where normalizing constant must be

$$c_n = \sum_{j=1}^{M} \frac{P_j}{\Theta_n(y_j)}$$

From (*) we realize that, during n-th cycle, the N samples will fall (on average) mostly in *under-estimated* bins $\Theta_n(y_i) << P_i$ and much less in *over-estimated* bins.

If the update $\underline{\Theta}_n \rightarrow \underline{\Theta}_{n+1}$ is based on flattening the visits histogram $\left\{ \hat{H}_i^{(n+1)}, i = 1,.., M \right\}$

(recall that $H_i^{(n+1)} = E^*[\hat{H}_i^{(n+1)}]$)

then equilibrium will be reached when $\Theta_n(y_i) = P_i$ in all bins.

Outline

- Historical Background
- Motivation
- Monte Carlo (MC)
- Importance Sampling (IS)
- Flat Histogram (FH) Methods
    - **Multicanonical Monte Carlo (MMC)**
    - Fast MMC
    - Wang Landau (WL)
- Generatig warped Random Variables
    - Rejection Method
    - Markov Chain Monte Carlo (MCMC)
- MMC with MCMC
- Conclusions

Multicanonical Monte Carlo (MMC)

- **First FH method, invented by physicist Berg in 1992.**
- **Update law based on a UWIS estimate:**

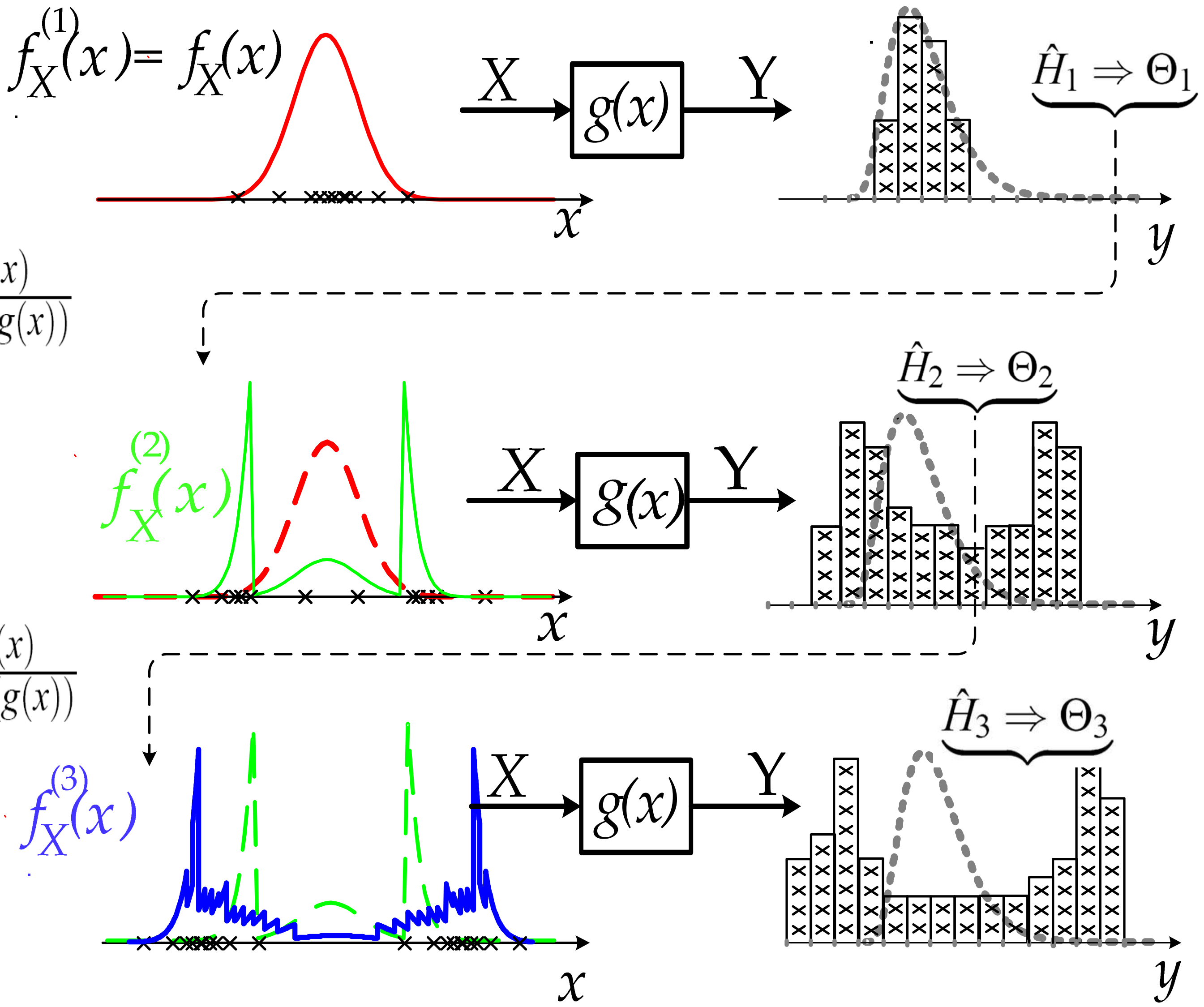
1. At step (or cycle) *n+1*, *N* samples drawn from $f_X^{(n+1)}(x) = \frac{f_X(x)}{c_n \Theta_n(g(x))}$
2. For every sample calculate $Y_j = g(X_j)$
3. from these evaluate visits histogram $\hat{H}_i^{(n+1)} = N_i^{(n+1)} / N$
4. updated IS estimate of PMF of Y is finally given by

$$\Theta_{n+1}(y_i) = \underbrace{\left(\frac{N_i^{(n+1)}}{N}\right)}_{\hat{H}_i^{(n+1)}} \underbrace{\left[\frac{1}{N_i^{(n+1)}} \sum_{n=1}^{N_i^{(n+1)}} w(X_n)\right]}_{c_n \Theta_n(y_i)}$$

UWIS estimate

Intuition on MMC

Θ_0
Uniform


$f_X^{(1)}(x) = f_X(x)$
x
X
$g(x)$
Y
y
$\hat{H}_1 \Rightarrow \Theta_1$
$f_X^{(2)}(x)$
x
X
$g(x)$
Y
y
$\hat{H}_2 \Rightarrow \Theta_2$
$f_X^{(3)}(x)$
x
X
$g(x)$
Y
y
$\hat{H}_3 \Rightarrow \Theta_3$

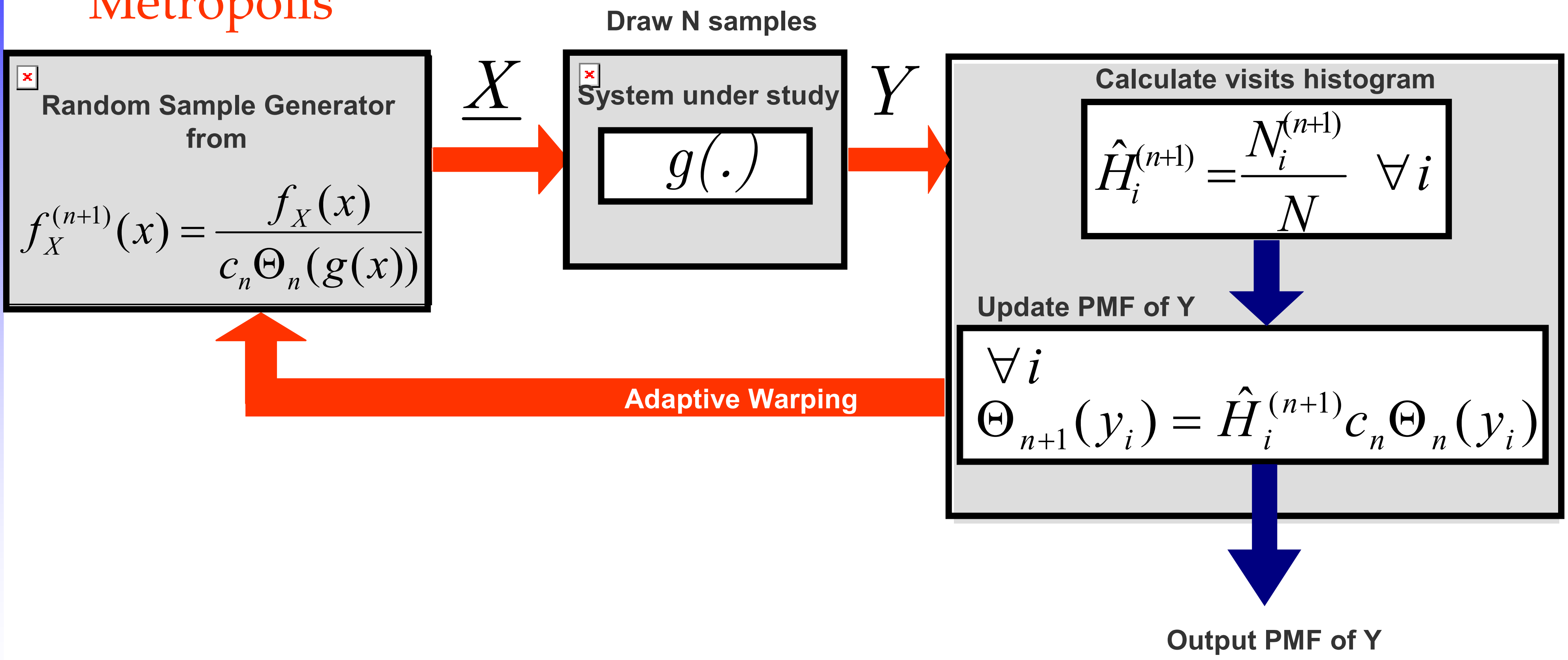
$$f_X^{(n+1)}(x) = \frac{f_X(x)}{c_n \Theta_n(g(x))}$$

$$f_X^{(n+1)}(x) = \frac{f_X(x)}{c_n \Theta_n(g(x))}$$

1 cycle is pure MC

MMC Block-Diagram

Metropolis



Random Sample Generator from

$$f_X^{(n+1)}(x) = \frac{f_X(x)}{c_n \Theta_n(g(x))}$$

$\underline{X}$

Draw N samples

System under study

$g(.)$

Y

Calculate visits histogram

$$\hat{H}_i^{(n+1)} = \frac{N_i^{(n+1)}}{N} \quad \forall i$$

Update PMF of Y

$$\forall i \quad \Theta_{n+1}(y_i) = \hat{H}_i^{(n+1)} c_n \Theta_n(y_i)$$

Adaptive Warping

Output PMF of Y

MMC “engine”: Metropolis

Metropolis is an algorithm that produces correlated samples $\{X_1,X_2,,....,X_N\}$ as a reversible Markov Chain whose steady-state distribution is the desired PDF $f^*_X(x)$.

At each time step m, if $X_{m-1}= x_i$ is the initial state, a next state is proposed as

$$x_f = x_i + U_m$$

where typically U_m is a uniform RV used to “explore” the state space around $x_{i..}$

The odds ratio for move $i \to f$ is formed as

$$R = \frac{f_X^*(x_f)}{f_X^*(x_i)}$$

Then the proposal is either accepted with probability min(1,R) and we set $X_m = x_f$ or else the proposal is rejected and we keep the initial value: $X_m = x_i$

MMC "engine": Metropolis

Hence in cycle (*n+1*) of MMC, Metropolis generation uses the odds ratio

$$R = \frac{f_X^{(n+1)}(x_f)}{f_X^{(n+1)}(x_i)} = \frac{f_X(x_f)}{\cancel{c_n}\Theta_n(g(x_f))} \bullet \frac{\cancel{c_n}\Theta_n(g(x_i))}{f_X(x_i)}$$

We make 3 important points:

1) The constant c_n cancels out and need not be computed

2) Samples from the UW warped PDF can be generated without knowledge of the domains D_i. In fact, R can be evaluated by simply computing $g(x_i)$, $g(x_f)$ and checking the bin they fall into.

3) Samples correlations increase the variance w.r.t IID samples. For instance, if an estimator has the form $\hat{E}_N = \frac{1}{N}\sum_{n=1}^{N} h(X_n)$ then its variance is

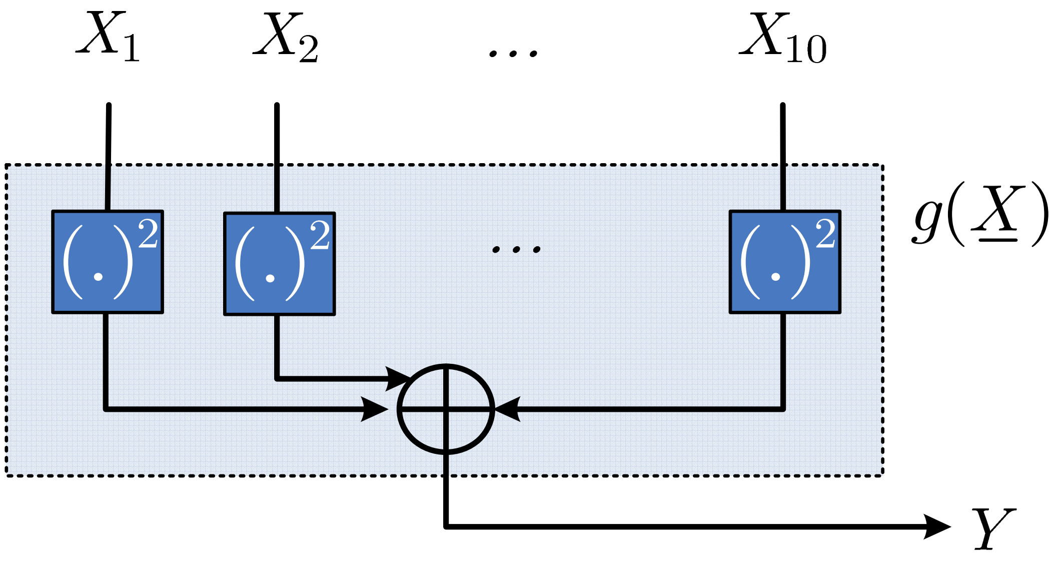
$$Var\left[\hat{E}_N\right] = \frac{1}{N}\left\{ Var\left[h(X_i)\right] + \underbrace{2\sum_{n=1}^{N-1}\frac{N-n}{N} Cov\left[h(X_i), h(X_{i+n})\right]}_{generally\ >0,\ \text{var}\ iance\quad increased\ \ !!} \right\}$$

Example: Nonlinear Memoryless System

Input: a 10-dimensional vector of normal IID RVs:

$$\underline{X} = \left[X_1 \quad X_2 \quad \dots \quad X_{10} \right]$$

$$X_i \sim N(0,1)$$

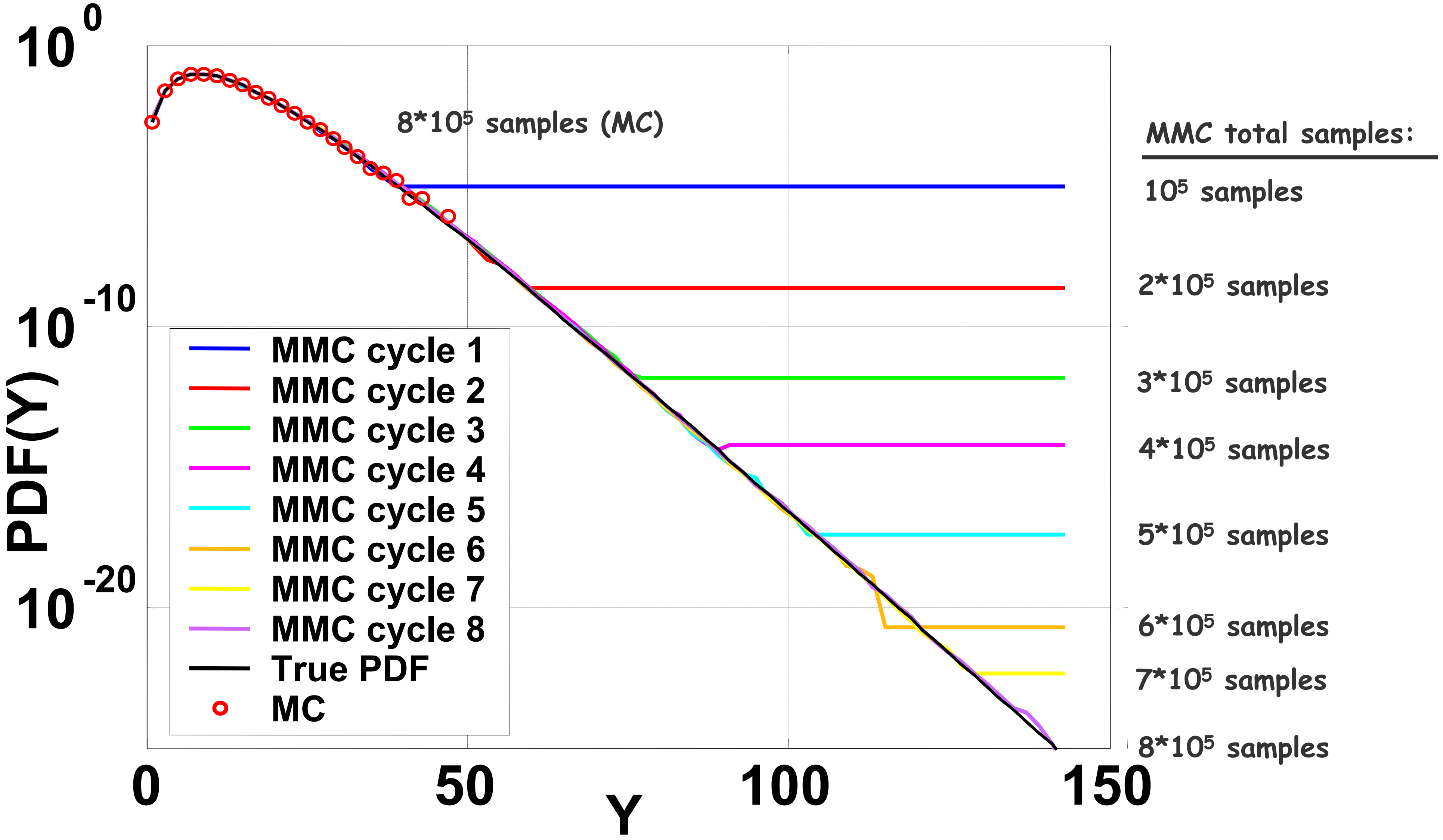


System: $g(\underline{X}) = \sum_{i=1}^{10} X_i^2$

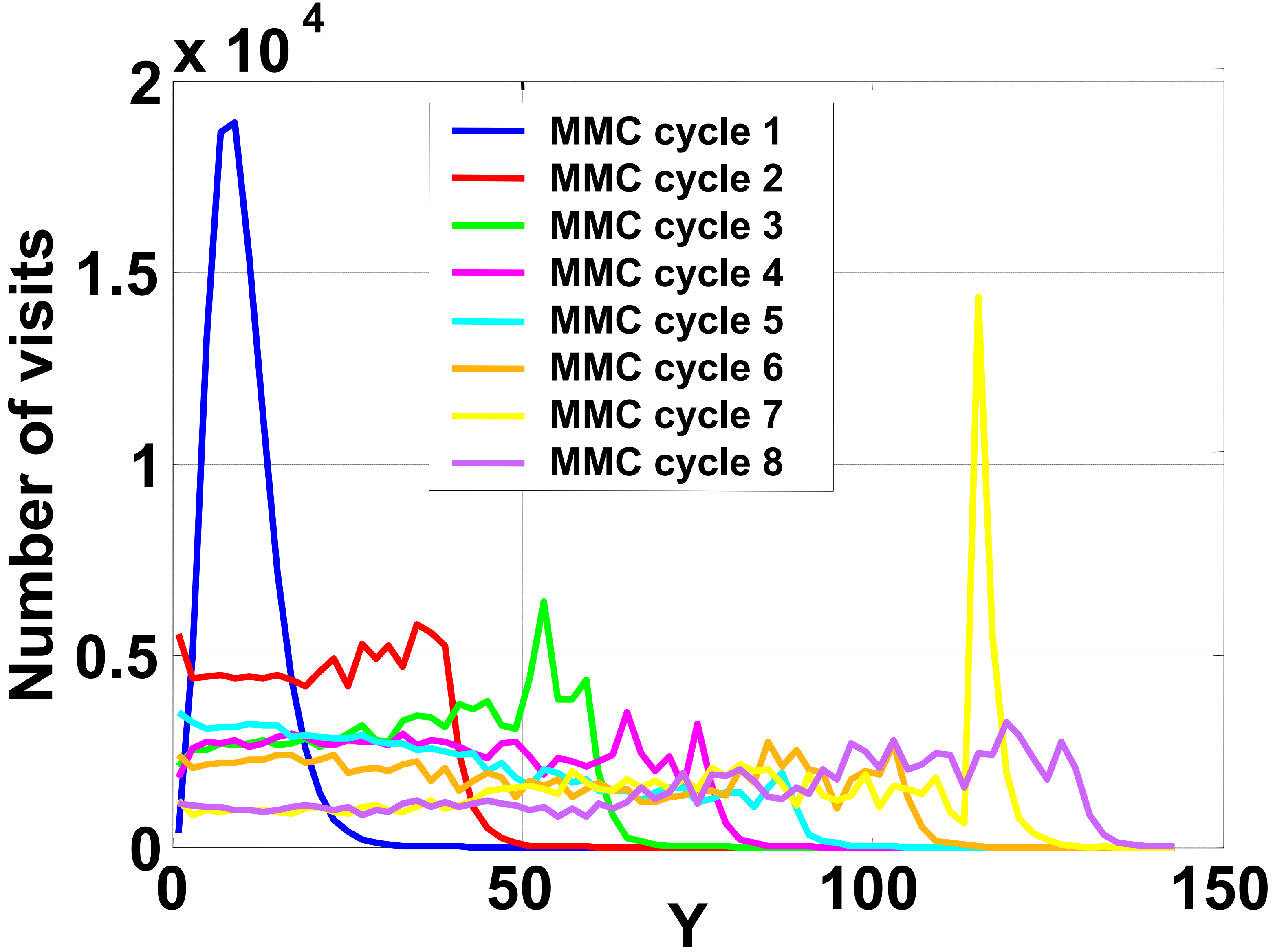
Output: $Y = g(\underline{X})$ of which we know the statistics:

$$Y \sim \chi^2(10)$$

Example 1: Standard MMC estimate



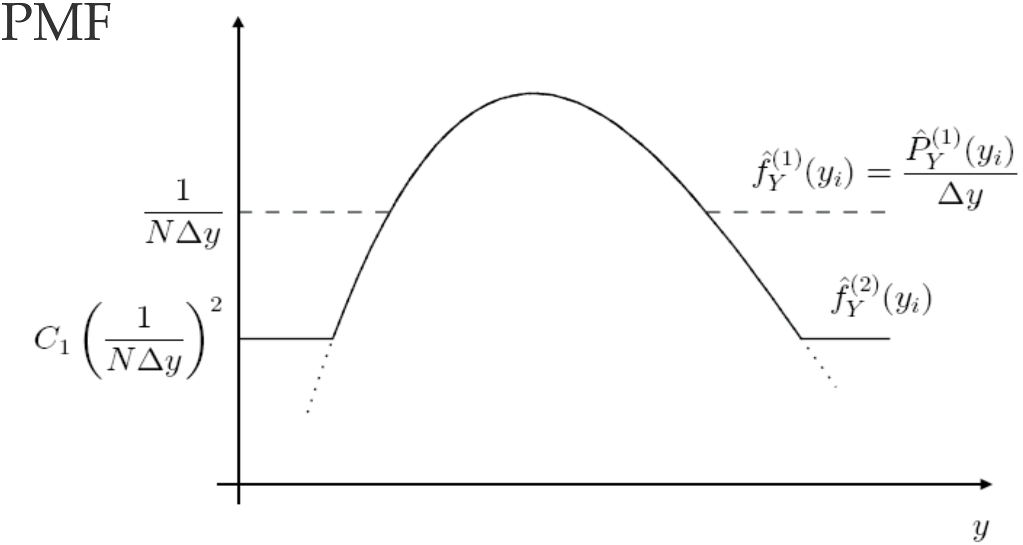
Example 1: Standard MMC Visits Histogram



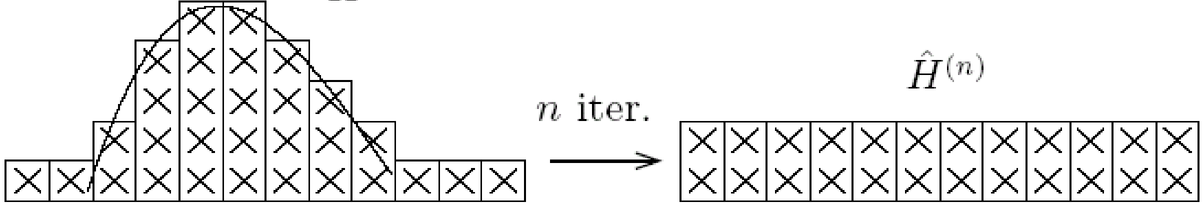
Standard MMC: Notes

- On bins in which no visits are present, the PMF of previous bin value is used: hence the “floors” in estimated PMF

- amount of “descent” per cycle scales with cycle size N



Empirical choice of N:



If a squared relative error ε^{FHIS} is desired and have M bins, then

$$N = \frac{M-1}{\varepsilon^{FHIS}}$$

Outline

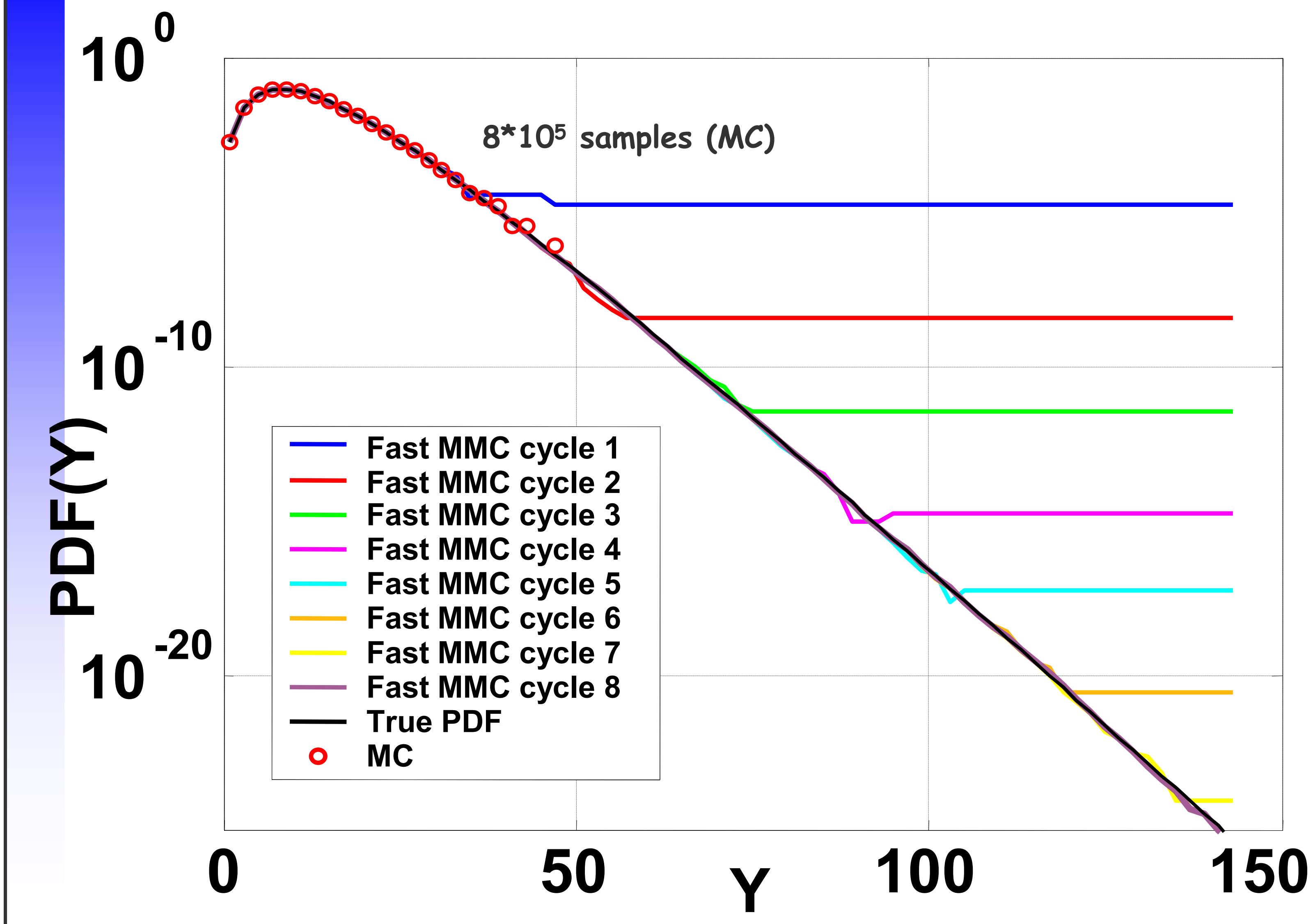
- Historical Background
- Motivation
- Monte Carlo (MC)
- Importance Sampling (IS)
- Flat Histogram (FH) Methods
  - Multicanonical Monte Carlo (MMC)
  - Fast MMC
  - Wang Landau (WL)
- Generatig warped Random Variables
  - Rejection Method
  - Markov Chain Monte Carlo (MCMC)
- MMC with MCMC
- Conclusions

Fast MMC

MMC has the evident drawback that modal region is visited and thus re-estimated at every cycle ⇒waste of samples!

We recently proposed a novel algorithm (A. Bononi *et al.*, Fotonica ‘07) that prevents MMC from visiting regions of Y range over which estimated PDF has already converged in previous cycles.

Example 1: Fast MMC estimate



FMMC total samples:

10^5 samples

$2*10^5$ samples

$3*10^5$ samples

$4*10^5$ samples

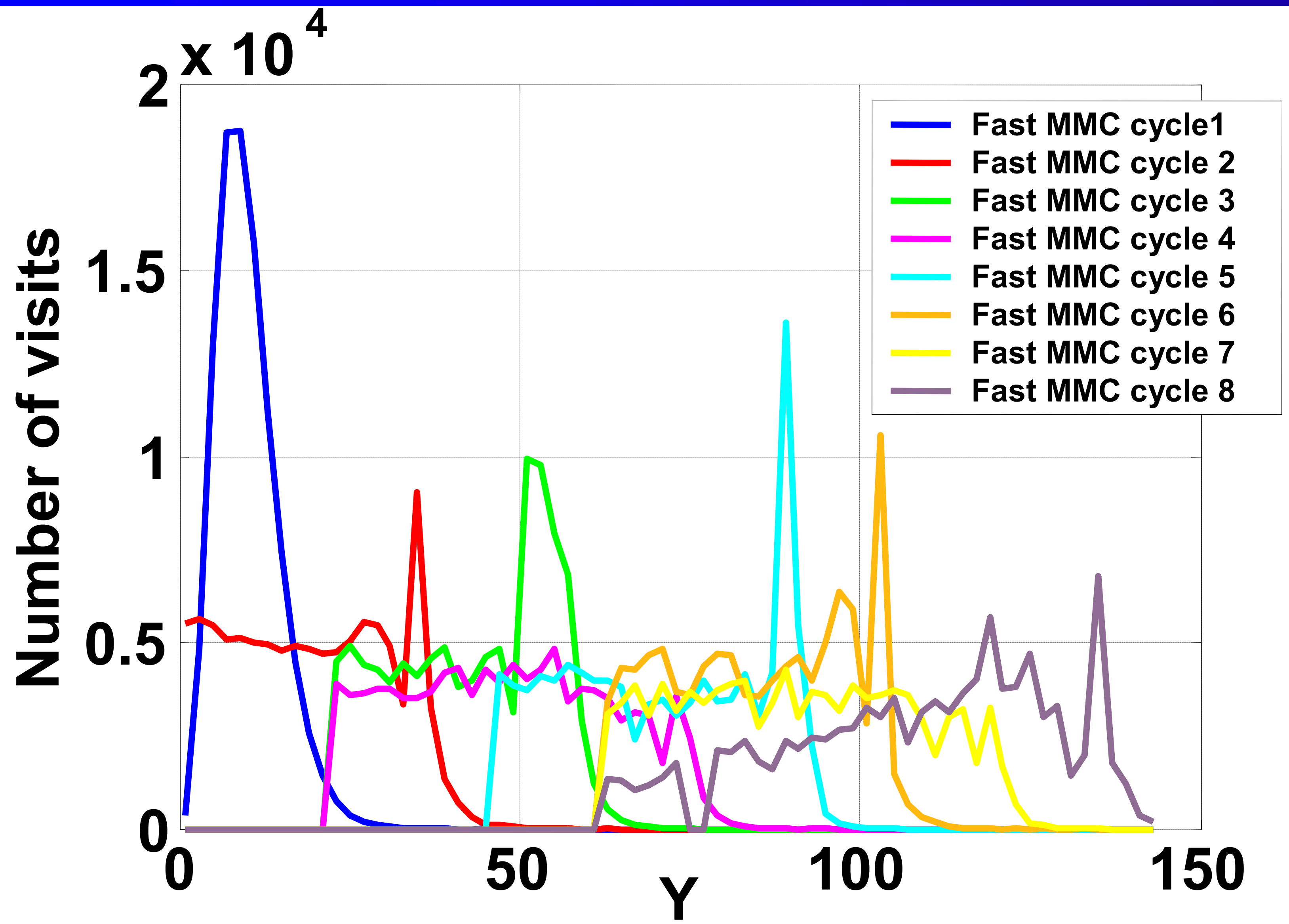
$5*10^5$ samples

$6*10^5$ samples

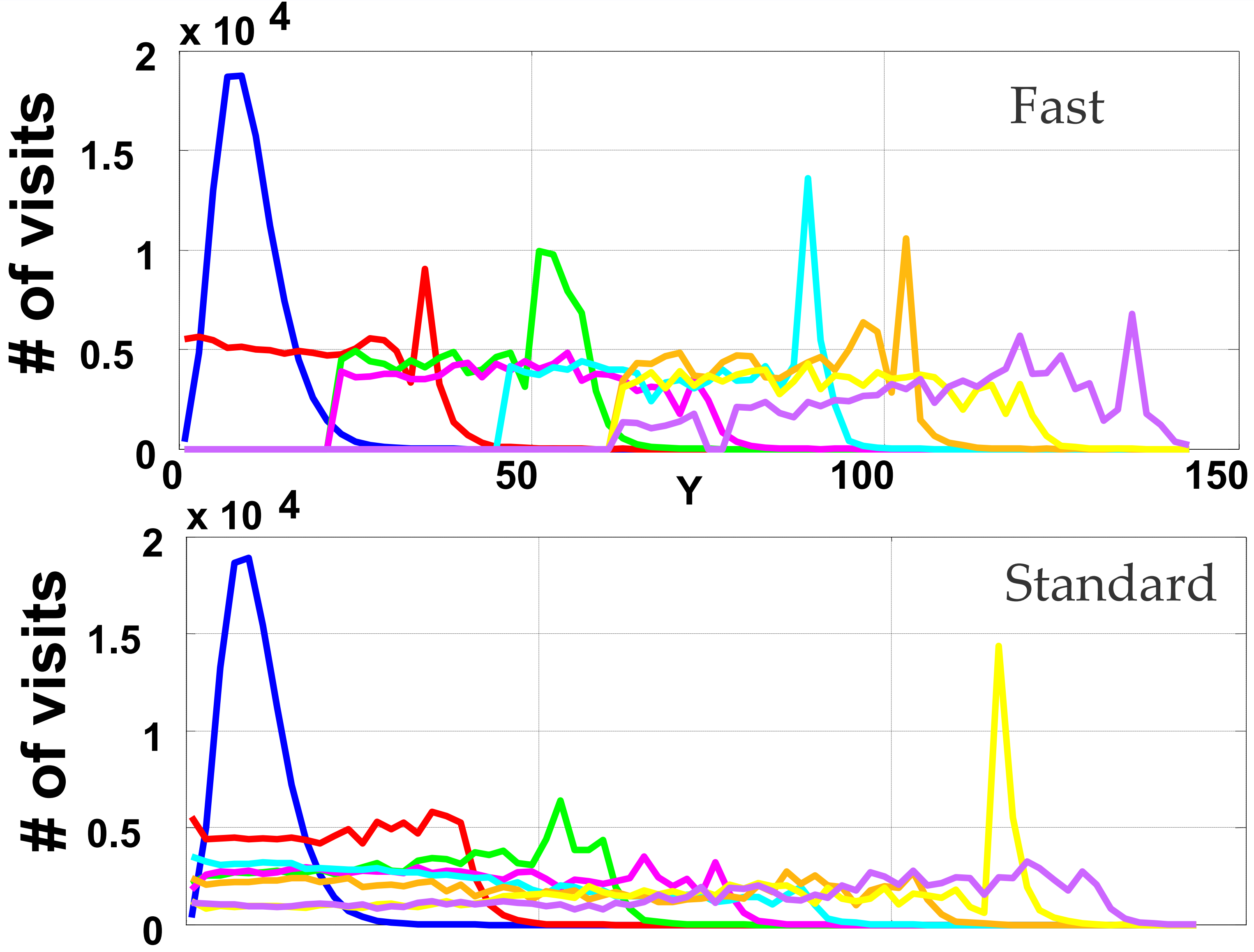
$7*10^5$ samples

$8*10^5$ samples

Example 1: Fast MMC Visits Histogram



Example 1: Fast MMC vs Standard MMC



Outline

- Historical Background
- Motivation
- Monte Carlo (MC)
- Importance Sampling (IS)
- Flat Histogram (FH) Methods
  - Multicanonical Monte Carlo (MMC)
  - Fast MMC
  - **Wang Landau (WL)**
- Generatig warped Random Variables
  - Rejection Method
  - Markov Chain Monte Carlo (MCMC)
- MMC with MCMC
- Conclusions

Wang-Landau (WL)

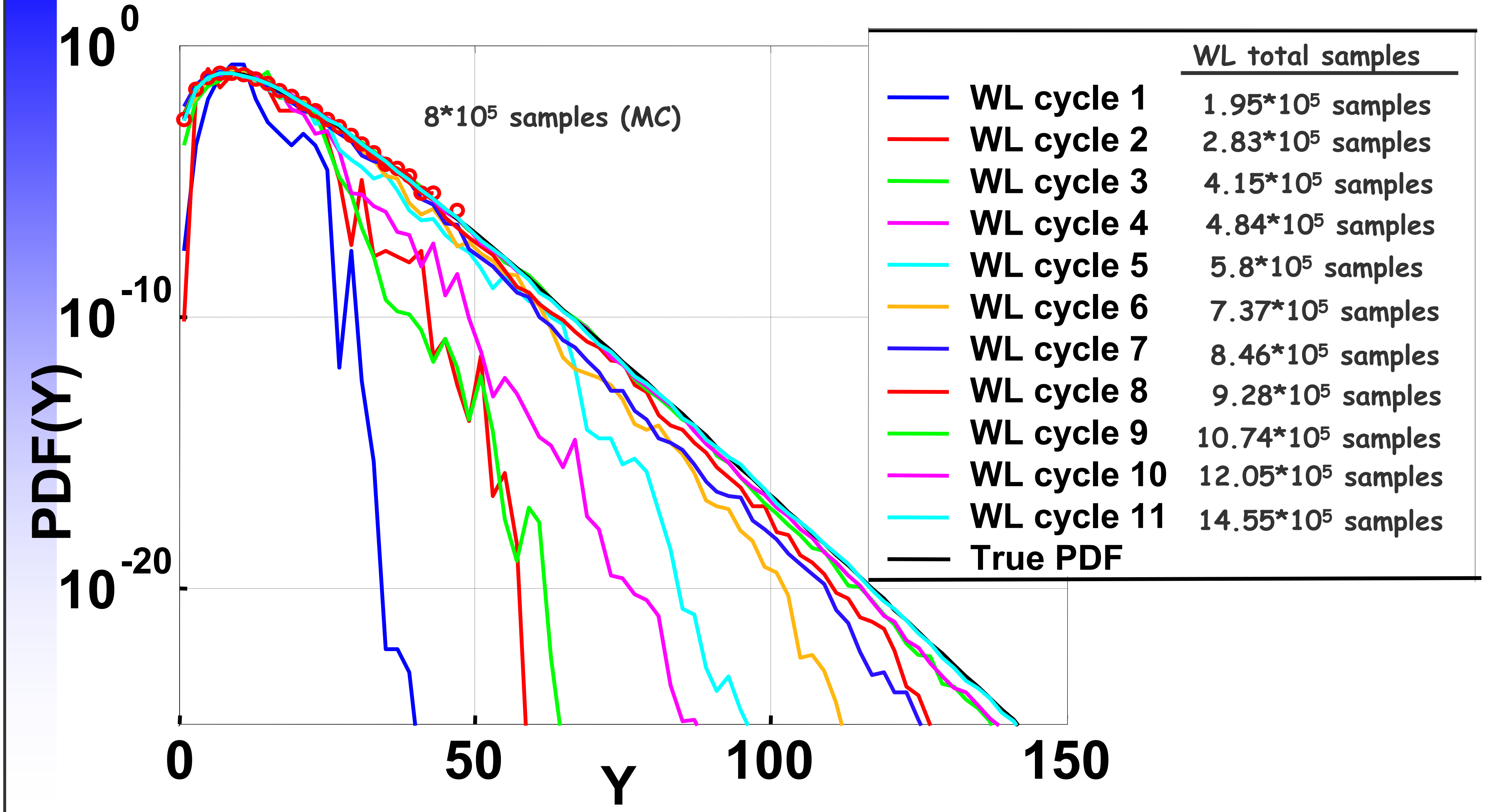
Wang-Landau updates at every time sample and provides biased estimates of PMF of Y. Like MMC, it works in cycles, but of variable length.

It uses a starting cycle precision parameter $f^0 = e \simeq 2718$

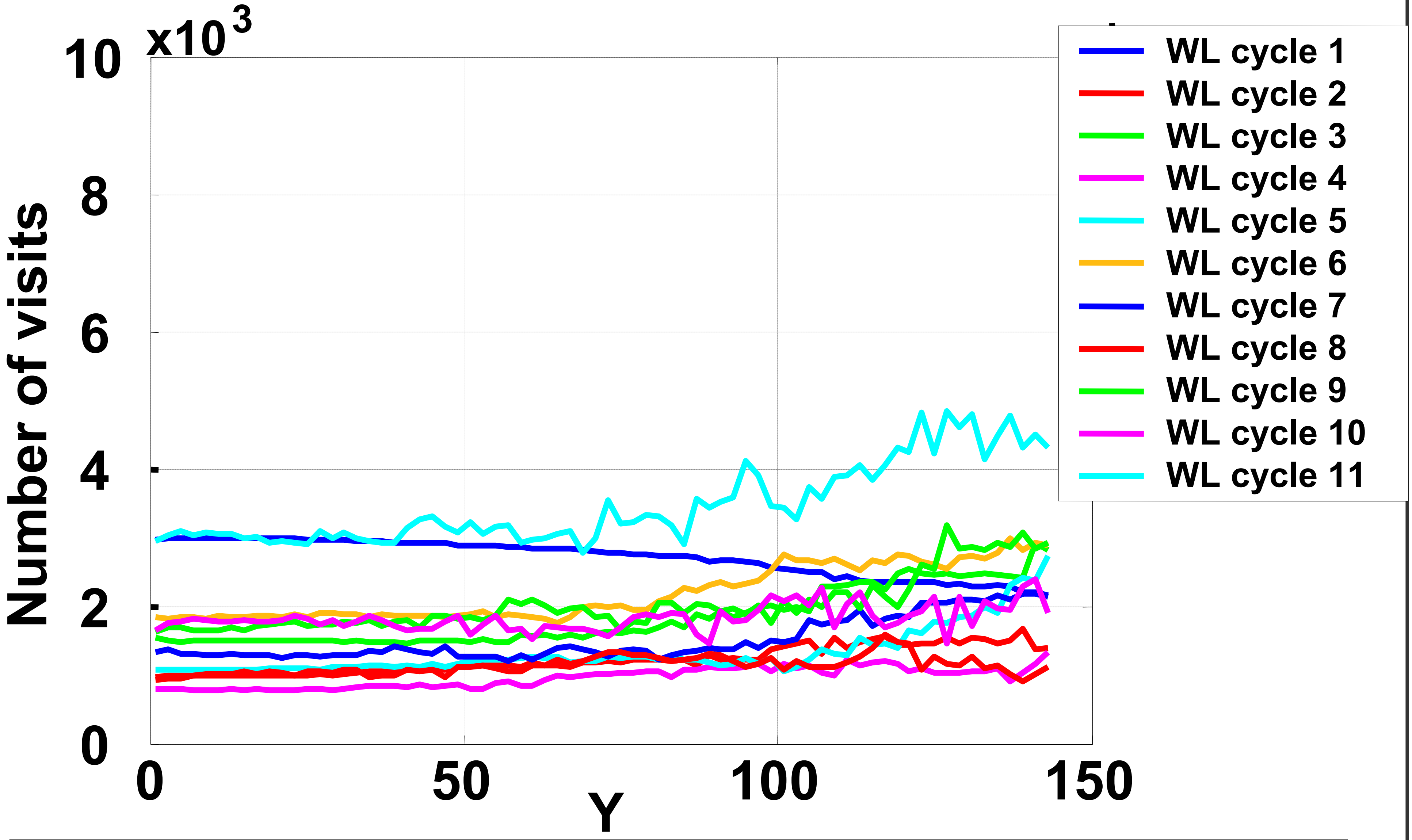
- **WL Algorithm**

1. At beginning of cycle *m*, reset the visits count and update the cycle precision parameter: $$f_m = \sqrt{f_{m-1}}$$
2. At time *n* of cycle *m*, draw a sample from $$f_X^{(n)}(x) = \frac{f_X(x)}{c_{n-1}\Theta_{n-1}(g(x))}$$
3. immediately update estimate of PMF of Y as
$$\Theta_n(y_i) = \begin{cases} f_m \cdot \Theta_{n-1}(y_i) & \text{if } g(X_n) \approx y_i \\ \Theta_{n-1}(y_i) & \text{else} \end{cases}$$
4. Update the visits histogram: increment by 1 count in visited bin.
5. If visits count is flat within desired tolerance (20%) go to next cycle *m+1*. else increment time and goto 2.

Example 1: WL estimate



Example 1: WL Visits Histogram



Outline

- Historical Background
- Motivation
- Monte Carlo (MC)
- Importance Sampling (IS)
- Flat Histogram (FH) Methods
    - Multicanonical Monte Carlo (MMC)
    - Fast MMC
    - Wang Landau (WL)
- Generatig warped Random Variables
    - Rejection Method
    - Markov Chain Monte Carlo (MCMC)
- MMC with MCMC
- Conclusions

Von Neumann Rejection Method

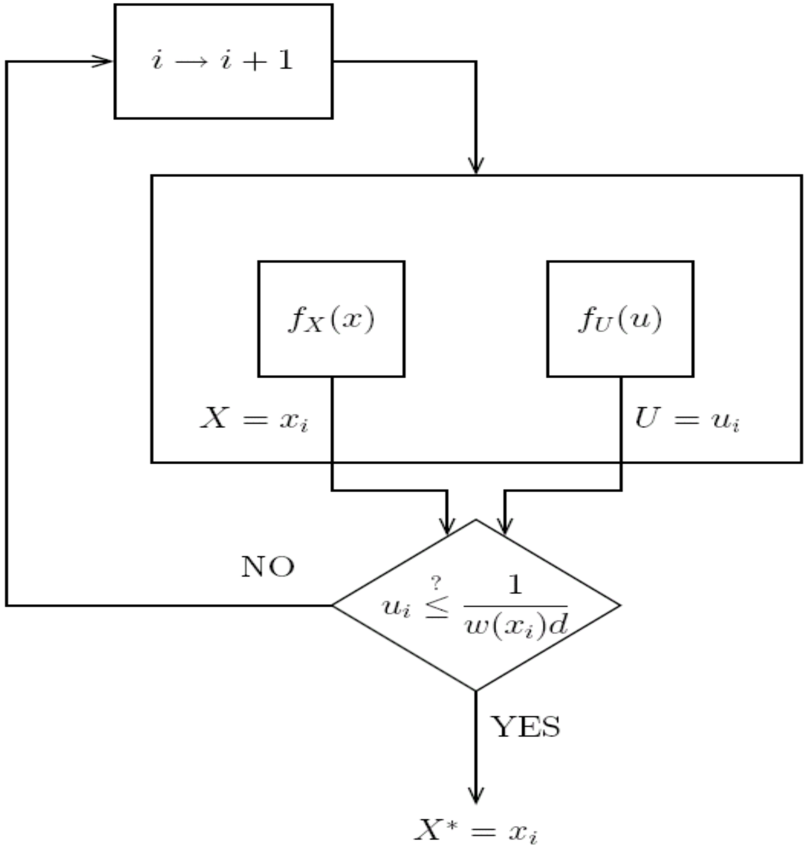
The idea: find a suitable event M, such that $f_X^*(x) = f_X(x \mid M)$

The condition: there should be a $d<\infty$ such that $\dfrac{f_X^*(x)}{d f_X(x)} \leq 1 \quad \forall x \in \Gamma$

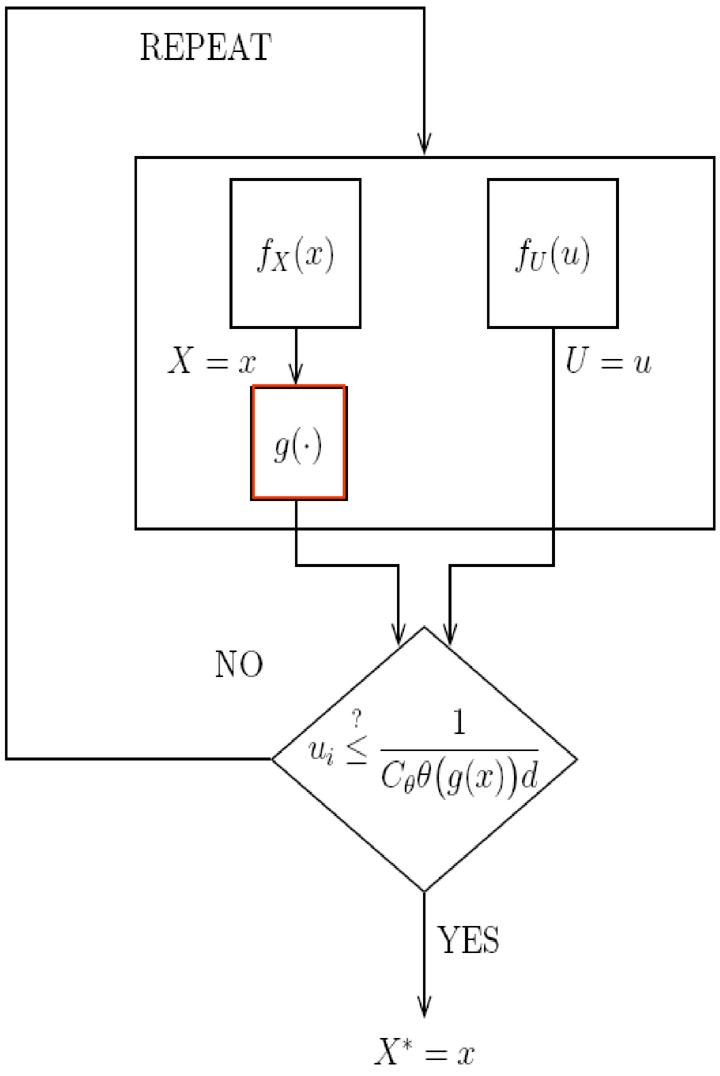
The sample generation algorithm:

```
1.  Generate $u \sim$ Uniform(0,1)
2.  Generate $x \sim f_X(x)$
3.  If (M)
        keep the sample.
    else
        reject it.
```

$$M \triangleq \left\{ u < \frac{f_X^*(x)}{d f_X(x)} \right\}$$



MMC with Rejection Method



1) To sample from $f_X^*(x) = \frac{f_X(x)}{w(x)}$, with weight $w(x) = c_\theta \Theta(y_i)$ *if* $x \in D_i$ it is not necessary to know domains $\{D_i\}, i = 1,..,M$ in state space: just need to verify, for each proposal *g(x)*, to which domain D_i it belongs and use the appropariate weight $c_\theta \Theta(y_i)$ for it.

2) Method gives IID samples.

3) The condition $\frac{f_X^*(x)}{d f_X(x)} \leq 1 \Rightarrow d \geq \frac{1}{w(x)}$, leads to choice

$$d = \max_{1 \leq i \leq M} \left(\frac{1}{c_\theta \Theta(y_i)} \right)$$

However

$$d = \frac{1}{P(M)} = \text{E[trials between successes]}$$

Hence when some of the $\Theta(y_i)$ to be estimted are very small, Rejecton method is too inefficient!

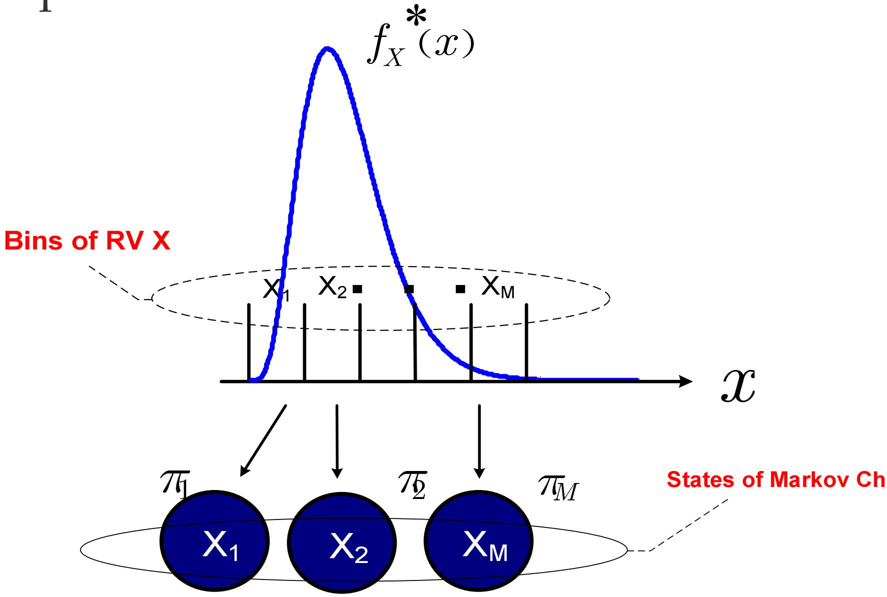
Outline

- Historical Background
- Motivation
- Monte Carlo (MC)
- Importance Sampling (IS)
- Flat Histogram (FH) Methods
  - Multicanonical Monte Carlo (MMC)
  - Fast MMC
  - Wang Landau (WL)
- Generatig warped Random Variables
  - Rejection Method
  - Markov Chain Monte Carlo (MCMC)
- MMC with MCMC
- Conclusions

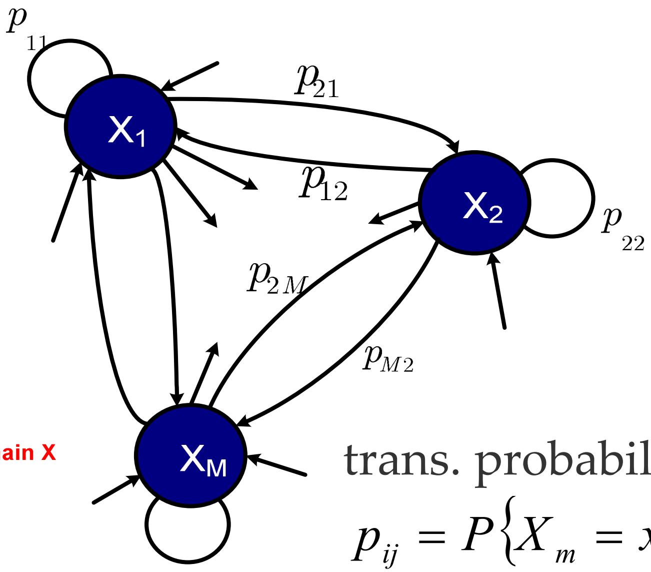
Markov Chain Monte Carlo (MCMC)

- Rigorous treatment requires advanced probability theory to deal with continuous-state Markov Chains. Here for simplicity I assume instead the state space is discretized.

- The idea is to have our samples $\{X_m, m \geq 1\}$ form an ergodic Discrete-Time Markov Chain (DTMC), whose (unique) steady-state distribution $\underline{\pi}$ coincides with the desired warped distribution:



$$\boxed{\pi_i = f_X^*(x_i)\Delta x}$$



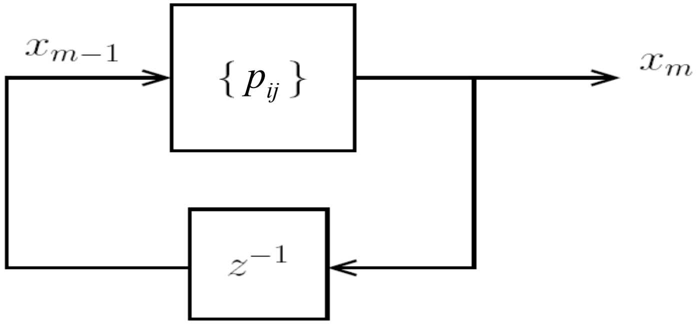
trans. probability j → i

$$p_{ij} = P\{X_m = x_i \mid X_{m-1} = x_j\}$$

trans. probability matrix

$$\underline{\underline{P}} = \{p_{ij}\}$$

Markov Chain Monte Carlo (MCMC)



I will visualize the DTMC generation mechanism as a "stochastic machine"

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

$$\underline{\underline{P}}\,\underline{\pi} = \underline{\pi}$$ global balance (GB)

In our problem the unknown is $\underline{\underline{P}}$. There are ∞ matrices that satisfy GB. We just need one!

To find a simple one, we impose that the DTMC be time-reversible, which is equivalent to imposing that $\underline{\underline{P}}$ satisfies:

$$\forall\, \textit{pair of states (i,j)} \qquad \underbrace{\pi_i p_{ji}}_{P\{X_m=x_j,X_{m-1}=x_i\}} = \underbrace{\pi_j p_{ij}}_{P\{X_m=x_i,X_{m-1}=x_j\}}$$ detailed balance (DB)

i.e., at equilibrium, the probability of being at x_i at time m-1 and moving to x_j at m must equal the probability of the reverse transition. Thus we get all unknowns $\{p_{ij}\}$.

Metropolis-Hastings (MH)

Metropolis-Hastings [Metropolis et al, *J Chem Phys 1953*; Hastings, *Biometrika 1970*] is a way to realize a time-reversible Markov chain with a pre-specified steady-state distribution:

1. Specify the desired $\underline{\pi}$
2. Pick an arbitrary $\underline{\underline{Q}} = \{q_{ij}\}$ (the "Candidate". I prefer to call it the "Explorer")
3. For any pair of states x_i and x_j :

either : a) $\overbrace{\pi_i q_{ji}}^{i \to j} > \overbrace{\pi_j q_{ij}}^{j \to i}$

or : b) $\overbrace{\pi_i q_{ji}}^{i \to j} < \overbrace{\pi_j q_{ij}}^{j \to i}$

In case (a) we accept the transitions $i \to j$ with a probability α_{ji} such that:

$$\pi_i q_{ji} \alpha_{ji} = \pi_j q_{ij}$$

In case (b) swap i and j.

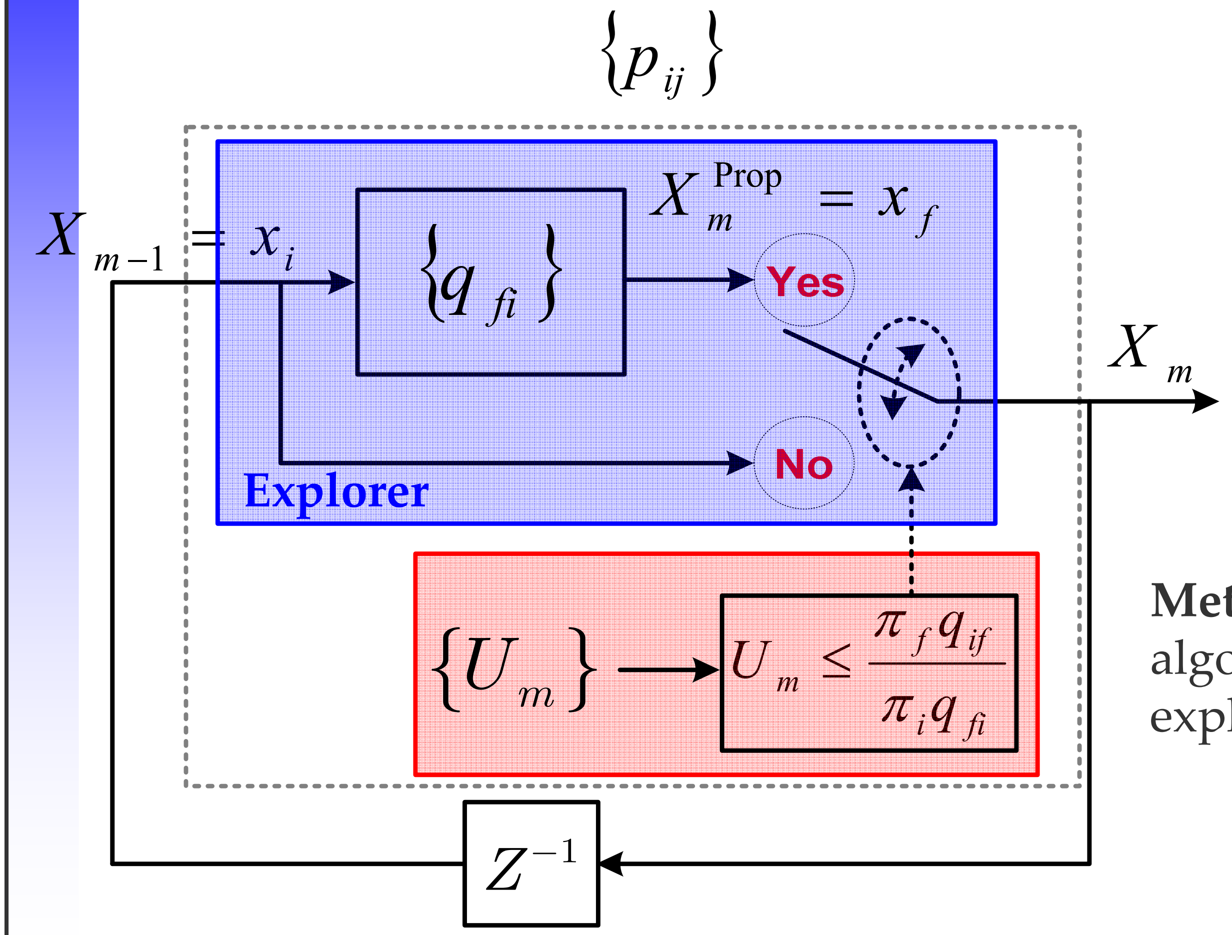
We force detailed balance !

Metropolis-Hastings

In summary:

Each time **the Explorer proposes** a move *i*→*f*, it is **accepted with probability**:

$$\alpha_{fi} = \min\left(\frac{\pi_f q_{if}}{\pi_i q_{fi}}, 1\right)$$



Doing so, we synthesize **by construction** a time- reversible ergodic Markov chain with steady state distribution $\underline{\pi}$

Metropolis algorithm is an MH algorithm in which $q_{if} = q_{fi}$ (symmetric explorer):

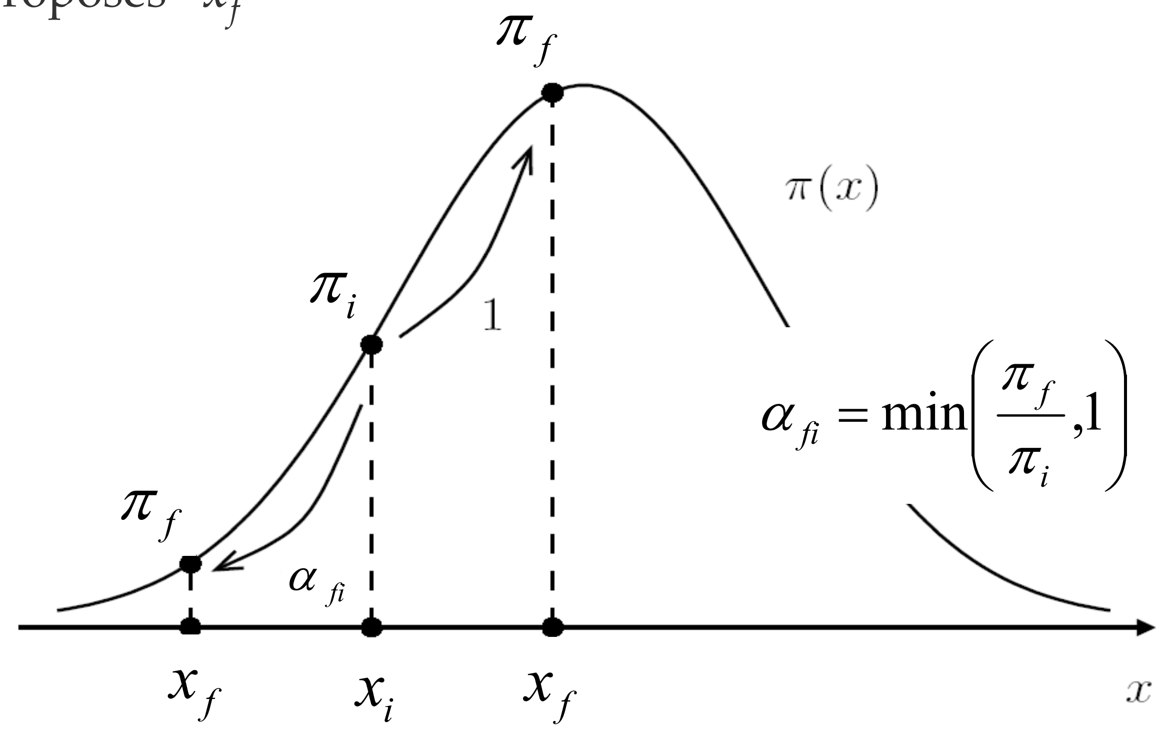
$$\alpha_{fi} = \min\left(\frac{\pi_f}{\pi_i}, 1\right)$$

Metropolis: intuition

Suppose $X_{m-1} = x_i$ and explorer proposes x_f

If $\pi_f > \pi_i$, proposal always accepted ($\alpha_{fi} = 1$)

So upward moves encouraged : visit more frequently modal range of π(x)



$$\alpha_{fi} = \min\left(\frac{\pi_f}{\pi_i}, 1\right)$$

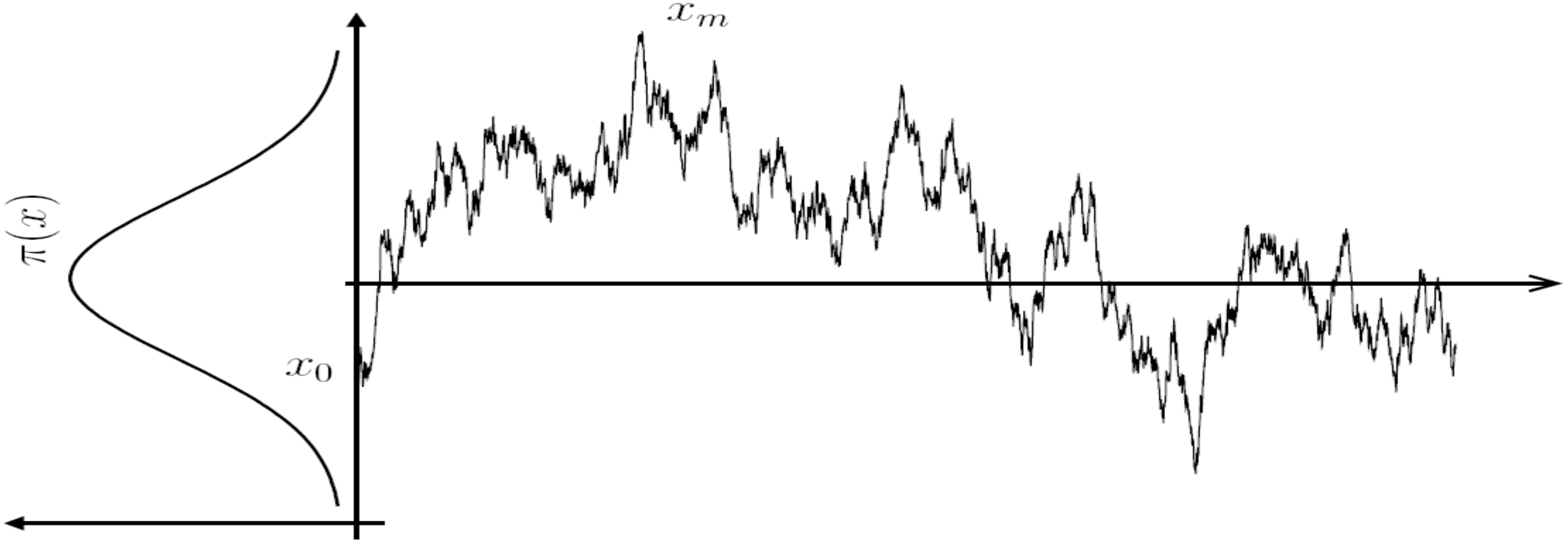
If $\pi_f << \pi_i$ proposal rejected most of the times ($\alpha_{fi} << 1$)

So big downward moves discouraged.

That's how the correct π(x) is actually sampled!

Ergodicity

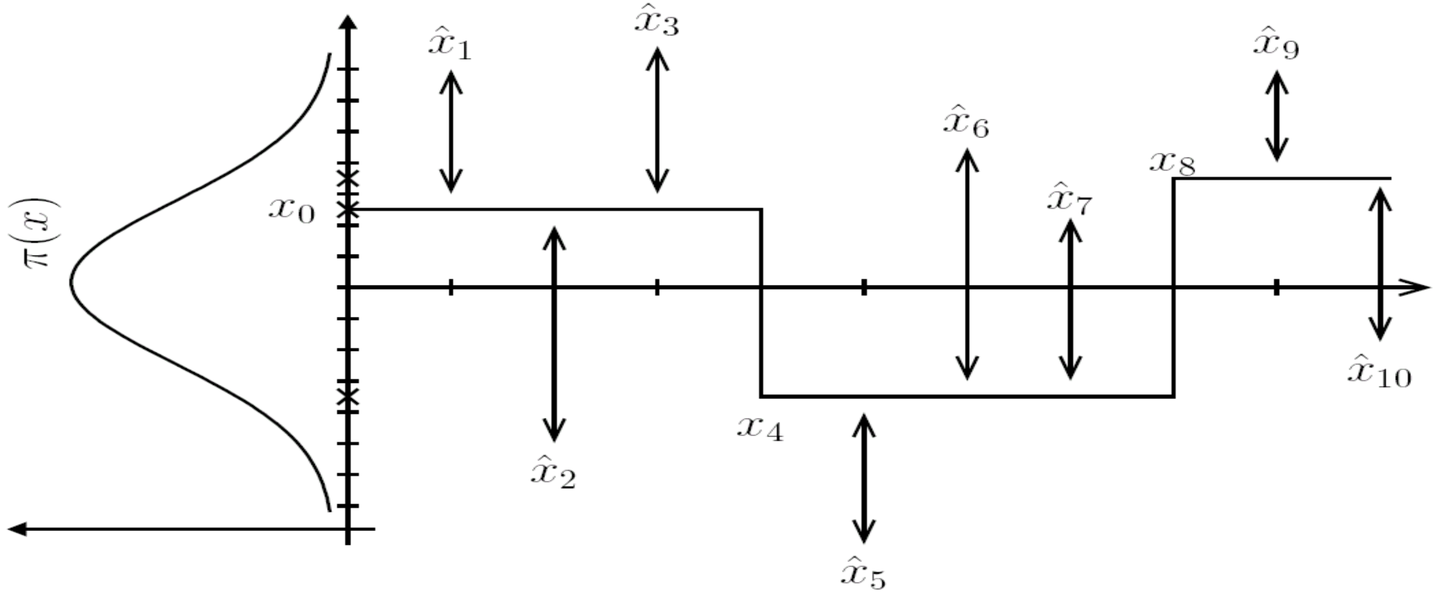
Q: When is a simulation long enough for sampling from π(x) ?



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

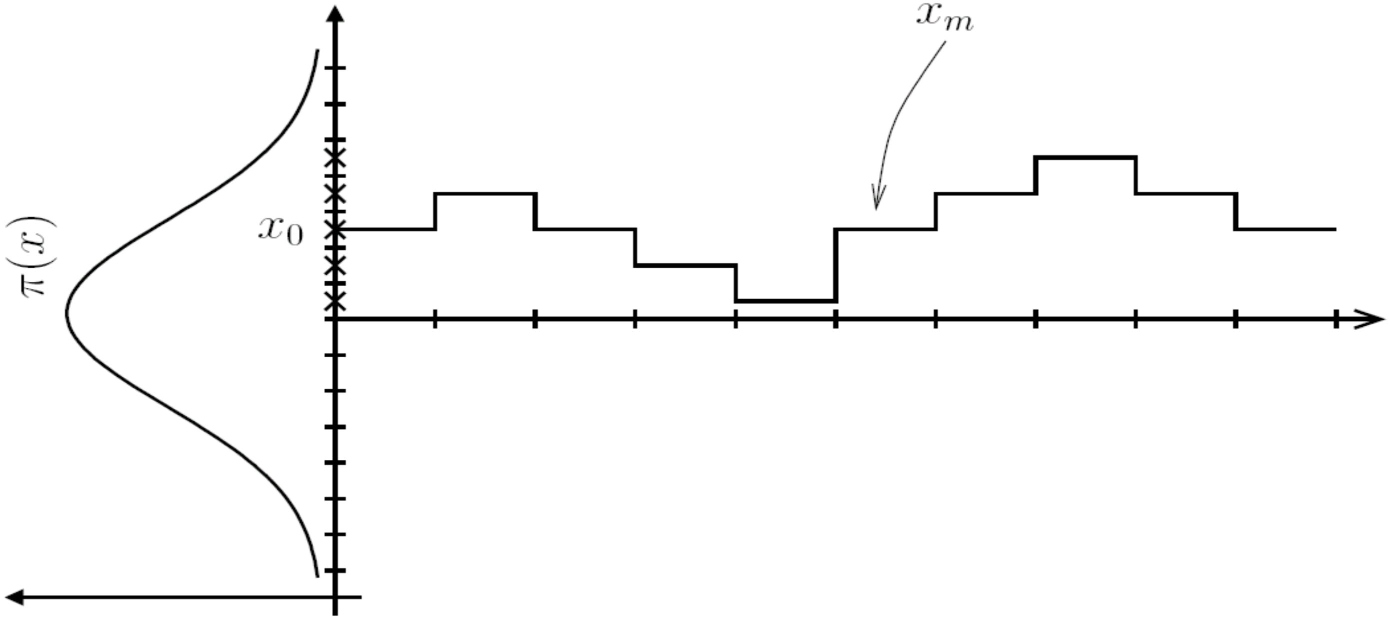
The Explorer

A symmetric explorer $q_{if} = q_{fi}$ is implemented by proposing $x_f = x_i + U$ where U is any zero-mean random vector with even PDF in all dimensions. What impact has its variance $\sigma^2{}_U$ on DTMC movement and thus on ergodicity?



If $\sigma^2{}_U$ too large, most proposals are rejected (samples with a hat ^ in figure): the chain moves too slowly. So a large rejection rate is not an indicator of efficient sampling.

The Explorer



If σ^2_U too small, most proposals are accepted: still, the chain moves too slowly
Hence also a too large acceptance rate is not a good indicator of "ergodic sampling"....

Aside: a particular non-symmetric explorer exists for which all moves are accepted :

$$q_{fi} = \pi_f \qquad \text{(In fact} \quad \alpha_{fi} = \min\left(\frac{\pi_f q_{if}}{\pi_i q_{fi}}, 1\right) = 1 \text{)}$$

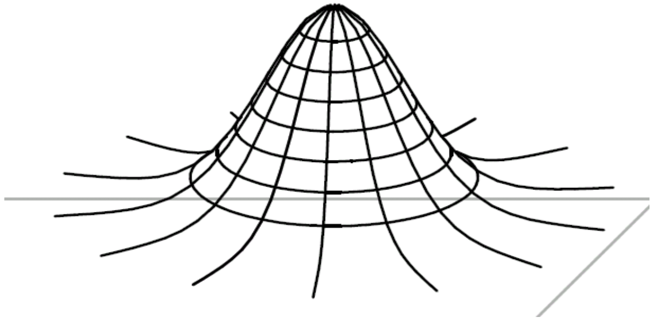
Explorers whose transition probability depends only on the final state generate DTMCs which are called independence chains.

The Explorer

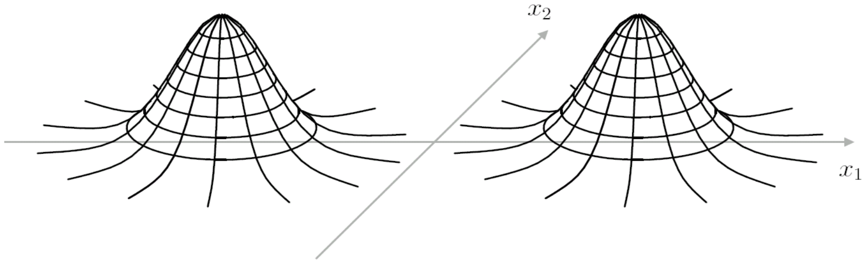
Q: What is a reasonable $\sigma^2{}_U$? What is a good acceptance rate?

Intuitively, the variance of U (in each dimension) should be some fraction of the variance of π(x) (in each dimension) if one hopes to correctly explore π(x) .

For unimodal π(x), this leads to reasonable acceptance rates (50% and higher)



For multimodal π(x), with almost zero probabilty among modes, this leads to very low acceptance rates!!
For such pathological cases, there is the following nice trick that one can use.......



Small-world MCMC [Y. Guan *et al.*, Stat. Comput. 2006]

The idea: at each step, propose a completely random state x_f with probability p, otherwise use a symmetric explorer to form a proposal of the form $x_f = x_i + U$, with U having "low-variance" (a fraction of that of a "mode")

One is thus able to accurately sample each mode with the "local explorer", being able to jump among modes using the "global explorer"

This amounts to using the following mixed explorer:

$$q_{fi} = (1-p)\tilde{q}_{fi} + p\frac{1}{|\Gamma|}$$

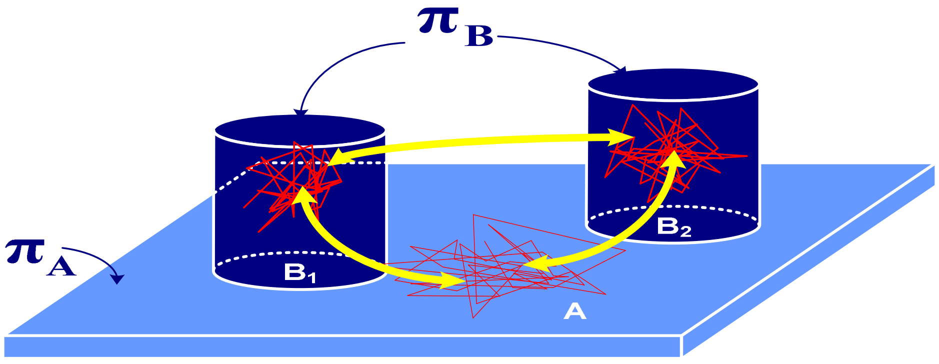
where $\tilde{q}_{fi}$ is the "local" low-variance symmetric explorer, and $|\Gamma|$ is the cardinality of the input space.

Small-world MCMC

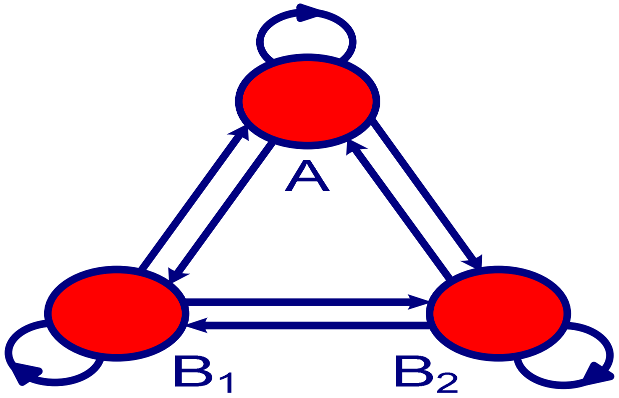
If one knows the total volume of the “important regions” (modes B_1, B_2, ...) relatve to $|\Gamma|$, and call it r, assumes that local explorer unable to jump among regions, and that once in $B=B_1 \cup B_2 \cup$ ..., the simulation is run for N time steps, then optimal value of p is

$$p \cong \frac{1}{\sqrt{Nr}}$$

Result obtained by stdying a markov chian over important and unimportant regions.

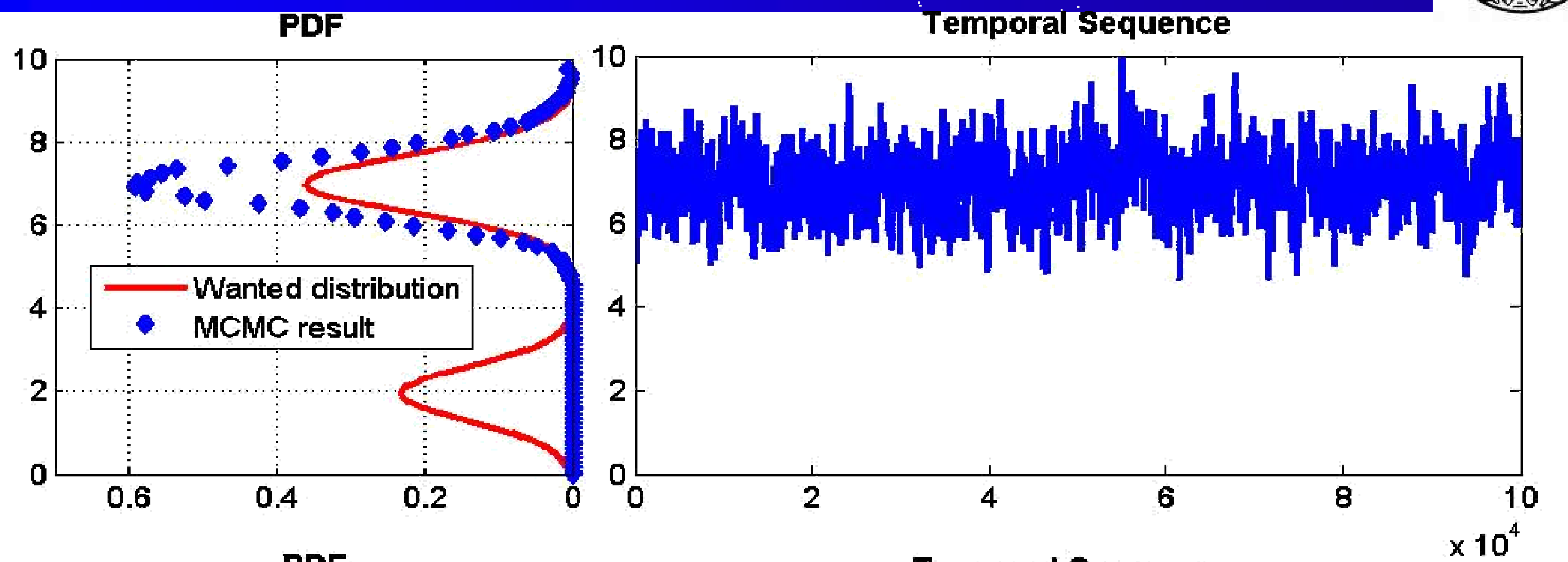


$$\Gamma = A \cup B_1 \cup B_2$$

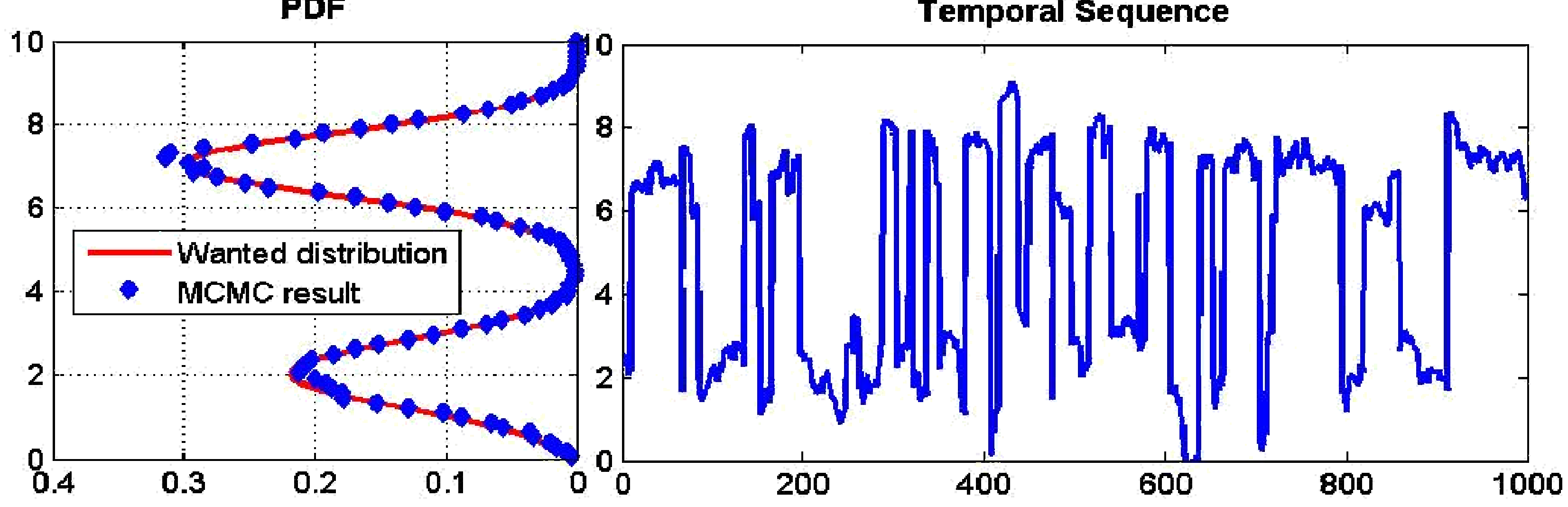


Small-world MCMC Examples

Without Small World



With Small World



Is Ergodicity necessary in Flat Histogram Algorithms ?

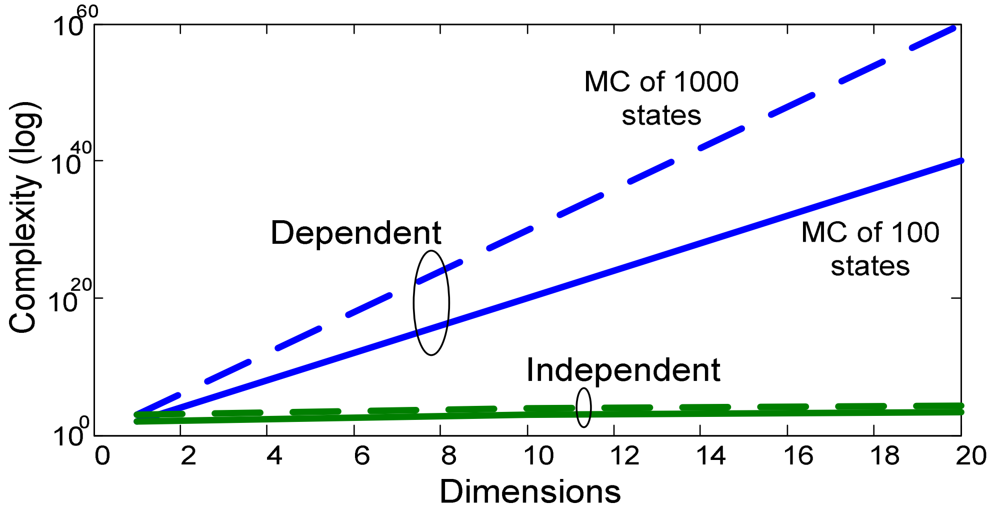
Since convergence is approached in cycles, and the task of each cycle is to get closer and closer to the flat output histogram, the issue of ergodicity is probably less important in FH PDF estimation than in other estimation applications that rely on a single MCMC run.

The convergence of the Wang-Landau FH is clear evidence of the above.

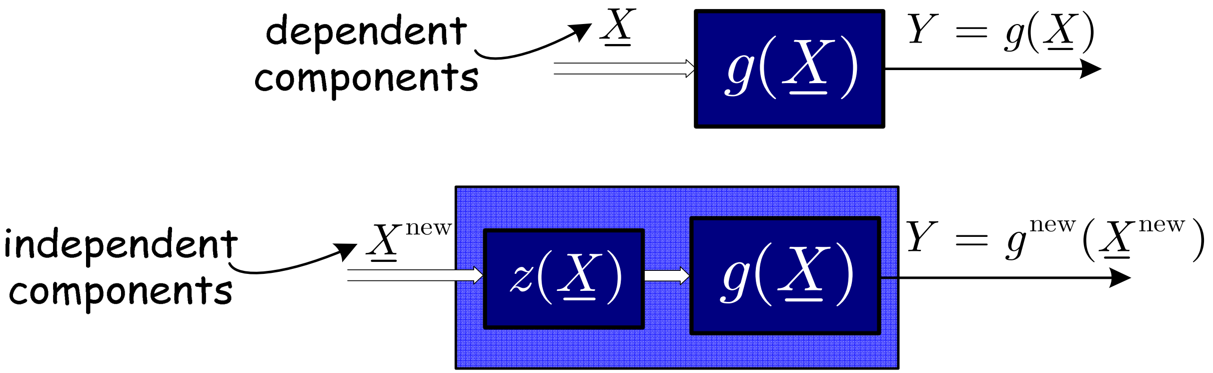
High Dimensionality: computation time

- Computation time grows exponentially with the dimension in case of **dependent** RVs, while only linearly for **independent** RVs.

Total number of bins of the X space



- We should formulate problems such that the input RV's are "almost" independent:



High Dimensionality: one-variable-at-a-time

When $\underline{X}=[X_1,..,X_d]$ has large dimension d, instead of jointly Metropolis-updating all its variables, it is more efficient to cyclically use several Metropolis-updates, one for each component (or group of components) of $\underline{X}$.

Ex: assume the components are IID, with common distribution $G(x)$. Hence the distribution we wish to sample from is $\pi(\underline{x})=G(x_1)*...*G(x_d)$. The "joint" explorer proposes a move to $\underline{x}_f=\underline{x}_i+\underline{U}$, where for instance $\underline{U}$ is uniform in a d-dimensional hypercube centerd at the origin. Then it will generally happen that some components of $\underline{U}$ are small, and some are large.

Then in joint Metropolis, the acceptance probability of move i→f will be

$$\alpha_{fi} = \min\left(1, \frac{\pi(\underline{x}_f)}{\pi(\underline{x}_i)}\right) = \min\left(1, \frac{G(x_{1,f})*...*G(x_{d,f})}{G(x_{1,i})*...*G(x_{d,i})}\right)$$

and if $\underline{x}_i$ is at a large probability, then the larger the dimension d, the smaller the product $\pi(\underline{x}_f) = G(x_{1,f})*...*G(x_{d,f})$ will be, and thus the larger the rejection rate.

Doing instead Metropolis on each component proposal $x_{k,f}=x_{k,i}+U_k$, (k=1,..,d) gives acceptance

$$\left(\alpha_{fi}\right)_k = \min\left(1, \frac{G(x_{k,f})}{G(x_{k,i})}\right)$$

so that movement in some components is always granted !

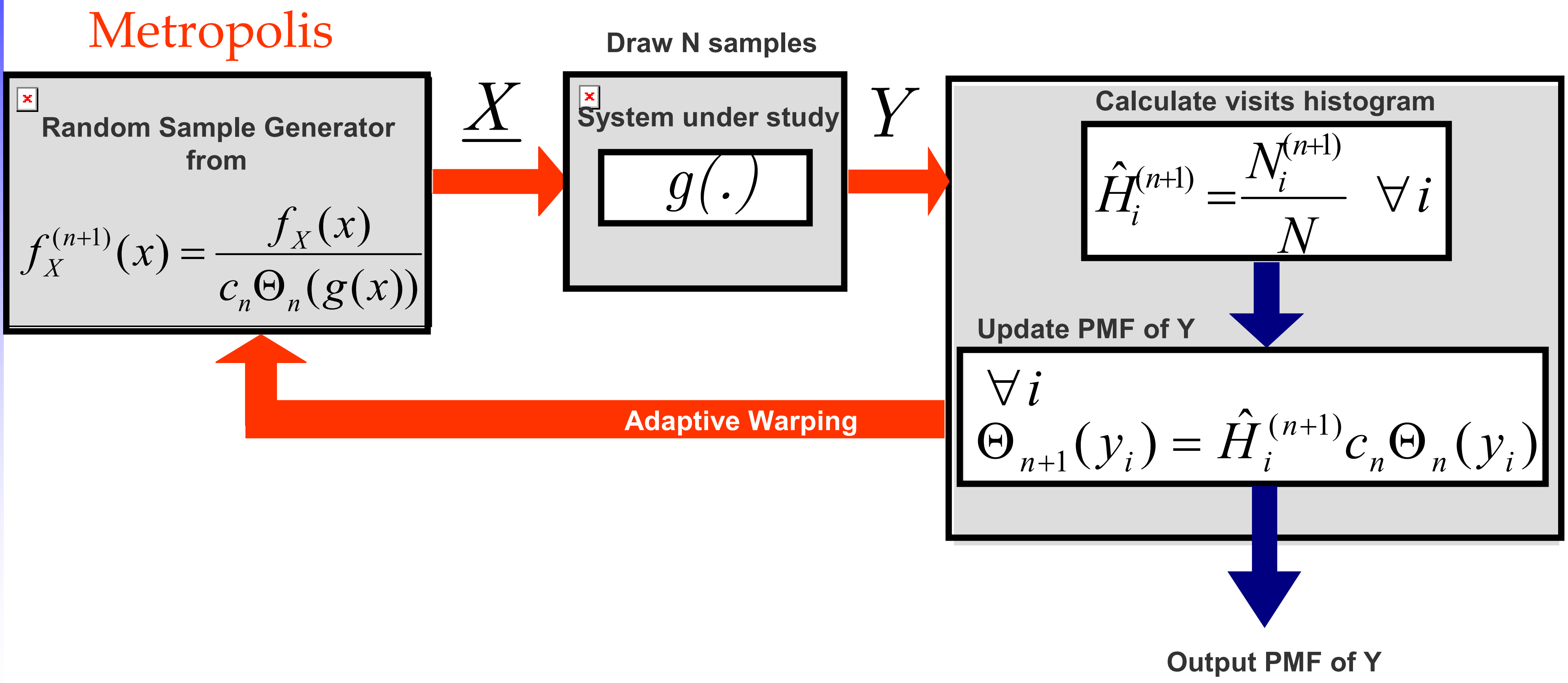
Outline

- Historical Background
- Motivation
- Monte Carlo (MC)
- Importance Sampling (IS)
- Flat Histogram (FH) Methods
  - Multicanonical Monte Carlo (MMC)
  - Fast MMC
  - Wang Landau (WL)
- Generatig warped Random Variables
  - Rejection Method
  - Markov Chain Monte Carlo (MCMC)
- **MMC with MCMC**
- Conclusions

MMC with MCMC

Let's see how the one-variable-at-a-time can be implemented in MMC.

Recall the block-diagram of MMC:



Metropolis

Random Sample Generator from

$$f_X^{(n+1)}(x) = \frac{f_X(x)}{c_n\Theta_n(g(x))}$$

$\underline{X}$

Draw N samples

System under study

$g(.)$

Y

Calculate visits histogram

$$\hat{H}_i^{(n+1)} = \frac{N_i^{(n+1)}}{N} \quad \forall i$$

Update PMF of Y

$\forall i$

$$\Theta_{n+1}(y_i) = \hat{H}_i^{(n+1)} c_n \Theta_n(y_i)$$

Adaptive Warping

Output PMF of Y

MMC using MCMC

The simple Metropolis mechanism described at pages 32-33 in inefficient in high-dimensional input spaces:

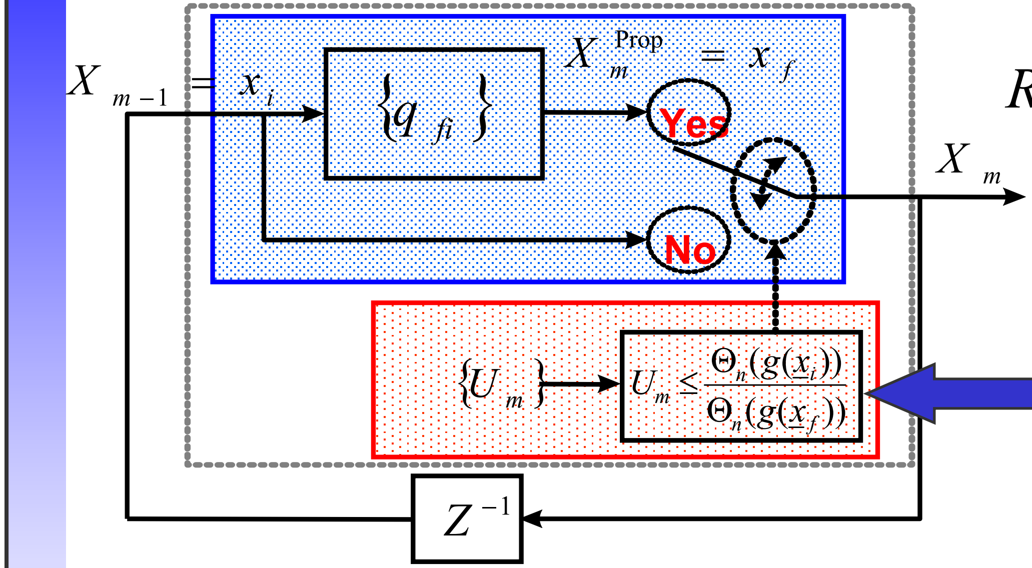
$$R=\frac{f_X^{(n+1)}(\underline{x}_f)}{f_X^{(n+1)}(\underline{x}_i)}=\frac{f_X(\underline{x}_f)}{c_n\Theta_n(g(\underline{x}_f))}\bullet\frac{c_n\Theta_n(g(\underline{x}_i))}{f_X(\underline{x}_i)}=\frac{\Theta_n(g(\underline{x}_i))}{\Theta_n(g(\underline{x}_f))}\bullet\frac{f_X(\underline{x}_f)}{f_X(\underline{x}_i)}$$

Acceptance probability $\alpha_{fi}=\max(1,R)$ too low when $\underline{x}_i$ is in mode of $f_X(\underline{x})$, unless $f_X(\underline{x})$ uniform, since it cancels in R.

- When input RVs are independent,
  one can use the following trick [Holzlohner *et al*, Opt. Lett. 2003]

MMC in high-dimensions

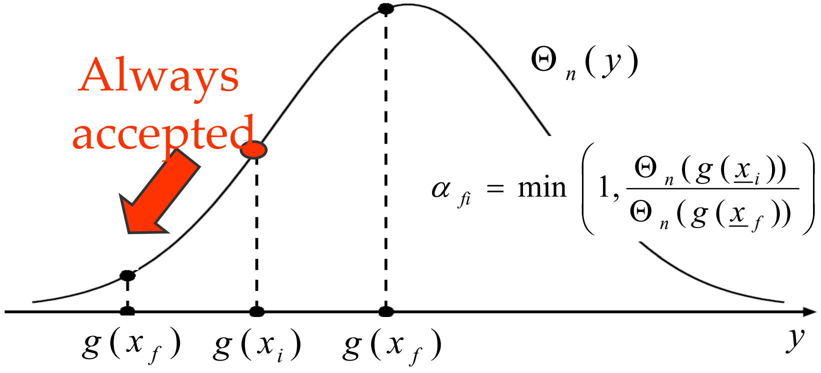
Use an independence chain explorer $q_{fi} = f_X(\underline{x}_f)\Delta x$ that samples directly from the initial distribution, so that the Hastings ratio becomes:


$X_{m-1} = x_i$
$\{q_{fi}\}$
$X_m^{\text{Prop}} = x_f$
Yes
No
X_m
$\{U_m\}$
$U_m \le \frac{\Theta_n(g(\underline{x}_i))}{\Theta_n(g(\underline{x}_f))}$
Z^{-1}

$$R = \frac{f_X^{(n+1)}(\underline{x}_f)q_{if}}{f_X^{(n+1)}(\underline{x}_i)q_{fi}} = \frac{\Theta_n(g(\underline{x}_i))}{\Theta_n(g(\underline{x}_f))} \bullet \frac{f_X(\underline{x}_f)q_{if}}{f_X(\underline{x}_i)q_{fi}} = \frac{\Theta_n(g(\underline{x}_i))}{\Theta_n(g(\underline{x}_f))}$$

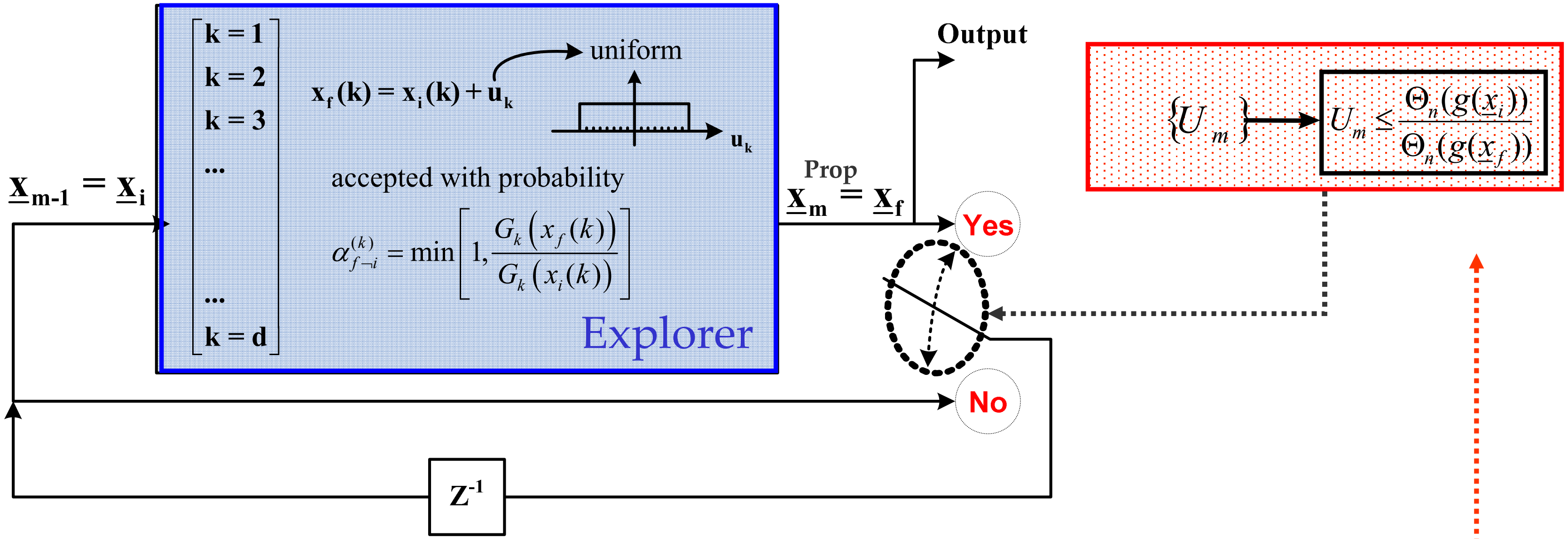
Reject/accept test now depends solely on the control variable $Y=g(X)$

Random walk on output RV Y is pushed towards tails of $\Theta_n(y)$



Tail exploration is however still too slow:
since proposals $\underline{x}_f$ are independent and drawn from distribution $f_X(\underline{x})$, then most sample proposals $g(\underline{x}_f)$ will fall in mode of $\Theta_n(y)$ and will thus be rejected.

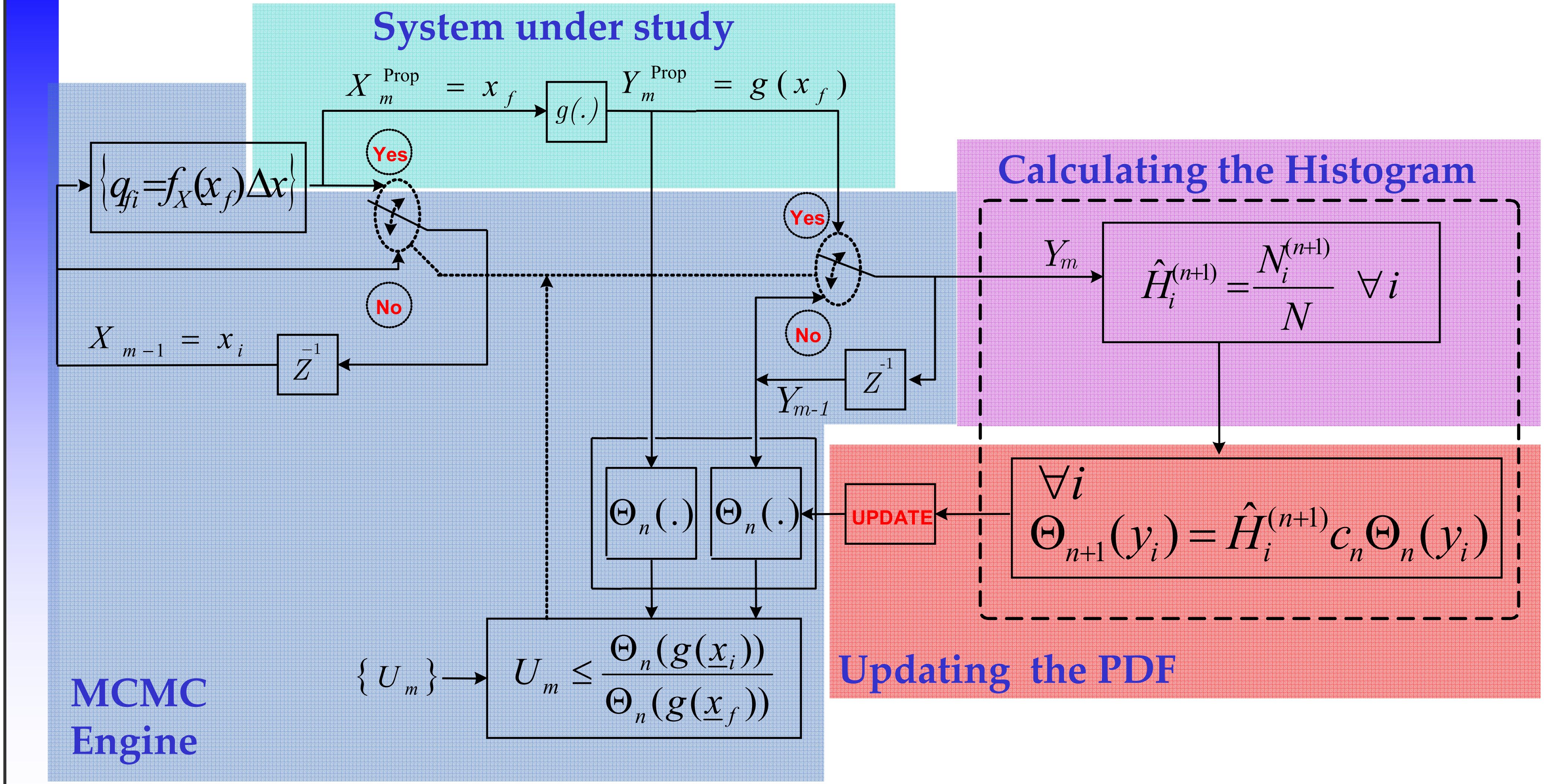
Hence the trick is to implement the independece chain explorer itself as an MCMC machine in which one-variable-at-a-time is performed !


Output
uniform
$x_f(k) = x_i(k) + u_k$
u_k
k = 1
k = 2
k = 3
...
...
k = d
accepted with probability
$\alpha_{f \to i}^{(k)} = \min\left[1, \frac{G_k(x_f(k))}{G_k(x_i(k))}\right]$
Explorer
$\underline{x}_{m-1} = \underline{x}_i$
Prop
$\underline{x}_m = \underline{x}_f$
Yes
No
Z^{-1}
$\{U_m\}$
$U_m \le \frac{\Theta_n(g(\underline{x}_i))}{\Theta_n(g(\underline{x}_f))}$

The d component-wise reject/accept Metropolis tests will thus produce correlated proposals $\underline{x}_f$ with the desired distribution $f_X(\underline{x})$, so that the outer reject/accept mechanism has now a reasonable acceptance probability

$$\alpha_{fi} = \min\left(1, \frac{\Theta_n(g(\underline{x}_i))}{\Theta_n(g(\underline{x}_f))}\right)$$

MMC machine: Global View


System under study
$X_m^{\text{Prop}} = x_f$
$g(.)$
$Y_m^{\text{Prop}} = g(x_f)$
$\{q_{fi} = f_X(x_f)\Delta x\}$
Yes
No
$X_{m-1} = x_i$
Z^{-1}
Yes
No
Y_m
Z^{-1}
Y_{m-1}
Calculating the Histogram
$\hat{H}_i^{(n+1)} = \frac{N_i^{(n+1)}}{N} \quad \forall i$
$\Theta_n(.)$
$\Theta_n(.)$
UPDATE
$\forall i$
$\Theta_{n+1}(y_i) = \hat{H}_i^{(n+1)} c_n \Theta_n(y_i)$
$\{U_m\}$
$U_m \le \frac{\Theta_n(g(\underline{x}_i))}{\Theta_n(g(\underline{x}_f))}$
MCMC Engine
Updating the PDF

Conclusions

- MMC is an IS-based adaptive Flat Histogram algorithm.

  Doesn't need almost any knowledge of specific physical problem!

  This is major difference with IS.

  Metropolis parameters (number of runs per cycle, explorer variance, small-world jump probability) are easily set. Number and density of bins can be dynamically adjusted cycle after cycle.

- WL doesn't seem to offer any advantages over MMC, but better FH methods may exist.

- There are particularly stiff problems (e.g. coded telecomm. systems) in which MMC is less efficient. This is a good topic for further research.

- FH are a tiny part of a vast topic from all sciences (physics, chemistry engineering, economics) dealing with statistical inference based on MCMC. Such a topic is becoming increasingly important in research and applications.

Thank You for your kind attention!

Questions?

References

- *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)*

- *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.*

- *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.*

- *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.*

- *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.*

- *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.*

- *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.*

- *B.A. Berg, “Introduction to Multicanonical Monte Carlo Simulations”, Fields Inst.Commun., vol. 26, n. 1, 2000.*

- *J. Gubernatis, and N. Hatano, “The multicanonical monte carlo method,” Computer Simulations, vol. 2, no. 2, pp. 95-102, Mar./Apr. 2000..*

References

- *R. Holzlohner, C. R. Menyuk, "Use of multicanonical Monte Carlo Simulations to Obtain Accurate Bit Error Rates in Optical Communication Systems", Opt. Lett., vol 28, no. 20, pp. 1894-1896, Oct. 2003.*

- *D. Yevick, "The Accuracy of Multicanonical System Models", Photon. Technol. Lett., vol. 15, pp. 224-226, Feb. 2003.*

- *F. Liang, "A theory on flat histogram monte carlo algorithms," J. Stat. Physics, vol. 122,no. 3, pp. 511-529, Feb. 2006.*

- *Y. F. Atchade, J. S. Liu, "The Wang-Landau algorithm for MC computation in eneral state spaces", Technical report, University of Ottawa, 2004 (available at www.mathstat.uottawa.ca/~yatch436/gwl.pdf.)*

- *N. Metropolis, A. W. Rosenbluth, A. H. Teller, E. Teller, "Equations of state calculations by fast computing machines", J. Chem. Phys, vol. 21, no. 6, pp. 1087-1092, June 1953.*

- *W. K. Hastings, "Monte Carlo sampling methods using Markov Chains and their applications", Biometrika, vol. 57, no. 1, pp. 97-109, Apr. 1970.*

- *C. J. Geyer, "Markov chain monte carlo lecture notes" Course notes originally used Spring Quarter 1998, University of Minnesota, last typeset November 21, 2005.*

- *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.*

- *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..*

For an extended presentation of this material, see also my course notes:

A. Bononi, "*Multicanonical Monte-Carlo and importance sampling: how are they related --*
A short course for Ph.D. students in information engineering", Parma University, Jan. 2007
(available at www.tlc.unipr.it/bononi/ricerca/seminars/MMCcourse.pdf)