“Monte Carlo unifying Frequentist and Bayesian inference”

Niels Jørgen Kjær, CERN, Switzerland

Caveats

- Searches are excluded from this talk at 95% CL.
- The presentation is not exhaustive.
- The presentation is biased towards LEP experience
- Thanks to many people for useful discussions.

Outline of topics

- Introduction
- What is Monte Carlo?
- How do HEP MC work?
- MC weights and reweighting
- Reweighting analyses
- Frequentist resampling
- Systematic errors
- Unifying everything
- Conclusions and outlook

Apologies for all the many topics I have left out!

Introduction

- The Monte Carlo technique has a long history
  - Buffon's Needle
  - Simple attempts on thermodynamics
  - Real use $\sim$ 1944 in development of nuclear weapons
  - Formalized $\sim$ 1970
  - Dominant in HEP from $\sim$ 1980
  - $>$90% of todays HEP analyses are MC based
- The Monte Carlo technique is widely used
  - Particle sciences
  - Finance
  - Military
  - Social sciences
- The Monte Carlo technique is used under many names
  - Markov Chain Monte Carlo (states)
  - Gibbs sampling, Metropolis, Hastings, etc. etc.
  - MC in HEP is the easiest and most explored example

What is a Monte Carlo in HEP?

$$< E >= \left(\int d\vec{\Omega}|\mathcal{M}(\vec{\Omega},\vec{p})|^2 E(\vec{\Omega})\right) \times \left(\int d\vec{\Omega}|\mathcal{M}(\vec{\Omega},\vec{p})|^2\right)^{-1}$$

$\vec{\Omega}$ observable space, $\vec{p}$ parameter space

Weighted or unweighted random sampling with “events”

If events, i, are generated according to $|\mathcal{M}(\vec{\Omega}_i,\vec{p})|^2$

$\Rightarrow < E >= \frac{1}{n}\sum_{i=1}^{n} E(\vec{\Omega}_i)$

Variance $\sim$ minimum if event weights are unity

Factorize into sequence of “chain rule” convolutions

$$P(\vec{\Omega},\vec{p}) = P(\vec{\Omega},\vec{a})\oplus P(\vec{a},\vec{b})\oplus P(\vec{b},\vec{c})\oplus P(\vec{c},\vec{p})$$

Treat possible integrable poles separately

Choose suitable parameter transformations

Optimization is important part of MC generators

Final MC generators should have:

Ability to generate events with unit weight

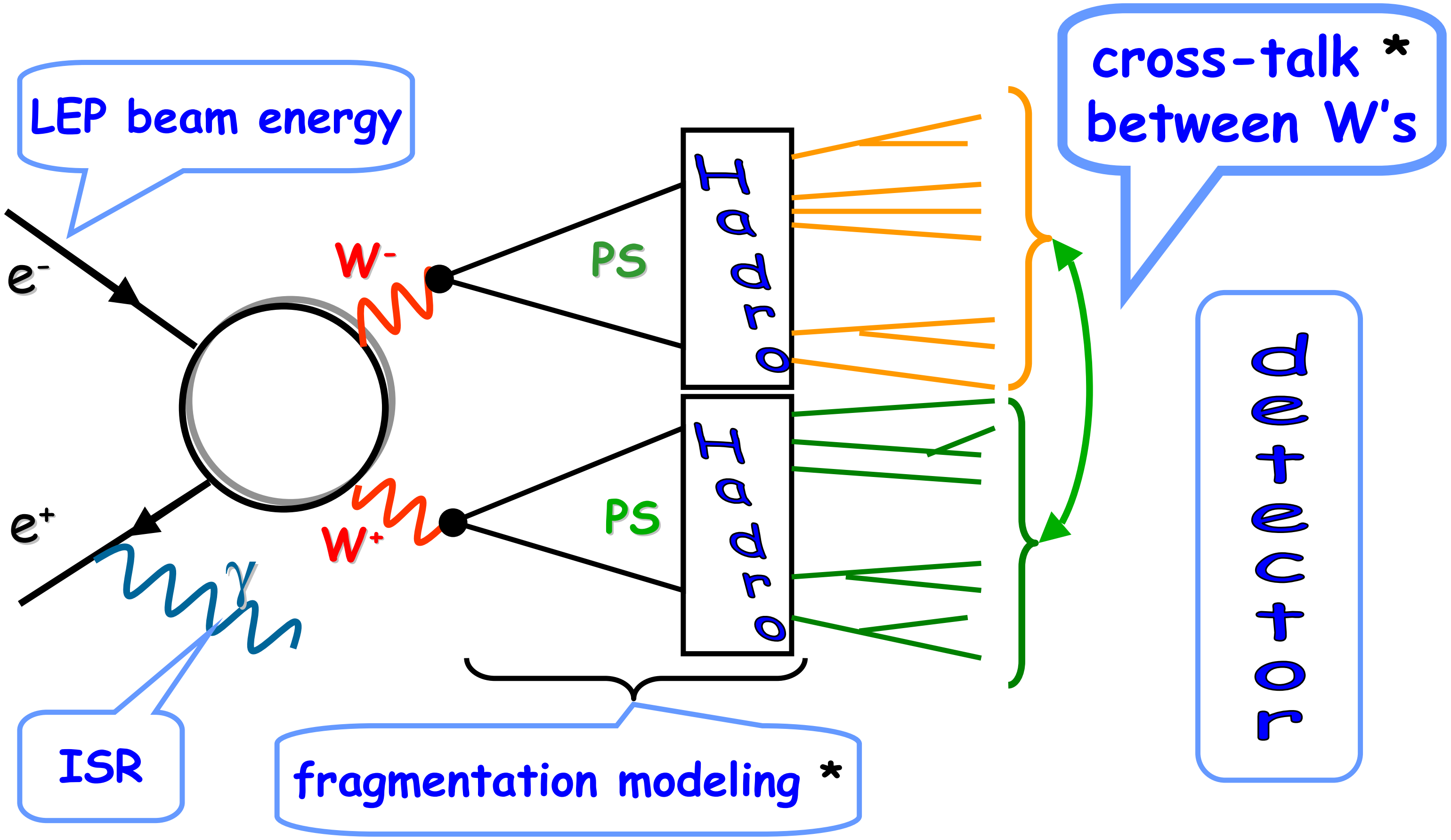
Estimations of numerical uncertainties

HEP events are independent and uncorrelated!

How does the HEP MC work?

- Electroweak part
  - Initial state, radiation and parton density functions
  - Hard Matrix element squared
  - Phase space
- Fragmentation and decays
  - Parton shower or ME approach
  - Decay tables
  - Polarization etc
  - Ends with generated “stable” particles
- Particle interactions with matter
  - Ionization loss, multiple scattering
  - hadronic interactions, delta rays, etc.
- Detector response description
  - Much more involved than the above
  - Many tunable parameters
  - Most CPU demanding

W Mass: Systematic Errors

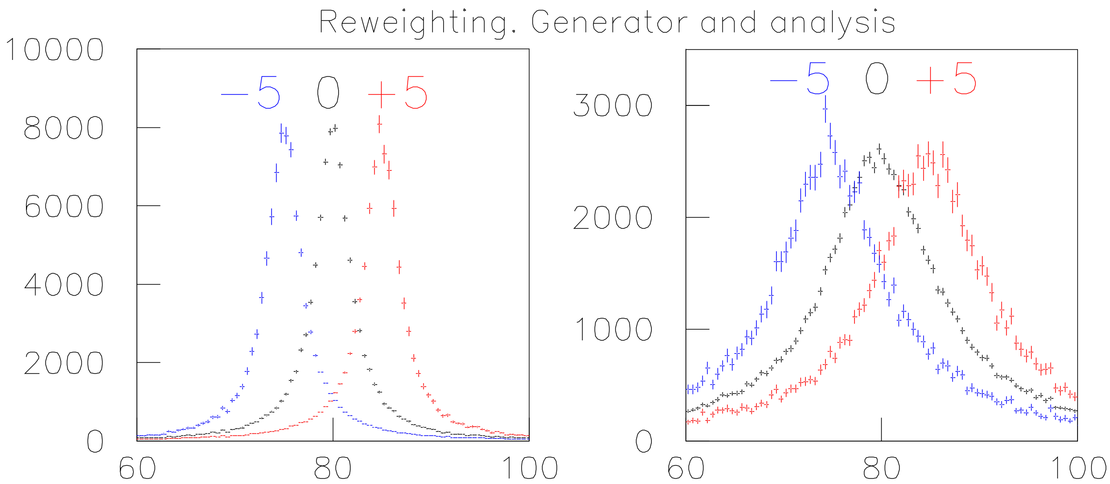


* processes not known from first principles

MC features

- Analytical functions
  - Fast and reliable
  - Integrable probabilities
- Numerical functions
  - Slower
  - Numerically less stable
- Decision trees
  - Make it hard to follow random numbers
  - Decorrelate random numbers
- Random number issues
  - EW: $\sim \mathbf{10^2}$ random numbers
  - QCD: $\sim \mathbf{10^4}$ random numbers
  - Detector $\sim \mathbf{10^6}$ random numbers
  - Pseudo random number quality only an issue for MC initialization. Others saved by decorrelation.

MC weights and reweighting



- Each used random number related to probability weights
  - When event weights are unity ⇒ generator default
  - Change any MC parameter ⇒ event weight changes
  - $w_i = P_i(p_{\text{new}})/P_i(p_{\text{gen}})$
  - When reweighting, variance increases
- Many parameters can easily be reweighted
  - Weights most often factorize
  - Event weights are normally computable
  - EW parameters are easiest $|\mathcal{M}(\vec{\Omega}, \vec{p})|^2$
  - QCD is most complicated
  - Reweighted MC is fully consistent

Weights in analyses

- HEP analyses are used to deal with weighted events
  - Samples generated with diff integrated luminosity
  - log-likelihood $\Rightarrow w_i \times$ log-likelihood
- Weighted events increase variance
  - $\mathtt{VAR} \propto (\sum_i w_i^2)/(\sum_i w_i)^2$
  - $n_{\mathrm{eq}} \propto (\sum_i w_i)^2/(\sum_i w_i^2)$
- Negative event weights are even allowed
  - Happens in some QCD/EW generators with negative interferences and matched calculations.
- Some minor complications:
  - Many weights and samples simultaneously
  - Large weights $\Leftrightarrow$ Large variance
    * Binned analyses become more difficult to control
    * Plots become less nice/convincing
    * Reweighted samples are highly correlated

Reweighting analyses

In the limit of infinite MC statistics a MC likelihood analysis is uniquely defined by:

$$L_i(\vec{\Omega}_i, \vec{p}) \propto \sum_{j \in i} P_j(\vec{\Omega}_j, \vec{p})$$

Generator probabilities provide analysis likelihoods

Selection can bias generator information

$$\sum_{\mathrm{allMC}} P_j(\vec{\Omega}_j, \vec{p}) = k(\vec{p})$$

$$L_i(\vec{\Omega}_i, \vec{p}) \propto (\sum_{j \in i} P_j(\vec{\Omega}_j, \vec{p}))/(\sum_{\mathrm{sel}} P(\vec{\Omega}, \vec{p}))$$

Full multi-dimensional likelihood $(\vec{p})$

Practical implementation

Choose few directions in observable space.

Bins and boxes, fixed or centered around data points

Too large: Undertraining and biases

Too small: Statistical fluctuations and overtraining

Challenges:

Overtraining means that there seems to be more information than there really is

Only fully consistent correction is to use independent MC samples to estimate unbiased performance

Correction procedures are non-trivial in general

Differential reweighting

Information and log-likelihoods can be expanded around the generated value, x_{gen}:

$$\ln(L_i(x)) = \ln(L_i(x_{\text{gen}}))$$
$$+\frac{\partial \ln(L_i(x))}{\partial x}(x-x_{\text{gen}})+\frac{1}{2}\frac{\partial^2 \ln(L_i(x))}{\partial x^2}(x-x_{\text{gen}})^2$$

$$a_i = \frac{\partial \ln(L_i(x))}{\partial x},\ a_{i2} = -\frac{\partial^2 \ln(L_i(x))}{\partial x^2}$$
$$\Rightarrow x - x_{\text{gen}} \simeq \left(\sum_i a_i\right)/\left(\sum_i a_{i2}\right)$$

Generator: $g_i = \frac{\partial \ln(P_i(x))}{\partial x},\ g_{i2} = -\frac{\partial^2 \ln(P_i(x))}{\partial x^2}$

A little algebra $\Rightarrow$

$$< a_i > = < g_i >,$$
$$< a_{i2} > = < g_{i2} > -(< g_i^2 > - < g_i >^2)$$

$< a_i >$ is "action" of event i: $\sim x_i/\sigma_i^2$

$< a_i >$ is easy to estimate and unbiased

$< a_{i2} >$ is "information" of event i: $\sim 1/\sigma_i^2$

$< a_{i2} >$ is fluctuating and overestimated $\mathcal{O}(1/n)$.

Reweighting and optimal observables

Some specific cases are simple to analyse

Triple Gauge Couplings in $e^+e^- \rightarrow W^+W^-$

Cross-sections are bilinear in the couplings

$$\frac{d\sigma}{d\vec{\Omega}} = \sum_{i,j} c_i c_j \frac{d\sigma_{ij}}{d\vec{\Omega}}$$

Few parameters $\Rightarrow$ full log-likelihood curves

$$\frac{d\sigma}{d\vec{\Omega}} = \frac{d\sigma_0}{d\vec{\Omega}} + c\frac{d\sigma_1}{d\vec{\Omega}} + c^2\frac{d\sigma_2}{d\vec{\Omega}}$$

$$O_1 = \frac{d\sigma_1}{d\vec{\Omega}} / \frac{d\sigma_0}{d\vec{\Omega}}, O_2 = \frac{d\sigma_2}{d\vec{\Omega}} / \frac{d\sigma_0}{d\vec{\Omega}}$$

O_1, O_2 are then optimal observables

The optimal observables act as projections of the event information on the coupling space.

When only first order terms are considered $\Rightarrow$ linear projections, which are identical to differential reweighting

Final analysis can be made a a full log-likelihood or as the sum of the optimal observables.

Large machinery which has been fully developed at LEP2

Is reweighting always possible?

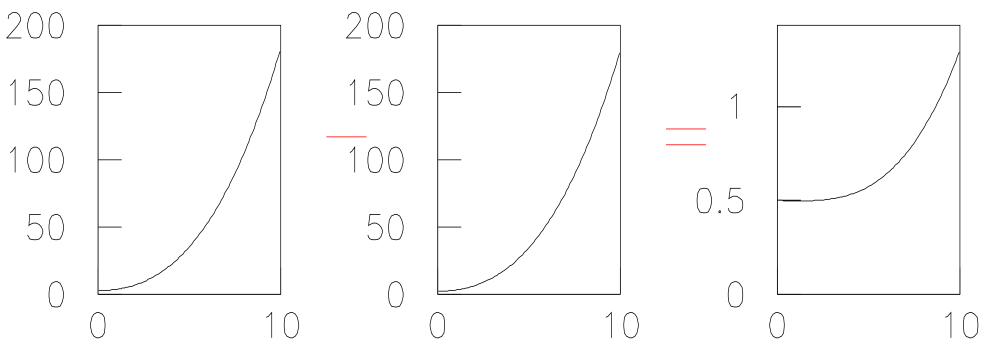
- EW type of reweighting
  - Matrix elements always available
  - Weights computable
  - Does not always work for exponentiated computations
- QCD type of reweighting
  - Single Probabilities available
  - Global parton shower weights not computable
- Matching of resolutions
  - If much more MC than data are available:
    * Include all parameters, changing them event-by-event
    * Choose generated distribution so that $\sigma_p < \sigma_a$ and $\sigma_p > \sigma_p(\mathrm{PDG})$
    * Reweight according to all parameters
  - MC cannot distinguish between reweighting physics quantities, PDG quantities, or event detector effects

Event by event analyses

The discussion assumes event-by-event analysis (log-likelihood curves). Are they always available?

- Global fits (Minuit)
  - Numerical problems when removing one event.
  - Covariance matrix $\sim$ invariant.
    Only change r.h.s. of optimization $\Longrightarrow$ stability
- Genuine correlations
  - Happens when purities are extracted from the data

Event-by-event likelihood can always be constructed by removing an event an see how the global results changes



Frequentist resampling

- The Bootstrap, the Jackknife, and sampling
  - Used to analyses the statistical properties of an analysis
  - Sampling of independent event samples
  - Bootstrap of event samples
  - Jackknife removing/adding event-by-event
- Control errors
  - Sampling always works, event extensive
  - Bootstrap does work $\sigma(\sigma) \sim \sigma/\sqrt{2n}$
  - Jackknife works on independent events, fastest
- Control analysis training
  - Works only for independent samples

Systematic errors I

- Assume MC is the whole “truth”
  - The data have no systematics
  - The analyses have no systematics
  - All systematics are in the MC model
- Systematics in MC
  - MC parameters
  - MC structures
- Can all systematics be included in one MC model?
  - Parameters: yes
  - Structures: More difficult
- Model independent measurements do not exist!
  - Is a set of observed events consistent with another set of events? It cannot be inferred without a model! Is the measurement of 2,6,10,14,18,22... consistent with 3,9,11,17,19...?
  - All models can be Monte Carlofied
  - All MC can be reweighted

Systematic errors II

- ElectroWeak like
  - Parameters $\Rightarrow$ Reweight parameters
  - $\mathcal{O}(\alpha) \Rightarrow$ Reweight events
  - Different schemes $\Rightarrow$ Use both as MC “truth”
- QCD like
  - QCD like parameters $\Rightarrow$ Generate and reweight
  - Gluon splitting $\Rightarrow$ reweight splittings
  - Different schemes $\Rightarrow$ Use both as MC “truth”
- Detector like
  - Bias correction on data
  - Detector shaking on MC $\Rightarrow$ Reweight resolutions

    Always make generated MC slightly “better” than data
  - More punch-through $\Rightarrow$ Reweight punch-through

Unifying I: What have we learned?

MC analysis is a nearly complete self-consistent paradigm

- MC analyzes data
- Data analyzes MC
- MC integrates in observable space
- MC differentiates in parameter space
- MC works with the chain rule
- MC handles all systematics
- MC analyzes analyses
- MC translates information from one data set to another
- MC is as model dependent as wished
- No model independent measurements exist

Unifying II: Analyzing analyses

HEP inference is a science

Nearly all our knowledge is built into MC programs

Inference inside the models is a technical process

Technical process control is engineering

- Linearized processes are easiest to control
- The Kalman filter is the optimal algorithm for controlling linear processes (See k2-filter talk)
- The analysis process must be controlled
- Combining multi-dimensional complex information is not unique to HEP
  - Learn from other fields?
  - Other fields learn from HEP?
  - HEP has be quite isolated w.r.t. other fields

Unifying III: What will this end up with?

- More Monte Carlo
- More reweightable parameters and more flexible simulation
- More Frequentist Bayesian collaboration?
- All analyses model dependent
  - Define “model” as MC or vice versa
  - Solve analysis problems like deconvolution and smoothing in this framework?
- HEP MC contains features we have grown familiar with and which are basically unknown to other fields
  - Phase space: The “mother” of MC.
    - Phase space should be there for all processes
    - Phase space dynamics is often very complex, e.g. constrained kinematical fits, correlations, compensations etc.
  - Process interferences
  - HEP is only unique in the sense that events are uncorrelated

Conclusions and outlook

- Development of MC analyses ongoing evolution
- Change of Paradigm, Model dependent, Systematics
- With MC/data you can do everything
- Within MC, inference is more a process than a science?
- Integrate and differentiate
- Intelligent and smart
- Well started is half finished
- Need good textbook for students

(In)famous last words

”God does not throw dice“

”Don’t tell God what to do”

”MC can simulate the World even if it is stochastic or not”

”Process control is a process which can be controlled by a process control which is a process which can be controlled by a process control which is a process which can be controlled by a process control which is a process which can be controlled by a process control which is a ...”

”Recursive algorithms are either tautologies,
not working, or only working by pure chance.”

”Think multi-dimensional/intuitively
Act mono-dimensional/logically”

”Alea jacta est”