Compression of Monte Carlo PDF replicas

Stefano Carrazza[1] and José Ignacio Latorre[2]

[1]University of Milan & CERN

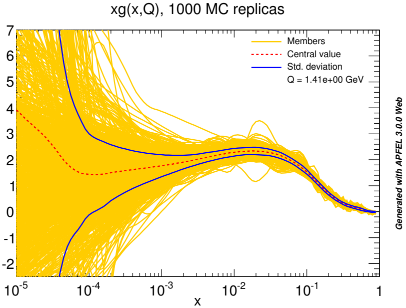
[2]University of Barcelona & National University of Singapore

PDF4LHC - CERN November 2014

Outline

Introducing the problem

Problem: **Reduce** the size of a PDF set of MC replicas with no **loss of information.**

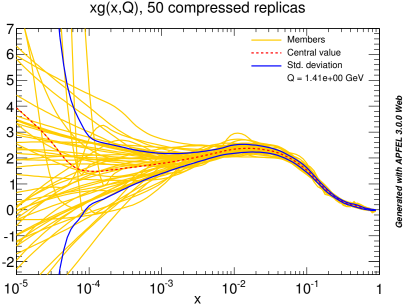


- Preserve the statistical properties of the prior PDF set.
- Avoid bias in the extrapolation region.
- Conserve physical requirements:
  - positivity
  - sum rules
  - PDF correlations

- No statistical properties conservation $\Rightarrow$ distortion of observables.
- Complex procedure, many features to identify and control.

Introducing the problem

Problem: **Reduce** the size of a PDF set of MC replicas with no **loss of information.**



- Preserve the statistical properties of the prior PDF set.
- Avoid bias in the extrapolation region.
- Conserve physical requirements:
  - positivity
  - sum rules
  - PDF correlations

- No statistical properties conservation ⇒ distortion of observables.
- Complex procedure, many features to identify and control.

Conservation of statistical properties of PDFs

Compress: Preserve as much as possible the underlying statistical distribution of a prior Monte Carlo PDF set.

- Starting from a large sample of N_{rep} Monte Carlo replicas:
  - find a compression algorithm to select $\tilde{N}_{rep} \ll N_{rep}$ replicas so that the basic properties of the underlying distribution are reproduced within some tolerance.

Conservation of statistical properties of PDFs

Compress: Preserve as much as possible the underlying statistical distribution of a prior Monte Carlo PDF set.

- Starting from a large sample of N_{rep} Monte Carlo replicas:
  - find a compression algorithm to select $\tilde{N}_{rep} \ll N_{rep}$ replicas so that the basic properties of the underlying distribution are reproduced within some tolerance.

Proposal:

Build a PDF set composed by Monte Carlo replicas from:

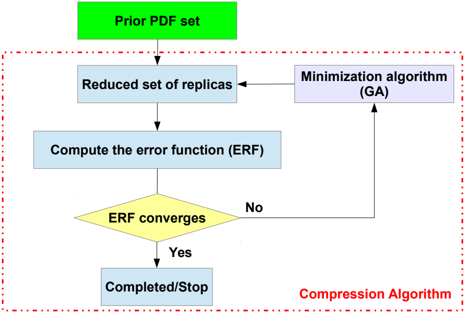
1. NNPDF, MMHT, CT14, HERAPDF2.0
   selected by some criterion (PDF errors, dataset, etc.)
2. Combine all these MC sets into a single distribution, e.g. as explained by Thorne and Watt, 1307.1347 Fig. 61.
3. Apply the compression algorithm and reduce the MC set

Towards a compression algorithm in 3 steps

- **Generate a prior MC PDF set:**
  - set with a large number of replicas

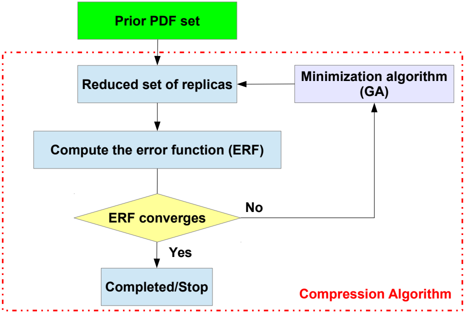
Towards a compression algorithm in 3 steps

- **Generate a prior MC PDF set:**
  - set with a large number of replicas
- **Select replicas** that minimize a convenient error function
  - minimization driven by a *genetic algorithm*



Towards a compression algorithm in 3 steps

- **Generate a prior MC PDF set:**
  - set with a large number of replicas
- **Select replicas** that minimize a convenient error function
  - minimization driven by a *genetic algorithm*



- **Validate the compressed set:**
  - verify estimators, PDF plots, predictions, χ^2, distances, etc.

Choosing the best error function

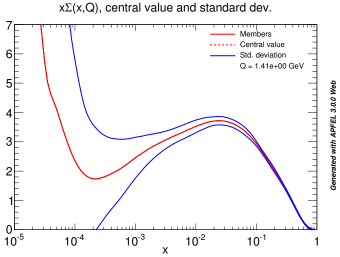
Possible ERF definitions:

Option A: Minimize the distance to the prior for the
$\Rightarrow$ Central Value and Standard Deviation

Choosing the best error function

Possible ERF definitions:

Option A: Minimize the distance to the prior for the
$\Rightarrow$ Central Value and Standard Deviation



Possibility to satisfy such criteria by selecting only 2 curves:

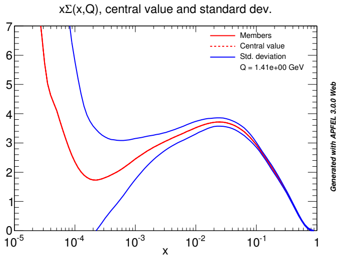
Bad Choice!

- bias of continuity, loss of structure
- dramatic loss of statistical information

Choosing the best error function

Possible ERF definitions:

Option A: Minimize the distance to the prior for the
$\Rightarrow$ Central Value and Standard Deviation



Possibility to satisfy such criteria by selecting only 2 curves:

Bad Choice!

- bias of continuity, loss of structure
- dramatic loss of statistical information

Problem: Higher moments not represented.

Choosing the best error function

Option B: Minimize the distance between 2 probability distributions:
$\Rightarrow$ Kolmogorov, Kullback, Chernhoff, L-distance
Higher moments are automatically adjusted

- **Kolmogorov:** simplest distance between probability distributions.
- **Kullback:** non-symmetric distance, encoding gain of information.
- **Chernhoff:** gives exponentially decreasing bounds on tail distributions of sums of independent random variables.
- **L-distance:** is a metric defined on a vector space where the distance between two vectors is the greatest of their differences along any coordinate dimension.

Choosing the best error function

Option B: Minimize the distance between 2 probability distributions:
$\Rightarrow$ Kolmogorov, Kullback, Chernhoff, L-distance
Higher moments are automatically adjusted

- **Kolmogorov:** simplest distance between probability distributions.
- **Kullback:** non-symmetric distance, encoding gain of information.
- **Chernhoff:** gives exponentially decreasing bounds on tail distributions of sums of independent random variables.
- **L-distance:** is a metric defined on a vector space where the distance between two vectors is the greatest of their differences along any coordinate dimension.

Problem: Ambiguity when defining the regions where the distance is computed. Large errors with few replicas.

Practical implementation

Practical Idea

Combine **options A and B** in a global error function.

Practical implementation

Practical Idea

Combine **options A and B** in a global error function.

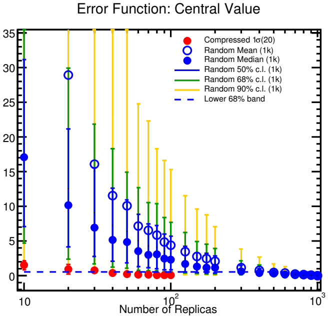
- We define the error function of an estimator E as

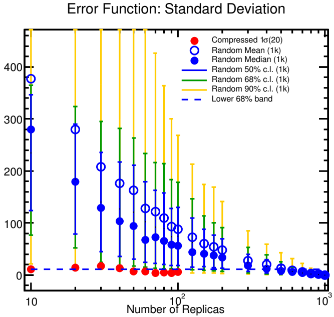
$$ERF_E = \frac{1}{N_E} \sum_{fl} \sum_{x} \left(E_{\mathrm{comp}} - \overline{E}_{\mathrm{prior}}\right)^2$$

  where N_E is a normalization weight.
- We construct a ERF which combines:
  - the first 4th moments: central value, std. dev., skewness, kurtosis
  - with the Kolmogorov distance
- Loop over all PDF flavors at the initial scale Q_0.
  In the next slides, ERF defined over 70 points in $x \in [10^{-5}, 0.9]$

Preliminary results

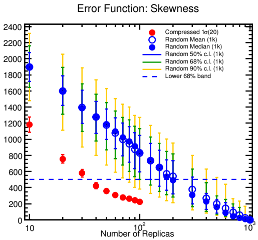
- Starting from a prior of 1000 replicas, we compare the ERF of:
  - 1k random sets, blue points.
  - compressed replicas, red points.

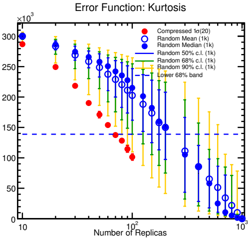


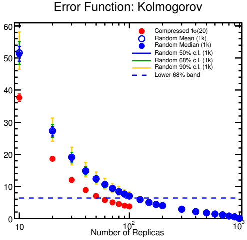


- Compression improves the description of CV and STD:
  - 100 random replicas $\sim$ 40-50 compressed replicas.

- Similar behavior is observed also for
  - skewness, kurtosis and Kolmogorov:



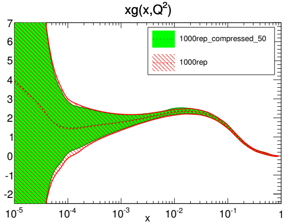


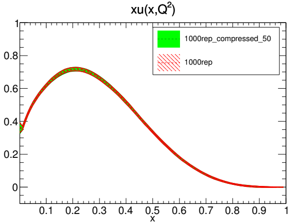


- Compression improves all estimators used in the ERF:
  - 100 random replicas $\sim$ 40-50 compressed replicas.

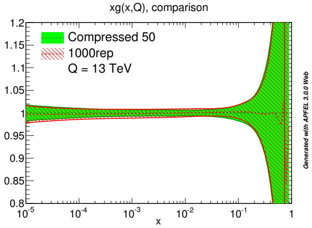
Validation: 1000→50 compressed

- Compression (1000→50) agreement at the level of PDF plots:



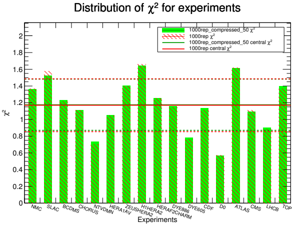


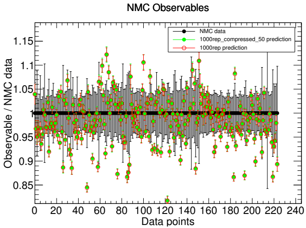
- Good agreement at initial and high Q values:



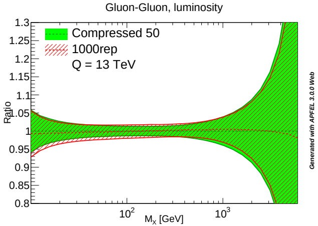
Validation: 1000→50 compressed

- Agreement at the level of χ^2 and predictions:



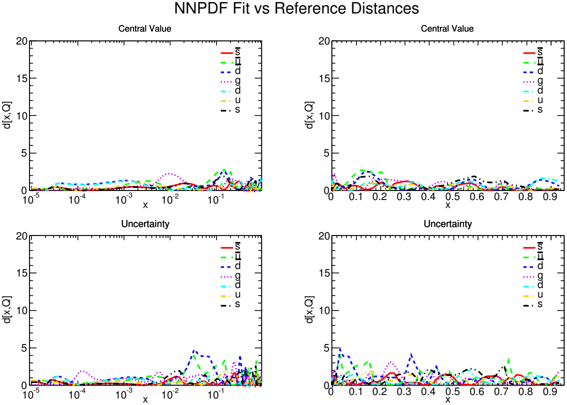


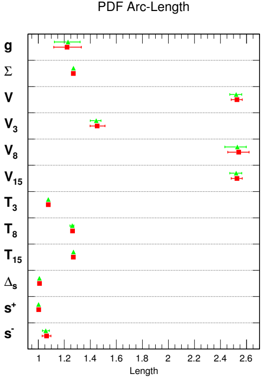
- Agreement at the level of luminosities:



Validation: 1000→50 compressed

- PDF distances and arc-length:

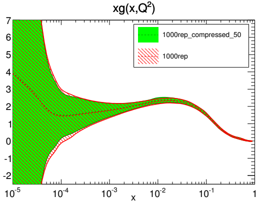




- The compression preserves the properties of the prior set.

Outlook

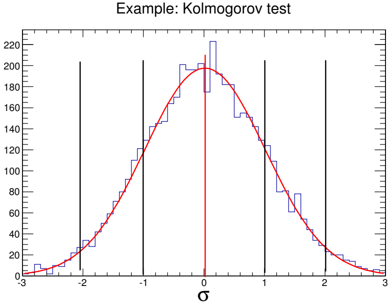
- Conclusion:
  - in this preliminary study we show that the reduction 1000 $\to$ 50 is possible to achieve with the compression algorithm.



- Outlook:
  - study the behavior when varying the *x* points for the ERF
  - study the Kolmogorov normalization impact
  - study other distances
  - start from a larger set of 10k replicas
  - analyze each PDF separately

Kolmogorov test

- For each PDF flavor, for each x point where the ERF is defined we divide the distribution in 6 regions delimited by multiples of the standard deviation.

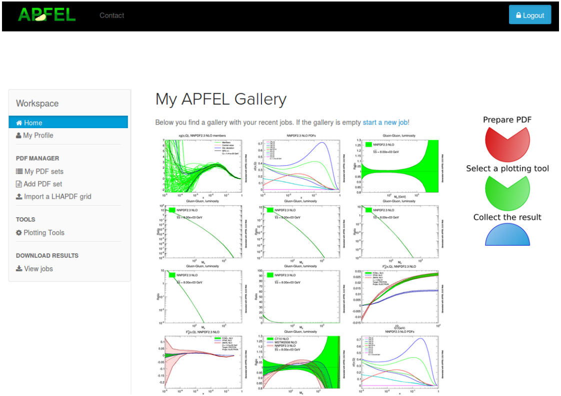


- We construct the rate of replicas in each region for the prior and the compressed set. This quantity is then introduced in the ERF.

APFEL Web

APFEL Web: a web-based application for the graphical visualization of PDFs.

http://apfel.mi.infn.it



See references: arXiv:1310.1394, arXiv:1410.5456.