The NNPDF4.0 global analysis of the proton structure

Roy Stegeman
University of Milan and INFN Milan

On behalf of the NNPDF Collaboration
Based on: arXiv:2109.02653

DIS 2022, 3 May 2022

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 740006.
High-precision: gluon

\[ \mathcal{L}_{ij}(M_X, y, \sqrt{s}) = \frac{1}{s} \sum_{i,j} f_i \left( \frac{M_X e^y}{\sqrt{s}}, M_X \right) f_j \left( \frac{M_X e^{-y}}{\sqrt{s}}, M_X \right) \]

Relative uncertainty for gg-luminosity
NNPDF3.1 (NNLO) - \( \sqrt{s} = 14000.0 \) GeV

Relative uncertainty for gg-luminosity
NNPDF4.0 (NNLO) - \( \sqrt{s} = 14000.0 \) GeV

How did we get here?
High-precision: singlet

\[ \mathcal{L}_{ij}(M_X, y, \sqrt{s}) = \frac{1}{s} \sum_{i,j} f_i \left( \frac{M_X e^y}{\sqrt{s}}, M_X \right) f_j \left( \frac{M_X e^{-y}}{\sqrt{s}}, M_X \right) \]

Relative uncertainty for qq-luminosity
NNPDF3.1 (NNLO) - \( \sqrt{s} = 14000.0 \) GeV

Relative uncertainty for qq-luminosity
NNPDF4.0 (NNLO) - \( \sqrt{s} = 14000.0 \) GeV

How did we get here?
Data
Data from NNPDF1.0 to NNPDF4.0

The number of datasets – normally corresponding to different processes – is generally more relevant than the number of datapoints.
Experimental data in NNPDF4.0

Kinematic coverage

New processes:
- direct photon
- single top
- dijets
- W+jet
- DIS jet
Methodology
Improved fitting methodology

- **Stochastic Gradient Descent** for NN training using TensorFlow
- Automated optimization of model hyperparameters
- Methodology is validated using closure tests (data region), future tests (extrapolation region), and parametrization basis independence

Physical constraints:
- PDF positivity
- Integrability of nonsinglet distributions (Gottfried sum rules)

\[
f_i \left( x, Q_0 \right) = x^{-\alpha_i} \left( 1 - x \right) ^{\beta_i} \text{NN}_i \left( x \right)
\]
Automated model selection

NNPDF aims to minimize sources of bias in the PDF:

- Functional form $\rightarrow$ Neural Network
- Model parameters $\rightarrow$ ?
Automated model selection

NNPDF aims to minimize sources of bias in the PDF:
- Functional form → Neural Network
- Model parameters → Hyperoptimization

Scan over thousands of hyperparameter combinations and select the best one

**k-fold cross-validation**: used to define the reward function based on a test dataset

Objective function:
\[ L = \text{mean}(\chi_1^2, \chi_3^2, \chi_2^2, \ldots, \chi_k^2) \]
<table>
<thead>
<tr>
<th>NNPDF4.0</th>
<th>Data</th>
<th>Methodology</th>
<th>Stability</th>
<th>LHC phenomenology</th>
<th>Open-source code</th>
<th>Summary and Outlook</th>
</tr>
</thead>
</table>

**Stability**
Parametrization basis independence

Evolution Basis:

\[ xV(x, Q_0) \propto NN_V(x) \]

\[ xT_3(x, Q_0) \propto NN_{T_3}(x) \]

Flavour Basis:

\[ xV(x, Q_0) \propto (NN_u(x) - NN_{\bar{u}}(x) + NN_d(x) - NN_{\bar{d}}(x) + NN_s(x) - NN_{\bar{s}}(x)) \]

\[ xT_3(x, Q_0) \propto (NN_u(x) + NN_{\bar{u}}(x) - NN_d(x) - NN_{\bar{d}}(x)) \]

Different strategies to parametrize the quark PDF flavour combinations leave the uncertainties essentially unchanged.
Impact of the new data

Individual datasets have a limited impact, but collectively they result in:

- Moderate reduction of PDF uncertainties
- Shifts in central value at the one-sigma level
Impact of the new fitting methodology

- Significant reduction of PDF uncertainties
- Good agreement between the central values

PDF uncertainties are validated using closure tests and future tests. Validation tests successful for both NNPDF4.0 and NNPDF3.1
LHC phenomenology
Implications for LHC phenomenology

Reduced luminosity uncertainties $\rightarrow$ Reduced uncertainty at the level of observables
Open-source code
The open-source NNPDF code

The full NNPDF code has been made public along with user friendly documentation

This includes: fitting, hyperoptimization, theory, data processing, visualization

It is possible to reproduce all results of NNPDF4.0 and more!

https://github.com/NNPDF/nnpdf
https://docs.nnpdf.science
Summary and Outlook
Summary and Outlook

- NNPDF4.0 is the latest release in the NNPDF family of PDF sets
- 44 new datasets from many new processes are included
- Improved methodology with Stochastic Gradient Descent and hyperoptimization
- Validation of PDF uncertainties using closure test, future test and parametrization basis independence
  ⇒ NNPDF4.0 achieves a high precision over a broad kinematic range

- The current level of PDF uncertainties challenges the accuracy of theoretical predictions and demands an increased effort towards the systematic inclusion in the fit of theoretical uncertainties (nuclear, higher orders, SM parameters, . . . ) and higher-order QCD and EW corrections
Summary and Outlook

- NNPDF4.0 is the latest release in the NNPDF family of PDF sets
- 44 new datasets from many new processes are included
- Improved methodology with Stochastic Gradient Descent and hyperoptimization
- Validation of PDF uncertainties using closure test, future test and parametrization basis independence

⇒ NNPDF4.0 achieves a high precision over a broad kinematic range

- The current level of PDF uncertainties challenges the accuracy of theoretical predictions and demands an increased effort towards the systematic inclusion in the fit of theoretical uncertainties (nuclear, higher orders, SM parameters, ... ) and higher-order QCD and EW corrections

Thank you!
Backup
Experimental data in NNPDF4.0

- 44 new datasets included
- 323 more data points in NNPDF4.0 than in NNPDF3.1
- New data is mostly from the LHC RUN II
NNPDF4.0 model
For more information see EPJ C79 (2019) 676

PDF = $A x^\alpha (1 - x)^\beta \text{NN}(x, \log x)$

Main changes:
- Python codebase
  - Easier and faster development
- Freedom to use external libraries (default: TensorFlow)
- Modularity $\Rightarrow$ ability to vary all aspects of the methodology
**Performance benefit - time per replica**

<table>
<thead>
<tr>
<th></th>
<th>NNPDF3.1</th>
<th>NNPDF4.0 (CPU)</th>
<th>NNPDF4.0 (GPU)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fit timing per replica</td>
<td>15.2 h</td>
<td>38 min</td>
<td>6.6 min</td>
</tr>
<tr>
<td>Speed up factor</td>
<td>1</td>
<td>24</td>
<td>140</td>
</tr>
<tr>
<td>RAM use</td>
<td>1.5 GB</td>
<td>6.1 GB</td>
<td>NA</td>
</tr>
</tbody>
</table>
Hyperoptimization: the reward function

Choosing as the hyperoptimization target the $\chi^2$ of fitted data results in overfitting.
Hyperoptimization: the reward function

Choosing as the hyperoptimization target the $\chi^2$ of fitted data results in overfitting.

We solve this using **k-fold cross-validation**:

1. Divide the data into $k$ representative subsets
2. Fit $k - 1$ sets and use $k$-th as test set  
   $\Rightarrow$ $k$ values of $\chi^2_{\text{test}}$
3. Optimize the average $\chi^2_{\text{test}}$ of the $k$ test sets  
   $\Rightarrow$ The hyperoptimization target is not based on data that entered the fit.

- No overfitting
- Compared to NNPDF3.1:
  - Increased stability
  - Reduced uncertainties
The (negligible) impact of datasets with tension

Excluding datasets with large \((\chi^2 - 1)/\sigma\chi^2\) one at a time and combining the resulting PDFs following the conservative PDF4LHC15 prescription shows stability at the level of statistical fluctuations.
Envelope of fits with different arametrisation bases

Different strategies to parametrize the PDF flavour combinations lead to the same result
Understanding the $\chi^2$ distribution

Experimental $\chi^2$

Experiments $\chi^2$ distribution

- $\chi^2_{\text{data}} = 1.161$
- $\chi^2_{\text{rep., data}} = 1.186 \pm 0.014$
- $\chi^2$ rep. distr.

$t_0 \chi^2$

Experiments $\chi^2$ distribution

- $\chi^2_{\text{data}} = 1.233$
- $\chi^2_{\text{rep., data}} = 1.258 \pm 0.011$
- $\chi^2$ rep. distr.
Impact of positivity on the PDFs

\[ \tilde{d}(x) \text{ at 1.7 GeV} \]

NNPDF4.0 data set, NNPDF3.1 positivity
NNPDF4.0 baseline
More implications for phenomenology