Towards a global NNPDF analysis

Maria Ubiali

School of Physics, University of Edinburgh

DIS 2009 Workshop
Madrid, 28 April 2009
Work in collaboration

**NNPDF collaboration**

R.D.Ball\(^1\), L.Del Debbio\(^1\), S.Forte\(^2\), A.Guffanti\(^3\), J.I.Latorre\(^4\), A. Piccione\(^2\), J. Rojo\(^2\), M.U.\(^1\)

\(^1\) PPT Group, School of Physics, University of Edinburgh
\(^2\) Dipartimento di Fisica, Universit`a di Milano
\(^3\) Physikalisches Institut, Albert-Ludwigs-Universit¨at Freiburg
\(^4\) Departament d’Estructura i Constituents de la Matèria, Universitat de Barcelona


NNPDF collaboration, [arXiv:0811.2288] **NNPDF1.1**

NNPDF collaboration, in preparation **NNPDF1.2**

NNPDF collaboration, in preparation **NNPDF2.0**

Maria Ubiali    NNPDF partons for LHC analyses
Outline

1 Introduction
   - Parton fits
   - NNPDF approach: the main ingredients

2 Results
   - NNPDF1.0
   - NNPDF1.1
   - NNPDF1.2
   - NNPDF2.0

3 NNPDF benchmark
   - Error propagation
   - Effect of the parametrization
   - Inclusion of new data
   - Incompatible data

4 Conclusions
Outline

1. Introduction
   - Parton fits
   - NNPDF approach: the main ingredients

2. Results
   - NNPDF1.0
   - NNPDF1.1
   - NNPDF1.2
   - NNPDF2.0

3. NNPDF benchmark
   - Error propagation
   - Effect of the parametrization
   - Inclusion of new data
   - Incompatible data

4. Conclusions
Parton Distribution Functions

- **Factorization Theorem** \( (Q^2 \gg \Lambda_{QCD}^2) \):

\[
\frac{d\sigma_H}{dX} = \sum_{a,b} \int dx_1 dx_2 f_a(x_1, \mu_f) f_b(x_2, \mu_f) \otimes \frac{d\hat{\sigma}}{dX}(\alpha_s(\mu_r), \mu_r, \mu_f, x_1, x_2, Q^2)
\]

- **DGLAP equations**:

\[
\frac{d}{dt} \begin{pmatrix} q \\ g \end{pmatrix} = \frac{\alpha_s}{2\pi} \begin{pmatrix} P_{qq} & P_{qg} \\ P_{gq} & P_{gg} \end{pmatrix} \otimes \begin{pmatrix} q \\ g \end{pmatrix} + O(\alpha_s^2)
\]

- Need robust input for analyses at LHC.
- Need statistically reliable interpretation for PDFs error bars: will provide dominant contribution to systematic uncertainties for some processes.
Introduction

Results

NNPDF benchmark

Conclusions

Parton fits

NNPDF approach

NNPDF approach

Experimental Data

MC generation

NN parametrization

Fi \ i=1,\ldots,N_{\text{data}}

NMC, BCDMS, SLAC, HERA, CHORUS...

\{Fi(1), Fi(2), Fi(N-1), Fi(N)\}

TRAINING

EVOLUTION

\{q0_{\text{net}}(1), q0_{\text{net}}(2), q0_{\text{net}}(N-1), q0_{\text{net}}(N)\}

REPRESENTATION OF PROBABILITY DENSITY

Maria Ubiali

NNPDF partons for LHC analyses
NNPDF approach

Determination of unbiased PDFs with faithful estimation of their uncertainties.

\[
\langle \mathcal{F}[f_i(x)] \rangle = \int [\mathcal{D}f_i] \mathcal{F}[f_i(x)] \mathcal{P}[f_i(x)] \rightarrow \frac{1}{N_{\text{rep}}} \sum_{k=1}^{N_{\text{rep}}} \mathcal{F}[f_i^{(k)(\text{net})}(x)]
\]

\[
\sigma_{\mathcal{F}[f(x)]} = \sqrt{\langle \mathcal{F}[f(x)]^2 \rangle - \langle \mathcal{F}[f(x)] \rangle^2}
\]

* The measure \( \mathcal{P}[f_i(x)] \) in space of PDFs is determined with MC method.
* Use all information contained in experiments.
* Redundant parametrization of PDFs: reduce bias.
* Statistic estimators to assess errors, correlations, stability and size of systematics.
* All details explained in hep-ph/0809.3716(NPB).
Outline

1 Introduction
   • Parton fits
   • NNPDF approach: the main ingredients

2 Results
   • NNPDF1.0
   • NNPDF1.1
   • NNPDF1.2
   • NNPDF2.0

3 NNPDF benchmark
   • Error propagation
   • Effect of the parametrization
   • Inclusion of new data
   • Incompatible data

4 Conclusions
NNPDF1.0: Experimental data

- **Kinematical cuts:**
  - $Q^2 > 2 \text{ GeV}^2$
  - $W^2 = Q^2 (1 - x)/x > 12.5 \text{ GeV}^2$
  - $\sim 3000$ points.
**NNPDF1.0: Parametrization**

**Parametrization** of 5 combinations of PDFs at $Q_0^2 = 2 \text{ GeV}^2$

- Singlet: $\Sigma(x) \mapsto NN_{\Sigma}(x)$ 2-5-3-1 37 pars
- Gluon: $g(x) \mapsto NN_{g}(x)$ 2-5-3-1 37 pars
- Total valence: $V(x) \equiv u_V(x) + d_V(x) \mapsto NN_{V}(x)$ 2-5-3-1 37 pars
- Non-singlet triplet: $T_3(x) \mapsto NN_{T3}(x)$ 2-5-3-1 37 pars
- Sea asymmetry: $\Delta_S(x) \equiv \bar{d}(x) - \bar{u}(x) \mapsto NN_{\Delta}(x)$ 2-5-3-1 37 pars

185 parameters
NNPDF1.0: Partons

Graphs showing the distribution of parton densities and their uncertainties for different parton sets (CTEQ6.5, MRST2001E, Alekhin02, NNPDF1.0) at various scales (x, Q^2).
NNPDF1.1: A consistency check

NNPDF1.0: flavor assumptions, symmetric strange sea proportional to non-strange sea according to $C_s \sim 0.5$ suggested by neutrino DIS data.

\[ s(x) = \bar{s}(x) \quad \bar{s}(x) = \frac{C_s}{2}(\bar{u}(x) + \bar{d}(x)) \]

NNPDF1.1: independent parametrization of the strange content of the nucleon.

Total strangeness: $s^+(x) \equiv (s(x) + \bar{s}(x))/2 \quad \rightarrow \quad \text{NN}_{(s^+)}(x) \quad 2-5-3-1 \ 37 \ \text{pars}$

Strangeness valence: $s^-(x) \equiv (s(x) - \bar{s}(x))/2 \quad \rightarrow \quad \text{NN}_{(s^-)}(x) \quad 2-5-3-1 \ 37 \ \text{pars}$

Randomized preprocessing.

Large uncertainty for strange PDFs.

Same $\chi^2$ and statistical features of the fit.

Check of stability and consistency of our statistically-sound approach.
NNPDF1.2: Constrain the strange distribution

- Neutrino and anti-neutrino dimuon production from NuTeV.
- HERA-II ZEUS data on NC and CC reduced xsec at large-$Q^2$.
- HERA-II ZEUS data on $xF_3^\gamma Z$.

\[ \bar{\sigma}^{\nu(\bar{\nu}),c} \propto (F_2^{\nu(\bar{\nu}),c}, F_3^{\nu(\bar{\nu}),c}, F_L^{\nu(\bar{\nu}),c}) \]
\[ F_2^{\nu,c} = x \left[ C_{2,q} \otimes 2 |V_{cs}|^2 s + \frac{1}{n_f} C_{2,g} \otimes g \right] \]
\[ F_2^{\bar{\nu},c} = x \left[ C_{2,q} \otimes 2 |V_{cs}|^2 \bar{s} + \frac{1}{n_f} C_{2,g} \otimes g \right] \]

Results
- No bias on the shape or normalization of strange valence and total strange.
- The only constraint comes from strange valence sum rule.
- Faithful determination of uncertainties of strange content of the nucleon.
- Interesting applications: NuTeV anomaly, $V_{cs}$ determination...

See Joan Rojo talk.
Towards a global neural fit: **NNPDF2.0**

- The inclusion of hadronic data is necessary to constrain large-$x$ gluon behavior, sea quarks, $u/d$ ratio at large $x$.

![Graph showing gluon distribution](image1.png)

![Graph showing parton distribution](image2.png)

- Upcoming **NNPDF2.0** is the first neural global fit including most available hadronic data.
- Fit soon competitive with MSTW and CTEQ parton fits.
Inclusion of fixet-target Drell-Yan data, Tevatron electroweak gauge boson production, Run II inclusive jet data from Tevatron.

\[ 800_{\text{DY}} + 200_{\text{JET}} = \mathcal{O}(1000) \text{ new data.} \]
Predictions evaluated with NNPDF1.2 and NNPDF2.0 (prel) error sets.
- NNPDF1.2: Large error bands on predictions, compatible with data.
- NNPDF2.0: Smaller error bands, data are well described.
Predictions evaluated with NNPDF1.2 and NNPDF2.0 (prel) error sets.

- **NNPDF1.2**: Large error bands on predictions, compatible with data.
- **NNPDF2.0**: Smaller error bands, data are well described.
NNPDF2.0: Predictions on Observables (preliminary)

- Predictions evaluated with NNPDF1.2 and NNPDF2.0 (prel) error sets.
- NNPDF1.2: Large error bands on predictions, compatible with data.
- NNPDF2.0: Smaller error bands, data are well described.
- Error on $V_{TOT}$ and $\bar{d}$ reduced due to inclusion of hadronic data.
- Partons compatible with previous determination.
The NLO computation of hadronic observables might be too slow for parton global fits.

Many parton fits rely on K-factor approximation, relatively fast.

K-factor depends on PDFs and it is not always a good approximation.

* NNPDF2.0 includes full NLO calculation of hadronic observables.
* Use available fastNLO interface for jet inclusive cross-sections.[hep-ph/0609285]
* Built up our own fastNLO-like evolution for Drell-Yan observables, not available in literature.
* Fast code easy to benchmark versus other slow codes.
Outline

1 Introduction
   - Parton fits
   - NNPDF approach: the main ingredients

2 Results
   - NNPDF1.0
   - NNPDF1.1
   - NNPDF1.2
   - NNPDF2.0

3 NNPDF benchmark
   - Error propagation
   - Effect of the parametrization
   - Inclusion of new data
   - Incompatible data

4 Conclusions
To understand differences between available approaches for determining parton uncertainties and to deal with incompatible data:

- Fix experimental data, QCD parameters, input parametrizations, error treatment.
- Assess the effect of the remaining assumptions.

**HERA-LHC 08 workshop**

- H1 benchmark $\rightarrow$ Hessian vs Monte Carlo error propagation.
- H1-NNPDF benchmark $\rightarrow$ Monte Carlo error propagation: polynomial vs NN parametrization.
- H1-NNPDF benchmark $\rightarrow$ Treatment of incompatible data.
- NNPDF-MRST/MSTW benchmark $\rightarrow$ Stability vs inclusion of new data. Confirmed by NNPDF2.0 plots.
I) H1 Benchmark: compare error propagation

- **Settings:** same theoretical and flavour assumptions, same parametrization, same data and kinematical cuts.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Data points</th>
<th>Observable</th>
</tr>
</thead>
<tbody>
<tr>
<td>H197mb</td>
<td>35</td>
<td>$\bar{\sigma}^{NC,+}$</td>
</tr>
<tr>
<td>H197lowQ2</td>
<td>80</td>
<td>$\bar{\sigma}^{NC,+}$</td>
</tr>
<tr>
<td>H197NC</td>
<td>130</td>
<td>$\bar{\sigma}^{NC,+}$</td>
</tr>
<tr>
<td>H197CC</td>
<td>25</td>
<td>$\bar{\sigma}^{CC,+}$</td>
</tr>
<tr>
<td>H199NC</td>
<td>126</td>
<td>$\bar{\sigma}^{NC,-}$</td>
</tr>
<tr>
<td>H199CC</td>
<td>28</td>
<td>$\bar{\sigma}^{CC,-}$</td>
</tr>
<tr>
<td>H199NChy</td>
<td>13</td>
<td>$\bar{\sigma}^{NC,-}$</td>
</tr>
<tr>
<td>H100NC</td>
<td>147</td>
<td>$\bar{\sigma}^{NC,+}$</td>
</tr>
<tr>
<td>H100CC</td>
<td>28</td>
<td>$\bar{\sigma}^{CC,+}$</td>
</tr>
</tbody>
</table>

- **Compare error propagation:** standard Hessian vs MC method (sampling of the DATA).

- **Results:** MC (gaussian and lognorm distrib) and standard error propagation give the same error!!!

(ArXiv:0901.2504 Feltesse, Glazov, Radescu)
II) H1-NNPDF Benchmark: compare parametrization

- **Settings**: same theoretical and flavour assumptions, same data and kinematical cuts.
- Same error propagation via Monte Carlo technique.
- Compare polynomial vs Neural Network redundant parametrization.
- **Results**: Larger spread for gluon due to flexibility of parametrization.
  
  ArXiv:0901.2504
III) NNPDF-MRST/MSTW Benchmark: stability versus data

- Same parametrization, flavour assumption and statistical treatment for benchmark and full DIS analysis.
- NNPDFbench is consistent with NNPDF1.0.
- Stability versus addition of new data. ArXiv:0901.2504
- Same stability in preliminary NNPDF2.0 parton distributions!!
IV) H1-NNPDF Benchmark: treatment of inconsistent data

CC cross-section: uncertainty shrinks because of the inclusion of CHORUS data (consistent to H1 data) in NNPDF1.0 analysis.

NC cross-section: inclusion of ZEUS data (slightly inconsistent to H1 data) in NNPDF1.0 analysis does not shrink uncertainty.

Systematic disagreement ZEUS vs H1 NC data is of the size of the uncertainty band.

Inconsistency in data does not allow to reduce uncertainty beyond size of inconsistency!!
Outline

1 Introduction
   - Parton fits
   - NNPDF approach: the main ingredients

2 Results
   - NNPDF1.0
   - NNPDF1.1
   - NNPDF1.2
   - NNPDF2.0

3 NNPDF benchmark
   - Error propagation
   - Effect of the parametrization
   - Inclusion of new data
   - Incompatible data

4 Conclusions
The first NNPDF1.0 parton set [arXiv:0808.1231] from a comprehensive DIS analysis is available on the common LHAPDF interface (http://projects.hepforge.org/lhapdf), the NNPDF1.1 is available on the NNPDF website (http://sophia.ecm.ub.es/nnpdf/)

- Inclusion of NuTev data constrains the strange distribution in the upcoming NNPDF1.2 fit.
- Inclusion of hadronic data (DY, jets, W asymmetry): first global NNPDF2.0 fit.
- Encouraging preliminary results!
- Implementation of a full fastNLO-like evolution strategy for hadronic observable, including Drell-Yan.
NNPDF APPROACH
Monte Carlo sample

Generate a $N_{\text{rep}}$ Monte Carlo sets of artificial data, or "pseudo-data" of the original $N_{\text{data}}$ data points

$$F_i^{(\text{exp})}(x_p, Q_p^2) \equiv F_i^{(\text{exp})} \rightarrow F_i^{(\text{art})(k)}(x_p, Q_p^2) \equiv F_i^{(\text{art})(k)}$$

$i = 1, ..., N_{\text{data}}$

$$k = 1, ..., N_{\text{rep}}$$

Multi-gaussian distribution centered on each data point:

$$F_i^{(\text{art})(k)} = S^{(k)}_{p,N} F_i^{(\text{exp})} \left( 1 + r_p^{(K)} \sigma_p^{\text{stat}} + \sum_{j=1}^{N_{\text{sys}}} r_{p,j}^{(k)} \sigma_{p,j}^{\text{sys}} \right)$$

If two points have correlated systematic uncertainties

$$r_{p,j}^{(k)} = r_{p',j}^{(k)}$$

Correlations are properly taken into account.
**Validation of the MC sample**

<table>
<thead>
<tr>
<th>Experiment</th>
<th>ZEUS</th>
<th>CHORUS</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\langle PE \left[ \langle F(\text{art}) \rangle_{\text{rep}} \right] \rangle_{\text{dat}}$</td>
<td>$8.5 \cdot 10^{-4}$</td>
<td>$1.8 \cdot 10^{-3}$</td>
<td>$7.1 \cdot 10^{-5}$</td>
</tr>
<tr>
<td>$r \left[ F(\text{art}) \right]$</td>
<td>1.000</td>
<td>1.000</td>
<td>0.980</td>
</tr>
<tr>
<td>$\langle PE \left[ \langle \sigma(\text{art}) \rangle_{\text{rep}} \right] \rangle_{\text{dat}}$</td>
<td>$9.6 \cdot 10^{-3}$</td>
<td>$1.8 \cdot 10^{-2}$</td>
<td>$3.0 \cdot 10^{-3}$</td>
</tr>
<tr>
<td>$\langle \sigma(\text{exp}) \rangle_{\text{dat}}$</td>
<td>0.0607</td>
<td>0.1088</td>
<td>0.0556</td>
</tr>
<tr>
<td>$\langle \sigma(\text{art}) \rangle_{\text{dat}}$</td>
<td>0.0603</td>
<td>0.1109</td>
<td>0.0562</td>
</tr>
<tr>
<td>$r \left[ \sigma(\text{art}) \right]$</td>
<td>1.000</td>
<td>0.998</td>
<td>0.980</td>
</tr>
<tr>
<td>$\langle \rho(\text{exp}) \rangle_{\text{dat}}$</td>
<td>0.079</td>
<td>0.650</td>
<td>0.145</td>
</tr>
<tr>
<td>$\langle \rho(\text{art}) \rangle_{\text{dat}}$</td>
<td>0.082</td>
<td>0.657</td>
<td>0.146</td>
</tr>
<tr>
<td>$r \left[ \rho(\text{art}) \right]$</td>
<td>0.982</td>
<td>0.996</td>
<td>0.996</td>
</tr>
<tr>
<td>$\langle \text{cov}(\text{exp}) \rangle_{\text{dat}}$</td>
<td>$1.53 \cdot 10^{-4}$</td>
<td>$2.03 \cdot 10^{-2}$</td>
<td>$1.07 \cdot 10^{-3}$</td>
</tr>
<tr>
<td>$\langle \text{cov}(\text{art}) \rangle_{\text{dat}}$</td>
<td>$1.57 \cdot 10^{-4}$</td>
<td>$2.11 \cdot 10^{-2}$</td>
<td>$1.01 \cdot 10^{-3}$</td>
</tr>
<tr>
<td>$r \left[ \text{cov}(\text{art}) \right]$</td>
<td>0.996</td>
<td>0.998</td>
<td>0.997</td>
</tr>
</tbody>
</table>

A MC sample with $\mathcal{O}(1000)$ replicas reproduces mean values, variances, correlations of experimental data within 1% accuracy.

**Convergence rate** increases with $N_{\text{rep}}$. 

*Maria Ubiali*  
NNPDF partons for LHC analyses
For each replica \( (k) \) of the experimental data we fit a set of independent PDFs. Ensemble of fitted replicas of PDFs: representation of the probability distribution in the space of PDFs.

Uncertainties, central values and any other statistical property (e.g. correlations) of the PDFs (or any function of them) can be evaluated using standard statistical methods.

\[
\langle \mathcal{F}[f(x)] \rangle = \frac{1}{N_{\text{rep}}} \sum_{k=1}^{N_{\text{rep}}} \mathcal{F}[f^{(k)}_{\text{net}}(x)]
\]

\[
\sigma_{\mathcal{F}[f(x)]} = \sqrt{\langle \mathcal{F}[f(x)]^2 \rangle - \langle \mathcal{F}[f(x)] \rangle^2}
\]

\[
\rho[f_a(x_1, Q_1^2), f_b(x_2, Q_2^2)] = \frac{\langle f_a(x_1, Q_1^2)f_b(x_2, Q_2^2) \rangle - \langle f_a(x_1, Q_1^2) \rangle \langle f_b(x_2, Q_2^2) \rangle}{\sigma_a(x_1, Q_1^2)\sigma_b(x_2, Q_2^2)}
\]
How PDFs uncertainties must be evaluated

- Monte Carlo prescription ([NNPDF](#))

$$
\sigma_{\mathcal{F}} = \left( \frac{N_{\text{set}}}{N_{\text{set}} - 1} \left( \langle \mathcal{F}[\{f\}]^2 \rangle - \langle \mathcal{F}[\{f\}] \rangle^2 \right) \right)^{1/2}
$$

- HEPDATA prescription ([CTEQ](#) and [MRST/MSTW](#))

$$
\sigma_{\mathcal{F}} = \frac{1}{2C_{90}} \left( \sum_{k=1}^{N_{\text{set}}/2} \left( \mathcal{F}[\{f^{(2k-1)}\}] - \mathcal{F}[\{f^{(2k)}\}] \right)^2 \right)^{1/2}, \quad C_{90} = 1.64485
$$

$C_{90}$ accounts for the fact that the upper and lower parton sets correspond to 90% confidence levels rather than to one-$\sigma$ uncertainties.

- HEPDATA* prescription ([Alekhin](#))

$$
\sigma_{\mathcal{F}} = \left( \sum_{k=1}^{N_{\text{set}}} \left( \mathcal{F}[\{f^{(k)}\}] - \mathcal{F}[\{f^{(0)}\}] \right)^2 \right)^{1/2}
$$
What are neural networks?

* Each neuron receives input from neurons in preceding layer.
* Activation determined by weights and thresholds according to a non linear function:

\[ \xi_i = g\left(\sum_j \omega_{ij} \xi_j - \theta_i\right), \quad g(x) = \frac{1}{1 + e^{-x}} \]

In a simple case (1-2-1) we have,

\[ \xi_1^{(3)} = \frac{1}{1 + e^{\theta_1^{(3)} - \sum_j \omega_{1j} \xi_j^{(2)} - \theta_1^{(2)} - \sum_j \omega_{2j} \xi_j^{(1)} - \theta_2^{(2)} - \sum_j \omega_{1j} \xi_j^{(1)} - \theta_1^{(1)}}} \]

7 parameters

...Just a convenient functional form which provides a redundant and flexible parametrization.

We want the best fit to be independent of any assumption made on the parametrization.
Our fitting strategy is very different from that of normally used: instead of a set of basis functions with a small number of pars, we have an unbiased basis of functions parameterized by a very large and redundant set of pars.

\[ \mathcal{O}(20) \text{ parm} \quad \text{CTEQ, MSTW, AKP} \]

\[ \mathcal{O}(200) \text{ parm} \quad \text{NNPDF} \]

Not trivial because ...
- A redundant parametrization might accommodate also random fluctuations of statistical data.
- Very large space of parameters

Ingredients of fitting procedure
- Flexible and redundant parametrization
- Genetic Algorithm minimization
- Dynamical stopping criterion
Dynamical Stopping Criterion

* GA is monotonically decreasing by construction.
* The best fit is given by an optimal training beyond which the figure of merit improves only because we are fitting statistical noise of the data.

Cross-validation method

* Divide data in two sets: training and validation.
* Random division for each replica \( f_t = f_v = 0.5 \).
* Minimisation is performed only on the training set. The validation \( \chi^2 \) for the set is computed.
* When the training \( \chi^2 \) still decreases while the validation \( \chi^2 \) stops decreasing \( \rightarrow \) STOP.
Definition of $\chi^2$

- Fully correlated $\chi^2$:

$$
\chi^2,(k) [\omega] = \frac{1}{N_{\text{dat}}} \sum_{i,j=1}^{N_{\text{dat}}} \left( F_i^{(\text{art})(k)} - F_i^{(\text{net})(k)} \right) \left( \left( \text{COV}^{(k)} \right)^{-1} \right)_{ij} \left( F_j^{(\text{art})(k)} - F_j^{(\text{net})(k)} \right),
$$

- The covariance matrix $\text{COV}^{(k)}$ is defined from the experimental covariance matrix which does not include normalization errors.

$$
\left( \text{COV}^{(k)} \right)_{ij} = \left( \text{COV}^{(\text{exp})} \right)^{-1}_{ij} S^{(k)}_{iN} S^{(k)}_{jN}
$$

$$
S^{(k)}_{pN} = \prod_{n=1}^{N_a} \left( 1 + r^{(k)}_{p,n} \sigma_{p,n} \right) \prod_{i=1}^{N_r} \sqrt{1 + r^{(k)}_{p,i} \sigma_{p,i}}
$$

- $F_i^{(\text{net})}$ is computed from PDFs using NLO, ZM-VFN scheme.
- $\alpha_s$ kept fixed.
- $N_{\text{rep}} = 100$-$1000$ to obtain accurate description of data.
NORMALIZATION AND SUM RULES
NNPDF1.2: Normalization and Sum Rules

\[\Sigma(x, Q^2_0) = (1 - x)^{m_\Sigma} x^{-n_\Sigma} N N_\Sigma(x),\]
\[V(x, Q^2_0) = A_V (1 - x)^{m_V} x^{-n_V} N N_V(x),\]
\[T_3(x, Q^2_0) = (1 - x)^{m_{T_3}} x^{-n_{T_3}} N N_{T_3}(x),\]
\[\Delta_S(x, Q^2_0) = A_{\Delta_S} (1 - x)^{m_{\Delta_S}} x^{-n_{\Delta_S}} N N_{\Delta_S}(x),\]
\[g(x, Q^2_0) = A_g (1 - x)^{m_g} x^{-n_g} N N_g(x),\]
\[s^+(x, Q^2_0) = (1 - x)^{m_+} x^{-n_+} N N_{s^+}(x),\]
\[s^-(x, Q^2_0) = (1 - x)^{m_-} x^{-n_-} N N_{s^-}(x) - A_{s^-} [x^r_{s^-} (1 - x)^{m_{t^-}}]\]

Normalization → Fixed by valence and momentum sum rules

\[\int_0^1 dx \ x (\Sigma(x) + g(x)) = 1\]
\[\int_0^1 dx \ (u(x) - \bar{u}(x)) = 2\]
\[\int_0^1 dx \ (d(x) - \bar{d}(x)) = 1\]
\[\int_0^1 dx \ (s(x) - \bar{s}(x)) = 0\]
NNPDF1.2: Sum Rules

For instance

$$A_V = \frac{3}{\int_0^1 dx \left((1 - x)^{m_V} x^{-n_V} NN_V(x)\right)}$$

For the strange sum rule it is slightly different:

$$A_{s-} = \frac{\Gamma (r_{s-} + t_{s-} + 2)}{\Gamma (r_{s-} + 1) \Gamma (t_{s-} + 1)} \int_0^1 dx \left((1 - x)^{m_{s-}} x^{-n_{s-}} NN_{s-}(x)\right)$$

When $A_{s-} = 0$ the valence sum rule constraint is removed.
Polynomial preprocessing functions are introduced in order to speed up the training but should not affect final results.

Default values for the preprocessing exponents, $\chi^2 = 1.34$.  

<table>
<thead>
<tr>
<th></th>
<th>m</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Sigma$</td>
<td>3</td>
<td>1.2</td>
</tr>
<tr>
<td>g</td>
<td>4</td>
<td>1.2</td>
</tr>
<tr>
<td>$T_3$</td>
<td>3</td>
<td>0.3</td>
</tr>
<tr>
<td>$V$</td>
<td>3</td>
<td>0.3</td>
</tr>
<tr>
<td>$\Delta_S$</td>
<td>3</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Stability checks under variation of exponents:

<table>
<thead>
<tr>
<th>Valence sector</th>
<th>$\chi^2$</th>
<th>Singlet sector</th>
<th>$\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_{T_3} = n_V = 0.1$</td>
<td>1.38</td>
<td>$n_{\Sigma} = n_g = 0.8$</td>
<td>1.39</td>
</tr>
<tr>
<td>$n_{T_3} = n_V = 0.5$</td>
<td>1.34</td>
<td>$n_{\Sigma} = n_g = 1.6$</td>
<td>1.52</td>
</tr>
<tr>
<td>$m_{T_3} = m_V = 2$</td>
<td>1.55</td>
<td>$m_{\Sigma} = m_g - 1 = 2$</td>
<td>1.37</td>
</tr>
<tr>
<td>$m_{T_3} = m_V = 4$</td>
<td>1.28</td>
<td>$m_{\Sigma} = m_g - 1 = 4$</td>
<td>1.41</td>
</tr>
</tbody>
</table>
Stability estimator: distance between MC ensembles.

* All features of the NNPDF parton set can be assessed by using standard statistical tools.

* Distances between two probability distributions:
  Quark $\left\{ f_{ik}^{(1)} = f_{k}^{(1)}(x_i, Q_0^2) \right\}$

$$
\langle d[f] \rangle = \sqrt{\left\langle \frac{\left( \langle f_i \rangle^{(1)} - \langle f_i \rangle^{(2)} \right)^2}{\sigma^2[f_i^{(1)}] + \sigma^2[f_i^{(2)}]} \right\rangle_{\text{pts}}}
$$

* With:

$$
\langle f_i \rangle^{(1)} \equiv \frac{1}{N_{\text{rep}}^{(1)}} \sum_{k=1}^{N_{\text{rep}}^{(1)}} f_{ik}^{(1)},
$$

$$
\sigma^2[f_i^{(1)}] \equiv \frac{1}{N_{\text{rep}}^{(1)}(N_{\text{rep}}^{(1)} - 1)} \sum_{k=1}^{N_{\text{rep}}^{(1)}} \left( f_{ik}^{(1)} - \langle f_i \rangle^{(1)} \right)^2
$$

* For statistically equivalent PDF sets: $\langle d[f] \rangle \sim \langle d[\sigma_f] \rangle \sim 1$
Stability versus preprocessing exponents

<table>
<thead>
<tr>
<th>Data region</th>
<th>$n_V = 0.1$</th>
<th>$n_V = 0.5$</th>
<th>$m_V = 2$</th>
<th>$m_V = 4$</th>
<th>$n_S = 0.8$</th>
<th>$n_S = 1.6$</th>
<th>$m_S = 2$</th>
<th>$m_S = 4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Sigma(x, Q^2)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\langle d[q] \rangle$</td>
<td>1.34</td>
<td>1.25</td>
<td>1.37</td>
<td>2.14</td>
<td>1.72</td>
<td>1.38</td>
<td>1.45</td>
<td>1.64</td>
</tr>
<tr>
<td>$\langle d[\sigma] \rangle$</td>
<td>1.45</td>
<td>1.44</td>
<td>1.25</td>
<td>1.44</td>
<td>2.03</td>
<td>2.66</td>
<td>0.95</td>
<td>1.35</td>
</tr>
<tr>
<td>$g(x, Q^2)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\langle d[q] \rangle$</td>
<td>1.31</td>
<td>1.30</td>
<td>2.69</td>
<td>1.15</td>
<td>3.06</td>
<td>2.08</td>
<td>1.20</td>
<td>1.74</td>
</tr>
<tr>
<td>$\langle d[\sigma] \rangle$</td>
<td>1.34</td>
<td>1.60</td>
<td>1.56</td>
<td>1.37</td>
<td>3.21</td>
<td>2.44</td>
<td>0.98</td>
<td>1.72</td>
</tr>
<tr>
<td>$T_3(x, Q^2)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\langle d[q] \rangle$</td>
<td>1.97</td>
<td>2.48</td>
<td>8.35</td>
<td>9.74</td>
<td>1.31</td>
<td>3.23</td>
<td>1.03</td>
<td>1.41</td>
</tr>
<tr>
<td>$\langle d[\sigma] \rangle$</td>
<td>1.10</td>
<td>1.47</td>
<td>1.98</td>
<td>1.53</td>
<td>1.10</td>
<td>2.66</td>
<td>1.76</td>
<td>1.99</td>
</tr>
<tr>
<td>$V(x, Q^2)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\langle d[q] \rangle$</td>
<td>11.03</td>
<td>1.55</td>
<td>3.61</td>
<td>5.60</td>
<td>0.94</td>
<td>2.12</td>
<td>1.25</td>
<td>3.54</td>
</tr>
<tr>
<td>$\langle d[\sigma] \rangle$</td>
<td>3.57</td>
<td>4.74</td>
<td>4.04</td>
<td>3.09</td>
<td>1.03</td>
<td>1.10</td>
<td>0.66</td>
<td>1.98</td>
</tr>
</tbody>
</table>

Extrapolation

<table>
<thead>
<tr>
<th>$n_V = 0.1$</th>
<th>$n_V = 0.5$</th>
<th>$m_V = 2$</th>
<th>$m_V = 4$</th>
<th>$n_S = 0.8$</th>
<th>$n_S = 1.6$</th>
<th>$m_S = 2$</th>
<th>$m_S = 4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Sigma(x, Q^2)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\langle d[q] \rangle$</td>
<td>1.06</td>
<td>1.69</td>
<td>1.49</td>
<td>1.84</td>
<td>7.72</td>
<td>4.67</td>
<td>0.87</td>
</tr>
<tr>
<td>$\langle d[\sigma] \rangle$</td>
<td>1.12</td>
<td>1.84</td>
<td>2.11</td>
<td>1.52</td>
<td>2.47</td>
<td>3.66</td>
<td>0.82</td>
</tr>
<tr>
<td>$g(x, Q^2)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\langle d[q] \rangle$</td>
<td>1.41</td>
<td>2.32</td>
<td>2.33</td>
<td>1.34</td>
<td>1.62</td>
<td>4.73</td>
<td>1.04</td>
</tr>
<tr>
<td>$\langle d[\sigma] \rangle$</td>
<td>1.41</td>
<td>1.86</td>
<td>1.95</td>
<td>1.30</td>
<td>2.15</td>
<td>2.72</td>
<td>0.81</td>
</tr>
<tr>
<td>$T_3(x, Q^2)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\langle d[q] \rangle$</td>
<td>1.71</td>
<td>2.70</td>
<td>7.40</td>
<td>1.60</td>
<td>1.36</td>
<td>2.37</td>
<td>0.78</td>
</tr>
<tr>
<td>$\langle d[\sigma] \rangle$</td>
<td>4.83</td>
<td>4.54</td>
<td>2.89</td>
<td>5.09</td>
<td>1.00</td>
<td>1.65</td>
<td>0.92</td>
</tr>
<tr>
<td>$V(x, Q^2)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\langle d[q] \rangle$</td>
<td>14.85</td>
<td>3.23</td>
<td>3.75</td>
<td>2.55</td>
<td>0.86</td>
<td>2.52</td>
<td>1.26</td>
</tr>
<tr>
<td>$\langle d[\sigma] \rangle$</td>
<td>2.65</td>
<td>5.08</td>
<td>3.94</td>
<td>2.78</td>
<td>1.20</td>
<td>0.87</td>
<td>0.62</td>
</tr>
</tbody>
</table>
NNPDF1.2: Randomized preprocessing

- Remarkable stability: in most cases variations are within 90% C.L.
- Exception given by valence and triplet: deviation $\sim 1.4\sigma$ from central value when varying exponents.
- Uncertainty on $V$ and $T_3$ underestimated by factor between 1 and 2.
- Note that we have full control on that!

NNPDF1.2: Randomized preprocessing!

- Bigger uncertainty on $\bar{u}$ and $u_v$! Will be reduced by DY data.
BENCHMARK PARTONS
HERA-LHC benchmark


<table>
<thead>
<tr>
<th>Set</th>
<th>$N_{\text{dat}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCDMSp</td>
<td>322</td>
</tr>
<tr>
<td>NMC</td>
<td>95</td>
</tr>
<tr>
<td>NMC-pd</td>
<td>73</td>
</tr>
<tr>
<td>Z97NC</td>
<td>206</td>
</tr>
<tr>
<td>H197low$Q^2$</td>
<td>77</td>
</tr>
</tbody>
</table>

3163 data $\rightarrow$ 773 data

$Q^2 > 9$ GeV$^2$

$W^2 > 15$ GeV$^2$
Comparison between collaborations and between benchmark/global partons.

$u(x, Q^2 = 2\text{GeV}^2)$: Data region
HERA-LHC benchmark

Comparison between collaborations and between benchmark/global partons.

$u(x, Q^2 = 2 \text{GeV}^2)$: Extrapolation Region
Dependence on data sets

HERA-LHC benchmark

- MRST01: benchmark partons and global partons do not agree within error!
- Input parametrization, flavor assumptions and statistical treatment ($\Delta \chi^2_{\text{global}} = 50$, $\Delta \chi^2_{\text{bench}} = 1$) are tuned to data.
- MSTW08: Bigger parametrization of PDFs and increased T($>1$) for benchmark fit, evaluated with dynamic tolerance criterion.
- Better compatibility, still problems with u,d valence and gluon

- NNPDF1.0 is consistent with MRST01 global fit.
- NNPDFbench is consistent with NNPDF1.0 and MRST01.
- Same parametrization and flavour assumption.
- Same statistical treatment.
- Underestimation of the error in the standard approach.
NNPDF2.0
NNPDF2.0: Predictions from previous fits

- Predictions evaluated with NNPDF1.0 error sets.
- Large error bands on predictions, compatible with data.
NNPDF2.0: FastNLO-like Drell-Yan

Schematically, take

\[ \sum_{i,j=1}^{N_q} a_{ij} \int_{x_1}^{1} dy_1 \int_{x_2}^{1} dy_2 f_i(y_1)f_j(y_2) C^{ij}(y_1, y_2) \]

Define an interpolation grid:

\[ x_{\text{min}} \equiv x_1 < x_2 < \ldots < x_{N_x-1} < x_{N_x} \equiv 1 \]

Around each of these grid points let us define a set of interpolating functions \( \mathcal{I}^{(\alpha)} \):

\[ \mathcal{I}^{(\alpha)}(x_{\alpha}) = 1 \quad \mathcal{I}^{(\alpha)}(x_{\beta}) = 0 \quad \sum_{\alpha=1}^{N_x} \mathcal{I}^{(\alpha)}(y)=1, \forall y \]

\[ f_j(y) = \sum_{\alpha=1}^{N_x} f_j(x_{\alpha}) \mathcal{I}^{(\alpha)}(y) \]

Therefore, like in FastNLO we get:

\[ \sum_{i,j=1}^{N_q} a_{ij} \sum_{\alpha, \beta=1}^{N_x} f_i(x^\alpha_1)f_j(x^\beta_2) \int_{x_1}^{1} dy_1 \int_{x_2}^{1} dy_2 \mathcal{I}^{(\alpha)}(y_1)\mathcal{I}^{(\beta)}(y_2) C^{QQ}(y_1, y_2) = \sum_{\alpha, \beta=1}^{N_x} f_i(x^\alpha_1)f_j(x^\beta_2)c_{ij\alpha\beta}(x_1, x_2) \]

with \( c_{ij\alpha\beta} \) and evolution precomputed (analogously).