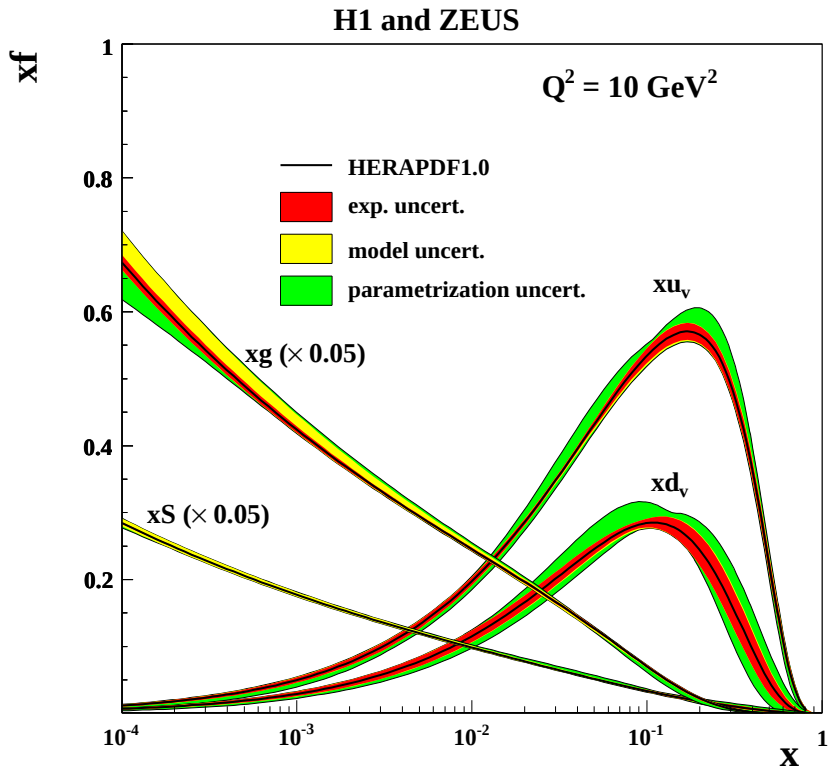
Uncertainty treatment in PDFs using HERAFitter

S. Glazov, PDF school, DESY, 23 Oct 2012

Outline

- Introduction
- Setup for the tutorial
- Experimental uncertainties
- Propagation of experimental uncertainties
- Theoretical (model) uncertainties
- Parametrization uncertainties

Introduction



HERAPDF fits determine valence, sea and gluon PDFs. The sources of uncertainty include **experimental** accuracy, **model** or theoretical uncertainties as well as uncertainties related to the choice of the **parametrization** form for PDFs. A way to estimate these uncertainties and propagate to PDFs is the subject of the lecture.

Setup for the tutorial

For the today's lecture-tutorial, we need a working version of the HERAFitter beta-2 release. The code should be compiled with the APPLGRID option:

```
./configure --enable-applgrid ;  make clean ; make -j 4 install
```

At tar file with a set of sample steering files is attached to the indico page. You need to unpack it in the HERAFitter directory

```
wget http://www.desy.de/~glazov/pdf_uncertainties.tar.gz
tar xfz pdf_uncertainties.tar.gz
```

We shall start with the "base" configuration:

```
cp steering.txt-base steering.txt
cp minuit.in.txt-base minuit.in.txt
cp ewparam.txt-base ewparam.txt
cp datafiles/hera/H1ZEUS_NC_e+p_HERA1.0.dat-base \
   datafiles/hera/H1ZEUS_NC_e+p_HERA1.0.dat
```

The base configuration will fit using HERA-I data, using 'RT FAST' treatment of heavy flavors and few tricks to make evolution run faster (smaller x grid and reading QCDNUM tables from previous run, see QCDNUM namelist).

Modeling experimental uncertainties: basics

The data are reported with their central values μ_i and experimental uncertainties Δ_i. The simplest situation is when the errors are Gaussian and there are no correlations among different data points. In this case finding an optimal PDF set corresponds to a χ^2 minimization:

$$\chi^2(p_k) = \sum_i \left(\frac{m_i(p_k) - \mu_i}{\Delta_i} \right)^2 .$$

Here p_k are parameters describing PDFs and $m_i(p_k)$ are predictions based on these parameters. The sum runs over all data points.

The "base" configuration uses uncorrelated treatment of the HERA data. We shall store it as a reference:

```
./bin/FitPDF
cp -r output output_base
```

Please check `output_base/Results.txt` which reports χ^2 values for experiments. Total χ^2 should be equal to 576.16.

Correlated experimental uncertainties

The measurements are, however, correlated with each other. Typically bin-to-bin correlations due to systematic uncertainties are more important than statistical correlations, which can be neglected. In this case a convenient way to represent these correlations is by using nuisance parameters.

The influence of correlated uncertainty sources j on data points i can be described by a matrix Γ_{ij} such that if a source moves up by 1σ all data points move by Γ_{ij}:

$$\mu_i \rightarrow \mu_i + \Gamma_{ij}$$

The 1σ shift of the systematic uncertainty should cause increase of the χ^2 by 1 unit which can be achieved by adding a penalty term. This defines the χ^2 for the data to theory comparison, using the nuisance parameters for the systematic sources:

$$\chi^2(p_k, b_j) = \sum_i \left(\frac{m_i(p_k) - \mu_i - \sum_j \Gamma_{ij} b_j}{\Delta_i} \right)^2 + \sum_j b_j^2$$

Here b_j are additional nuisance parameters representing shifts of data due to systematic errors.

Adding correlated systematics for HERA data

We can now include correlation information for HERA data. For illustration, we shall do that for $e + p$ data only. In principle since the data are correlated across the data sets, this is not correct. You can try to do it more consistently, for all sets, as an exercise.

Prepare the data file:

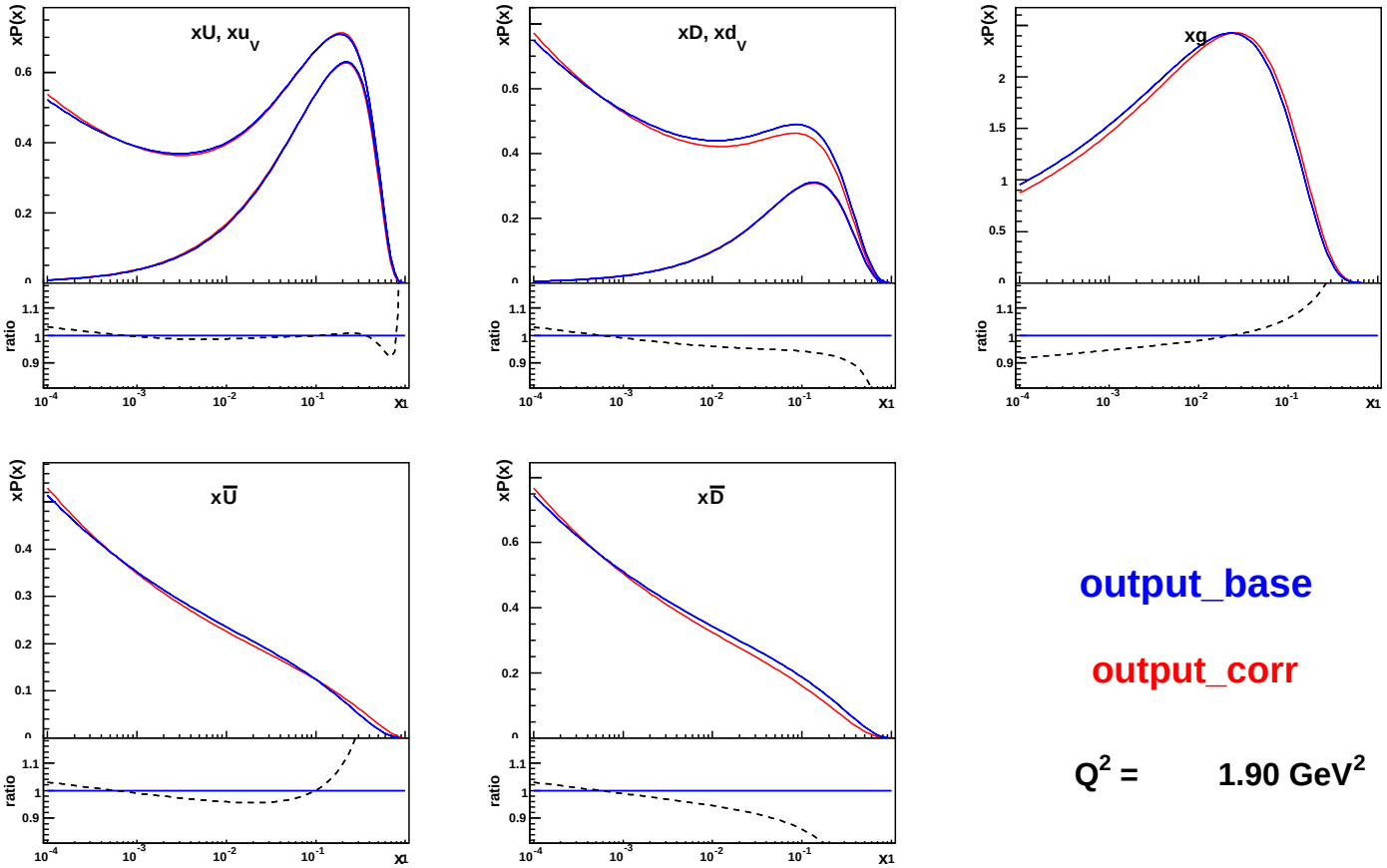
```
cp datafiles/hera/H1ZEUS_NC_e+p_HERA1.0.dat-corr \
   datafiles/hera/H1ZEUS_NC_e+p_HERA1.0.dat
```

(check DIFF between CORR and BASE file).

And run:

```
./bin/FitPDF
cp -r output output_corr
./bin/DrawResults output_corr output_base
. ./restore.cmd
gv DrawResults.ps
```

Adding correlation to HERA data: PDFs



Some changes in PDFs when correlations are added. As we will see in the moment, they are within uncertainties. Other changes can be observed by checking OUTPUT/RESULTS.TXT file which lists shifts of nuisance parameters.

Propagation of experimental uncertainties

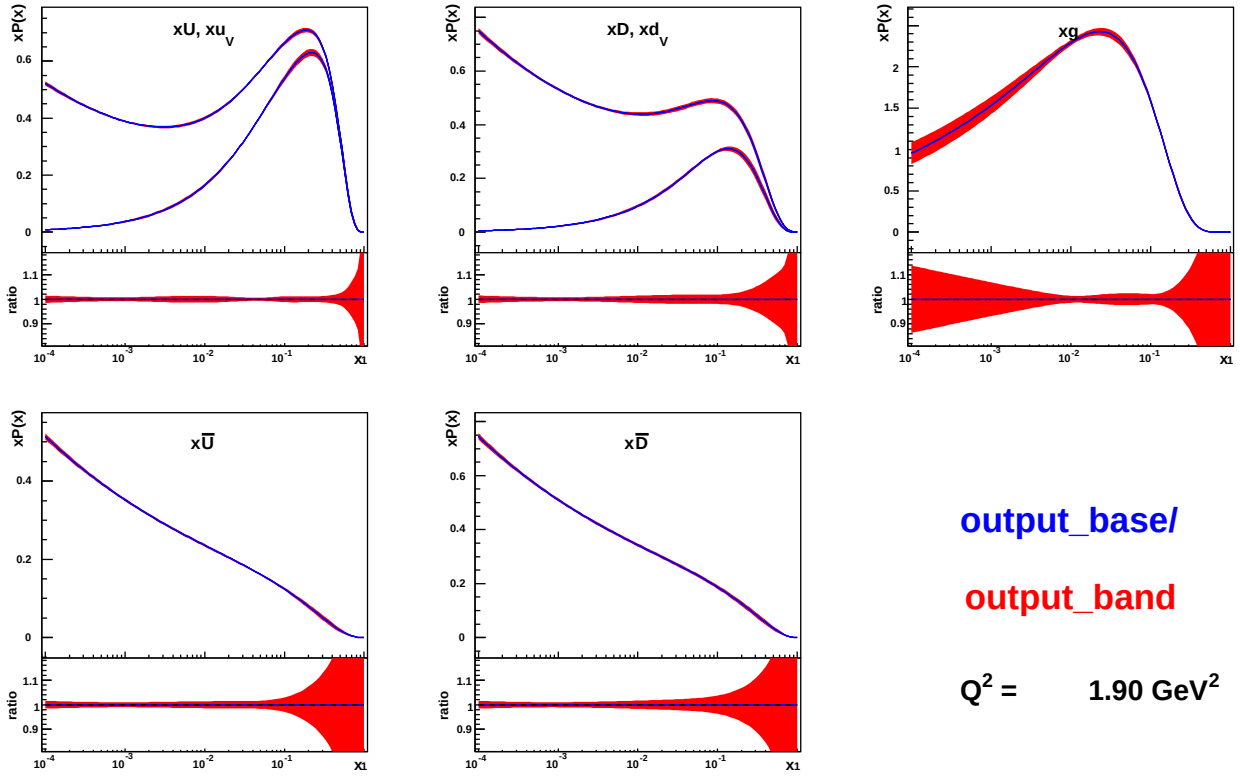
Inspecting the MINUIT output file, OUTPUT_BASE/MINUIT.OUT.TXT, one can observe that the parameters used to describe PDFs are often highly correlated with each other.

```
 PARAMETER  CORRELATION COEFFICIENTS
       NO.  GLOBAL     2     3    12    13    15    23    33    41    42    43
        2  0.99688  1.000 0.893 0.396 0.069-0.331 0.455-0.133-0.596-0.692-0.503
        3  0.99893  0.893 1.000 0.247 0.191-0.252 0.685-0.233-0.558-0.587-0.776
       12  0.99835  0.396 0.247 1.000-0.593-0.795-0.264 0.759 0.316 0.187 0.291
......
```

To estimate PDF uncertainties, ideally one wants to have uncorrelated PDF sets. This can be achieved by diagonalisation procedure. A robust way of doing this, taking into account potentially asymmetric shape of the χ^2 function around the minimum is introduced by J. Pumplin, Phys.Rev.D65:014011,2001. This is activated in `HERAFitter` by choosing DoBANDS = TRUE. Let's try this for HERA data:

```
cp steering.txt-band steering.txt
./bin/FitPDF
cp -r output output_band
./bin/DrawResults --bands output_band output_base
gv DrawResults.ps
```

Results of the “bands” output



output_base/

output_band

Q^2 = 1.90 GeV^2

The bands indicate the experimental uncertainties. The fit produced a set of LHAPDF files which can be also used to propagate them to predictions.

You can also produce root graphs if you need better plotting styles, use BIN/STOREGRAPHS –BANDS OUTPUT_BAND which produces OUTPUT_BAND/GRAPHS.ROOT file. Check TOOLS/ROOT/PLOT_PDF.C as well.

MC method

Alternative way to propagate experimental uncertainties is to use Toy MC method. In this method, input data are allowed to fluctuate according to their experimental and systematic uncertainties. For example, for Gaussian uncertainties, random numbers are prepared for all data points and systematic error sources and the data are modified as

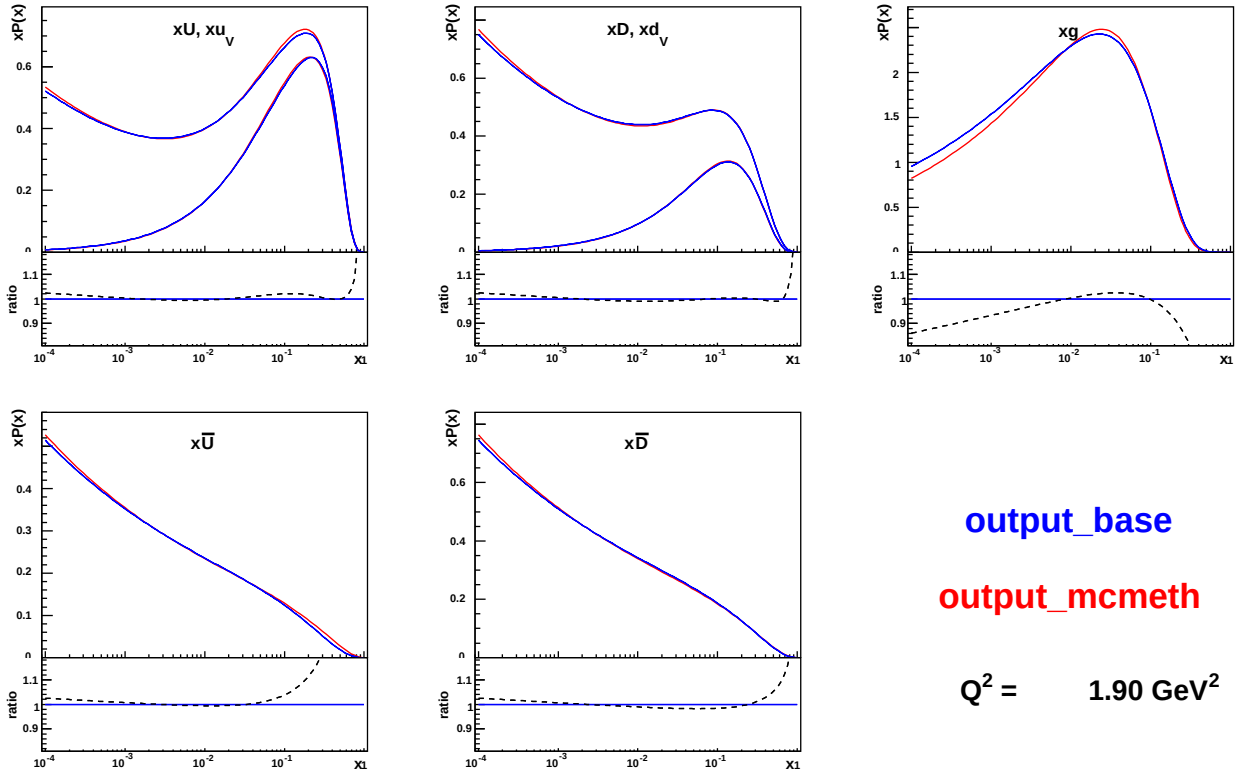
$$\mu_i \to \mu_i(1 + r_i\Delta_i + \sum_j r_j\gamma_{ij}) \,,$$

where $\gamma_{ij} = \Gamma_{ij}/\mu_i$ and r_i are random Gaussian numbers.

In HERAFitter, Toy MC method is activated by setting lRAND = True. The random number sequence is controlled by ISeedMC. Systematic errors can be treated as Gaussian, uniform and LogNormal.

```
. ./restore.cmd
cp steering.txt-mcmeth steering.txt
./bin/FitPDF
cp -r output output_mcmeth
./bin/DrawResults output_mcmeth output_base
gv DrawResults.ps
```

MC replica vs central fit



To fully try MC method one needs to run at least ~ 100 replica (on a farm) and use mean/rms over them. To generate root histograms, one can use BIN/STOREGRAPHSMC DIR1 ... DIRN command.

Please note that χ^2 values at the minimum should increase for the MC method by about N_{data} units. Compare OUTPUT_BASE/RESULTS.TXT and OUTPUT_MCMETH/RESULTS.TXT

Bias corrections

It is important to make sure that fits to the data are not biased. Biases may occur from non-Gaussian distribution of the data uncertainties, e.g. normalization errors are typically multiplicative and follow log-normal rather than Gaussian distribution.

Bias corrections are built in the `HERAFitter`. The bias control is steered by CHI2STYLE parameter. In all cases correlated systematic uncertainties are assumed to be multiplicative. For H12000 style, uncorrelated uncertainties are Gaussian; for HERAPDF style correction is applied for Poisson-like statistical errors and multiplicative uncorrelated systematic uncertainties:

$$\Delta_i^2(m_i) = \Delta_{i,\mathrm{stat}}^2 \frac{m_i}{\mu_i} + \Delta_{i,\mathrm{uncor}}^2 \left(\frac{m_i}{\mu_i}\right)^2 .$$

Finally, H12011 style includes $+2\log\frac{\Delta_i(m_i)}{\Delta_i(\mu_i)}$ correction which compensates a bias due to modified denominator in χ^2 formula.

In any case it is important to make sure that the fit gives unbiased representation of the data.

Fit to Predictions and Closure test

Basic tests allowing to study biases are included in the `HERAFitter`. The idea is to perform fits to the theory expectation, using uncertainties as obtained from the data. This can turned on by setting LRANDDATA = FALSE flag in the MCERRORS namelist. Typically it is more interesting to use Poisson uncertainties by setting STATYPE = 4. This option works the best if the data uncertainties are estimated based on $\sqrt{N_{\text{data}}}$ approximation, however should work fine for more complex unfoldig too.

If you have more complex data model, the best is to prepare dedicated input data files. This is not complicated, provided the data files are plain text files.

Let's try the built in fitting machinery:

```
. ./restore.cmd
cp steering.txt-mcmeth_gen steering.txt
./bin/FitPDF
cp -r output output_mcmeth_gen
cat output_mcmeth_gen/Results.txt
```

Results of the fit to the prediction

Here is the content of the RESULTS.TXT file:

```
 First iteration    594.45560150274389              582   1.0214013771524808
After minimisation     581.23   582     0.999

  Partial chi2s
Dataset    1    144.12   145  NC cross section HERA-I H1-ZEUS combined e-p.
Dataset    2    375.53   379  NC cross section HERA-I H1-ZEUS combined e+p.
Dataset    3     23.51    34  CC cross section HERA-I H1-ZEUS combined e-p.
Dataset    4     38.07    34  CC cross section HERA-I H1-ZEUS combined e+p.
```

As expected, χ^2 value is close to N_{DF}. For a systematic study of biases, one needs to perform runs for at least $\sim$ 100 Toy MC replica.

Model uncertainties

Model as well as theory uncertainties can be studied by varying model parameters within their uncertainties. For example, HERAPDF1.0 fit considers the following variations:

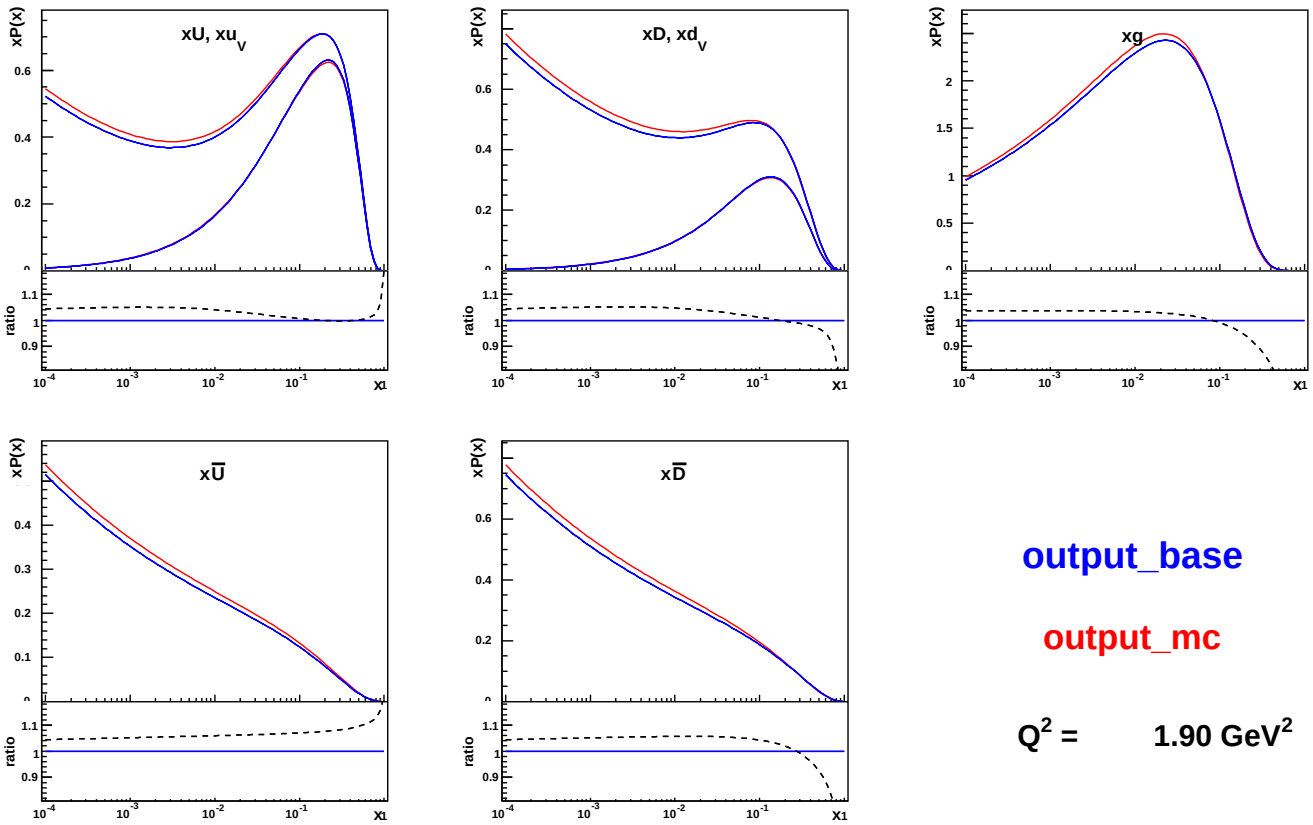
Variation	Standard Value	Lower Limit	Upper Limit
f_s	0.31	0.23	0.38
m_c [GeV]	1.4	1.35 ($Q_0^2 = 1.8$)	1.65
m_b [GeV]	4.75	4.3	5.0
Q^2_{min} [GeV^2]	3.5	2.5	5.0
Q_0^2 [GeV^2]	1.9	1.5 ($f_s = 0.29$)	2.5 ($m_c = 1.6$, $f_s = 0.34$)

where f_s is the strangeness fraction ($x\bar{s} = f_s x\bar{D}$ and $x\bar{d} = (1 - f_s)x\bar{D}$), m_c, m_b are c, b-quark masses, Q^2_{min} is the minimal value accepted for the data and Q_0^2 is the evolution starting scale.

Let's see the effect of the charm mass variation to $m_c = 1.65$ GeV:

```
. ./restore.cmd
cp ewparam.txt-mc ewparam.txt
./bin/FitPDF
cp -r output output_mc
./bin/DrawResults output_mc output_base
```

Effect of the charm mass variation



output_base

output_mc

Q^2 = 1.90 GeV^2

Increase of m_c leads to suppression of charm density and thus increase of u and d densities, to compensate for this. Note also that initial χ^2 for OUTPUT_MC/RESULTS.TXT is pretty bad while χ^2 after minimization is close to the nominal fit.

Scale uncertainty

To study effects of the scale uncertainty, we will use a reference which includes APPLGRID based predictions. A good example are ATLAS inclusive jets with $R = 0.6$, publication based on 2010 data (Phys.Rev. D86 (2012) 014022).

```
. ./restore.cmd
cp steering.txt-base-jets steering.txt
cp minuit.in.txt-jets minuit.in.txt
./bin/FitPDF
cp -r output output_jets
```

The scales are controlled by SCALES namelist, variables DATASETMUR and DATASETMUF. For example, we can set $\mu_F = 2$ for all ATLAS samples:

```
cp steering.txt-scale-jets steering.txt
./bin/FitPDF
cp -r output output_jets_scale
```

Effect of the scale variation

Since we did not run full fit, the effect of the scale variation can be checked by comparing χ^2 values at the first iteration and also checking the theory predictions, stored in OUTPUT/FITTEDRESULTS.TXT files. These are the last 8 lines for the two files:

TAIL -8 OUTPUT_JETS_SCALE/FITTEDRESULTS.TXT:

```
ATLAS Jet data 3.6 <= |y| < 4.4
pt1         pt2         YBinSize       data      +- uncor     +- tot     th orig         th mod     pull    iset
0.25000E+02 0.25000E+02 0.16000E+01 0.15080E+07 0.10651E+06 0.10143E+07 0.16111E+07 0.14886E+07 0.18198E+00   33
0.37500E+02 0.37500E+02 0.16000E+01 0.13848E+06 0.17482E+05 0.97680E+05 0.15061E+06 0.13042E+06 0.46064E+00   33
0.52500E+02 0.52500E+02 0.16000E+01 0.12918E+05 0.10312E+04 0.56777E+04 0.14968E+05 0.12649E+05 0.26140E+00   33
0.70000E+02 0.70000E+02 0.16000E+01 0.12747E+04 0.10305E+03 0.41566E+03 0.17089E+04 0.13310E+04 -.54593E+00   33
0.95000E+02 0.95000E+02 0.16000E+01 0.76973E+02 0.82460E+01 0.22815E+02 0.10751E+03 0.78783E+02 -.21949E+00   33
0.13500E+03 0.13500E+03 0.16000E+01 0.82810E+00 0.18516E+00 0.33739E+00 0.15060E+01 0.87098E+00 -.23156E+00   33
```

TAIL -8 OUTPUT_JETS/FITTEDRESULTS.TXT:

```
ATLAS Jet data 3.6 <= |y| < 4.4
pt1         pt2         YBinSize       data      +- uncor     +- tot     th orig         th mod     pull    iset
0.25000E+02 0.25000E+02 0.16000E+01 0.15080E+07 0.11563E+06 0.10143E+07 0.17501E+07 0.14938E+07 0.12282E+00   33
0.37500E+02 0.37500E+02 0.16000E+01 0.13848E+06 0.18630E+05 0.97680E+05 0.16052E+06 0.12956E+06 0.47848E+00   33
0.52500E+02 0.52500E+02 0.16000E+01 0.12918E+05 0.10844E+04 0.56777E+04 0.15772E+05 0.12678E+05 0.22187E+00   33
0.70000E+02 0.70000E+02 0.16000E+01 0.12747E+04 0.10725E+03 0.41566E+03 0.17822E+04 0.13354E+04 -.56572E+00   33
0.95000E+02 0.95000E+02 0.16000E+01 0.76973E+02 0.83946E+01 0.22815E+02 0.10957E+03 0.78521E+02 -.18438E+00   33
0.13500E+03 0.13500E+03 0.16000E+01 0.82810E+00 0.18104E+00 0.33739E+00 0.14601E+01 0.85068E+00 -.12472E+00   33
```

See changes in `th orig`.

Parametrization uncertainty

The usual parameterisation of PDFs

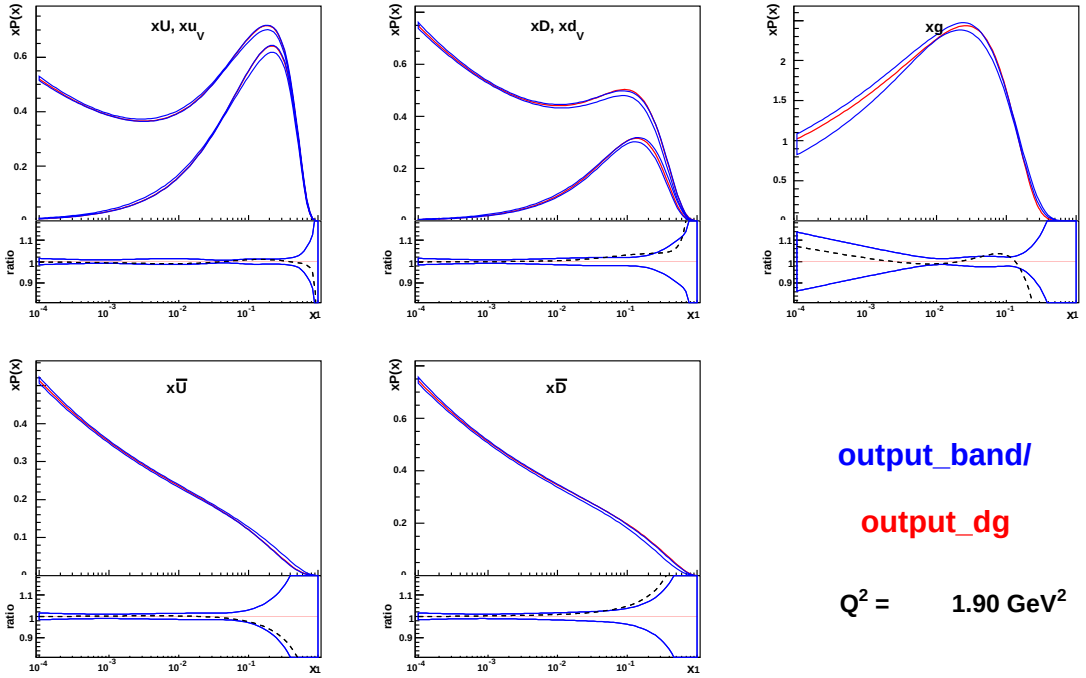
$$x f(x) = A x^B (1-x)^C \left(1 + Dx + Ex^2 + \ldots\right)$$

has a nice feature that it allows to describe PDFs with just a few parameters. The main ones are normalisation A, low x power $-1 < B$ ($B > 0$ for valence-like shapes and $B < 0$ for the sea) and suppression parameter for $x \to 0$, $C > 0$.

The parameters D, E are used to interpolate PDF behaviour at intermediate x. How many of them should be used to describe the data is, however, not well defind problem which requires requlirastion.

A simple regularisation prescription is given by HERAPDF. One starts with the basic form ($D, E, \ldots = 0$) and adds one parameter at a time. They are used for the central fit if the χ^2 is improved significantly. If the χ^2 is not improved significantly anymore, central value is fixed. Variations which improve χ^2 by more than ~ 1 unit are used to estimate parameterisation uncertainty by building envelope of PDF variations.

Effect of parametrization uncertainties



output_band/

output_dg

Q^2 = 1.90 GeV^2

Try opening *D* parameter for the gluon:

```
. ./restore.cmd
cp minuit.in.txt-Dg minuit.in.txt
./bin/FitPDF; cp -r output output_dg
./bin/DrawResults --bands output_dg output_band/
gv DrawResults.ps
```

χ^2 is decreased by $\sim$ 1.5 units, some impact for gluon density.

Recap

- Experimental uncertainties can be estimated using Hessian or Monte Carlo method. Correlation information for the experimental data are intruduced using nuisance parameters
- χ^2 fits may be biased because of non-Gaussian behavior of the uncertainties. The biases can be corrected using various methods. It is important, however, to perform closure tests, which can be run using toy MC method.
- Model uncertainties can be estimated by varying input parameters. A couple of variations are available in `HERAFitter` using QCDNUM, various heavy flavour models as well as APPLGRID and FASTNLO.
- Parameterisation uncertainties can be studied using `HERAPDF` approach: adding one extra parameter at a time and building an envelope over the variations.

Features for the next Release

- More complex data model:
  - Covariance matrix
  - Alternative minimization
  - More diverse bias correction procedure: e.g. individual systematic errors can be declared “additive”.
  - Toy MC with asymmetric errors.
- New methods for parameterisation uncertainty estimates:
  - Flexible parameterisation with explicit penalty for deviation from the “basic form”.
  - NNPDF-like overtraining protection method.

→ Stay tuned !