Sources of PDF uncertainty

Jon Pumplin

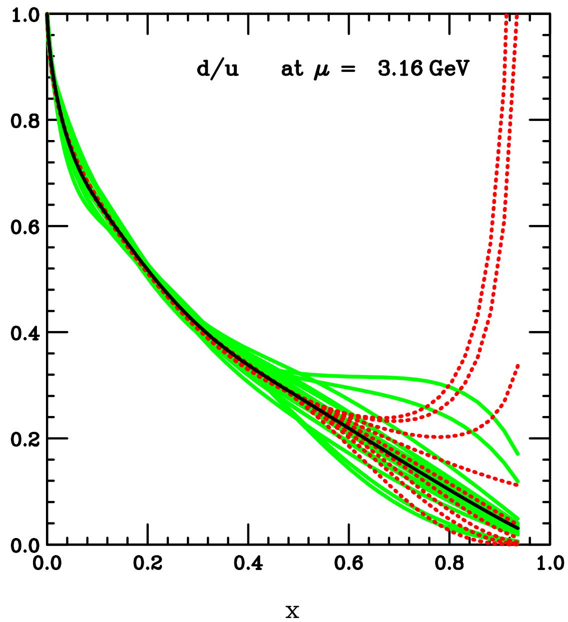
CTEQ meeting at Northwestern (20–21 November 2009)

1. Consistency between data sets used in PDF fitting (arXiv:0909.0268,arXiv:0904.2425)

2. Uncertainties caused by parametrization choices (arXiv:0909.5176: heavily revised version in progress)

Parametrization dependence: Uncertainty of $d(x)/u(x)$ at large x

((Slide from a CTEQ6.5–era talk that shows large parametrization uncertainty of the kind discussed in Thia's talk.))



Black: CTEQ6.5 central fit
Green: 40 CTEQ6.5 eigenvector uncertainty sets
Red: results from equally-acceptable alternative parametrizations

In CTEQ6.5, we assumed $d_v(x) \sim (1-x)^{a_d}$ and $u_v(x) \sim (1-x)^{a_u}$ at $x \to 1$, with constraint $a_d - a_u = +1$. This constraint was imposed (for the best fit and for all eigenvector sets) because $a_d - a_u$ is very weakly constrained by χ^2 ("flat direction")

Red dotted curves are fits made with a variety of choices for $a_d - a_u$. They are all very good fits, so the behavior of d/u is completely unconstrained by the experiments included here for $x > 0.8$.

Measuring internal consistency of the fit

Partition the data into two subsets:

$$\chi^2 = \chi^2_S + \chi^2_{\bar{S}}$$

where subset S can, for example, be chosen as

- any single experiment (reported here)
- all of the jet experiments
- all of the low-Q data points (to look for higher twist)
- all of the low-x data points (to look for BFKL)
- all experiments with deuteron corrections
- all of the neutrino experiments (to look for nuclear corrections)

A method I call Data Set Diagonalization which was first proposed in my HERA/LHC talk in March 2004 directly answers the questions

1. What does subset S measure?
2. Is subset S consistent with the rest of the data?

Data Set Diagonalization

The DSD method is an extension of the Hessian method. It works by transforming the contributions χ_S^2 and $\chi_{\bar{S}}^2$ to χ^2 into a form where they can be interpreted as independent measurements of N quantities.

The essential point is that the linear transformation that leads to

$$\chi^2 = \chi_0^2 + \sum_{i=1}^{N} z_i^2$$

is not unique, because any further orthogonal transform of the z_i will preserve it. Such an orthogonal transformation can be defined using the eigenvectors of any symmetric matrix. After this second linear transformation of the coordinates, the chosen symmetric matrix will then be diagonal in the resulting new coordinates.

This freedom is exploited in the DSD method by taking the symmetric matrix from the quadratic form that describes the contribution to χ^2 from the subset S of the data that is chosen for study. Then . . .

DSD method – continued

$$
\begin{aligned}
\chi^2 &= \chi_S^2 + \chi_{\overline{S}}^2 + \text{const} \\
\chi_S^2 &= \sum_{i=1}^{N} [(z_i - A_i)/B_i]^2 \\
\chi_{\overline{S}}^2 &= \sum_{i=1}^{N} [(z_i - C_i)/D_i]^2
\end{aligned}
$$

Thus the subset S of the data and its complement $\overline{S}$

take the form of independent measurements of the N variables z_i, with results

$$
\begin{aligned}
\mathbf{S}: \; z_i &= A_i \pm B_i \\
\overline{S}: \; z_i &= C_i \pm D_i
\end{aligned}
$$

DSD method – continued

$$\begin{aligned}
\chi^2 &= \chi^2_S + \chi^2_{\overline{S}} + \text{const} \\
\chi^2_S &= \sum_{i=1}^{N} [(z_i - A_i)/B_i]^2 \\
\chi^2_{\overline{S}} &= \sum_{i=1}^{N} [(z_i - C_i)/D_i]^2
\end{aligned}$$

This decomposition answers the question "What is

measured by data subset S?" — it is those parameters z_i for which the $B_i \lesssim D_i$. The fraction of the measurement of z_i contributed by S is

$$\gamma_i = \frac{D_i^2}{B_i^2 + D_i^2}\,.$$

The decomposition also measures the compatibility between S and the rest of the data $\overline{S}$: the disagreement between the two is

$$\sigma_i = \frac{|A_i - C_i|}{\sqrt{(B_i^2 + C_i^2)}}\,.$$

Experiments that provide at least one measurement with $\gamma_i > 0.1$

Process	Expt	N	$\sum_i \gamma_i$
$e^+ p \to e^+ X$	H1 NC	115	2.10
$e^- p \to e^- X$	H1 NC	126	0.30
$e^+ p \to e^+ X$	H1 NC	147	0.37
$e^+ p \to e^+ X$	H1 CC	25	0.24
$e^- p \to \nu X$	H1 CC	28	0.13
$e^+ p \to e^+ X$	ZEUS NC	227	1.69
$e^+ p \to e^+ X$	ZEUS NC	90	0.36
$e^+ p \to \nu X$	ZEUS CC	29	0.55
$e^+ p \to \bar{\nu} X$	ZEUS CC	30	0.32
$e^- p \to \nu X$	ZEUS CC	26	0.12
$\mu p \to \mu X$	BCDMS F_2p	339	2.21
$\mu d \to \mu X$	BCDMS F_2d	251	0.90
$\mu p \to \mu X$	NMC F_2p	201	0.49
$\mu p/d \to \mu X$	NMC F_2p/d	123	2.17
$p\,\mathrm{Cu} \to \mu^+\mu^- X$	E605	119	1.52
$pp, pd \to \mu^+\mu^- X$	E866 pp/pd	15	1.92
$pp \to \mu^+\mu^- X$	E866 pp	184	1.52
$\bar{p}p \to (W \to \ell\nu)X$	CDF I Wasy	11	0.91
$\bar{p}p \to (W \to \ell\nu)X$	CDF II Wasy	11	0.16
$\bar{p}\,p \to \mathrm{jet}\,X$	CDF II Jet	72	0.92
$\bar{p}\,p \to \mathrm{jet}\,X$	D0 II Jet	110	0.68
$\nu\, Fe \to \mu X$	NuTeV F_2	69	0.84
$\nu\, Fe \to \mu X$	NuTeV F_3	86	0.61
$\nu\, Fe \to \mu X$	CDHSW	96	0.13
$\nu\, Fe \to \mu X$	CDHSW	85	0.11
$\nu\,\mathrm{Fe} \to \mu^+\mu^-\mathrm{X}$	NuTeV	38	0.68
$\bar{\nu}\,\mathrm{Fe} \to \mu^+\mu^-\mathrm{X}$	NuTeV	33	0.56
$\nu\,\mathrm{Fe} \to \mu^+\mu^-\mathrm{X}$	CCFR	40	0.41
$\bar{\nu}\,\mathrm{Fe} \to \mu^+\mu^-\mathrm{X}$	CCFR	38	0.14

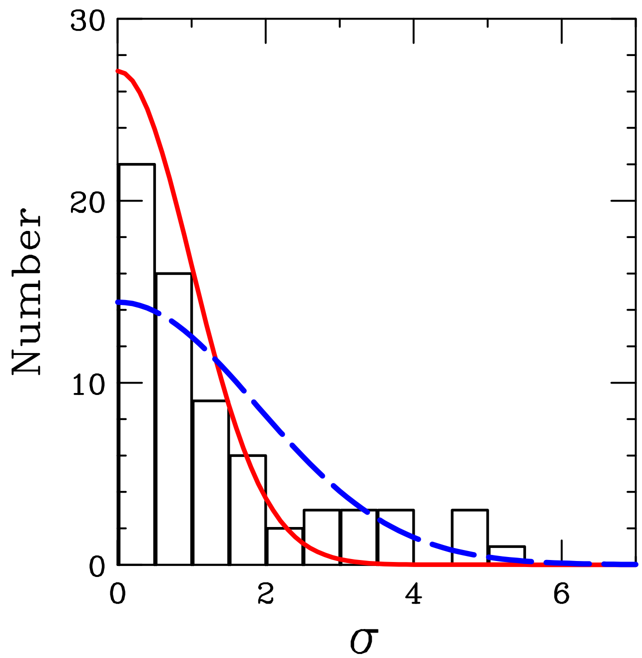
Total of $\sum \gamma_i = 23$ is close to the actual number of fit parameters.

H1+ZEUS measure 6.2 of the parameters — fewer than in HERA-only fits as expected.

Consistency tests: measurements that conflict strongly with the other experiments ($\sigma_i > 3$) are shown in red.

Expt	$\sum_i \gamma_i$	$(\gamma_1, \sigma_1), (\gamma_2, \sigma_2), \ldots$
H1 NC	2.10	(0.72, 0.01) (0.59, 3.02) (0.43, 0.20) (0.36, 1.37)
H1 NC	0.30	(0.30, 0.02)
H1 NC	0.37	(0.21, 0.06) (0.16, 0.83)
H1 CC	0.24	(0.24, 0.00)
H1 CC	0.13	(0.13, 0.00)
ZEUS NC	1.69	(0.45, 3.13) (0.42, 0.32) (0.35, 3.20) (0.29, 0.80) (0.18, 0.64)
ZEUS NC	0.36	(0.22, 0.01) (0.14, 1.61)
ZEUS CC	0.55	(0.55, 0.04)
ZEUS CC	0.32	(0.32, 0.10)
ZEUS CC	0.12	(0.12, 0.02)
BCDMS F_2p	2.21	(0.68, 0.50) (0.63, 1.63) (0.43, 0.80) (0.34, 4.93) (0.13, 0.94)
BCDMS F_2d	0.90	(0.32, 0.67) (0.24, 2.49) (0.19, 2.09) (0.16, 5.22)
NMC F_2p	0.49	(0.20, 4.56) (0.17, 4.76) (0.12, 0.50)
NMC F_2p/d	2.17	(0.61, 1.11) (0.56, 3.60) (0.43, 0.90) (0.36, 0.79) (0.21, 1.41)
E605 DY	1.52	(0.91, 1.29) (0.38, 1.12) (0.23, 0.31)
E866 pp/pd	1.92	(0.88, 0.57) (0.69, 1.15) (0.35, 1.80)
E866 pp	1.52	(0.75, 0.04) (0.39, 1.79) (0.23, 1.94) (0.14, 3.57)
CDF Wasy	0.91	(0.57, 0.33) (0.34, 0.51)
CDF Wasy	0.16	(0.16, 2.84)
CDF Jet	0.92	(0.48, 0.47) (0.44, 3.86)
D0 Jet	0.68	(0.39, 1.70) (0.29, 0.76)
NuTeV F_2	0.84	(0.37, 2.75) (0.29, 0.42) (0.18, 0.97)
NuTeV F_3	0.61	(0.30, 0.50) (0.16, 1.35) (0.15, 0.30)
CDHSW	0.13	(0.13, 0.04)
CDHSW	0.11	(0.11, 1.32)
NuTeV	0.68	(0.39, 0.31) (0.29, 0.66)
NuTeV	0.56	(0.32, 0.18) (0.24, 2.56)
CCFR	0.41	(0.24, 1.37) (0.17, 0.12)
CCFR	0.14	(0.14, 0.79)

Consistency of measurements in a global fit



Histogram of the consistency measure σ_i for the 68 significant ($\gamma_i > 0.1$) measurements provided by the 37 experiments in a typical global fit.

Solid curve is the absolute Gaussian prediction

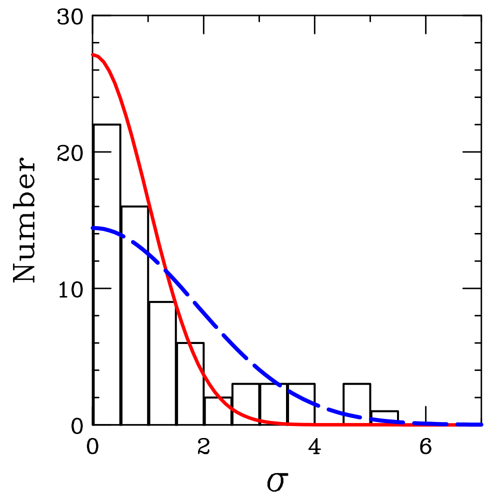
$$\frac{dP}{d\sigma} = \sqrt{\frac{2}{\pi}} \exp(-\sigma^2/2)\ .$$

Dashed curve is a scaled Gaussian with $c = 1.9$:

$$\frac{dP}{d\sigma} = \sqrt{\frac{2}{\pi\, c^2}} \exp(-\sigma^2/(2\, c^2))$$

Conclude: Disagreements among the experiments are larger than predicted by standard Gaussian statistics; but less than a factor of 2 larger.

Conclusion from the consistency study



This fit provided direct evidence of a significant source of discrepancy associated with fixed-target DIS experiments for large x at small Q. (Higher-twist effects had been seen there previously; but not taken into account in PDF fitting — at least by CTEQ.) Removing those data by a kinematic cut makes the average disagreement smaller, but it still does not become consistent with the absolute Gaussian.

In hep-ph/0909.0268, I argue that this suggests a "tolerance criterion"

$$\Delta\chi^2 \approx (1 \times 1.64 \times 2)^2 \approx 10$$

for 90% confidence uncertainty estimation.

It is possible that other uncertainties in the analysis require larger $\Delta\chi^2$; but the experimental inconsistencies do not.

Parametrization dependence

The PDF for each flavor at μ_0 is an unknown continuous function of x. We approximate it by some simple analytic form with 5 or 6 free parameters. This introduces a systematic error called parametrization dependence.

How big is the parametrization dependence?

Let parameter z represents a physical observable, after a linear transformation to make central value 0 and S.D. 1, e.g.

$$\sigma_{t\bar{t}} = a + b\,z$$

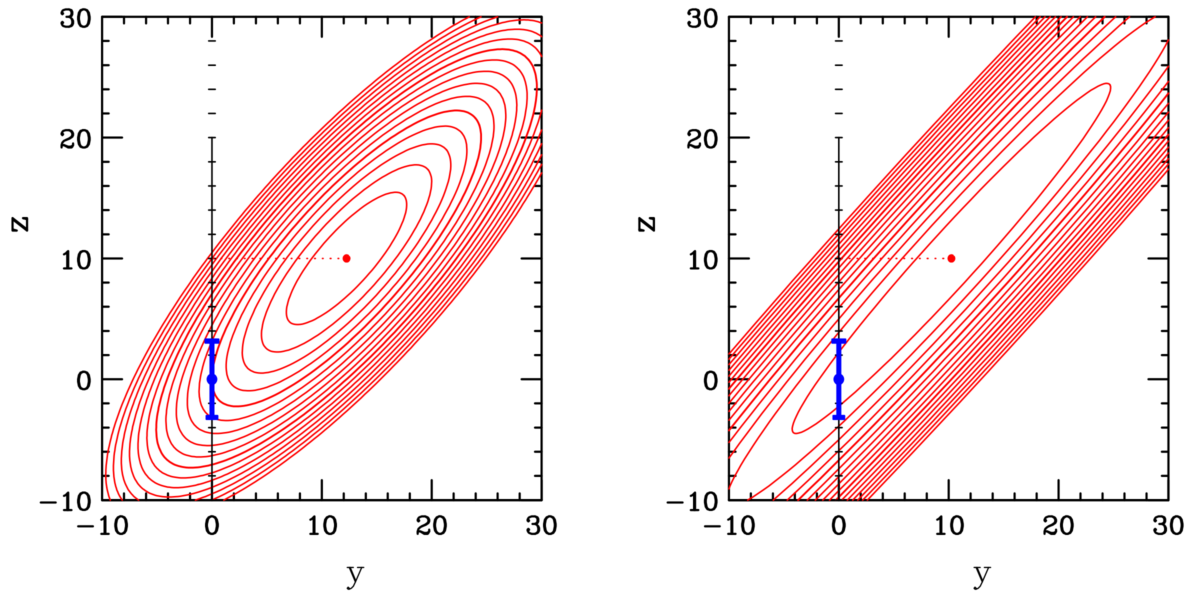
or

$$g(x,\mu) = c + d\,z\;.$$

Let parameter y represent the displacement along a direction in an expanded fitting space that was neglected in the parametrization choice.

Contours of $\chi^2 = 3010$, 3020, 3030,. . . , 3110 with minimum $\chi^2 = 3000$ for two hypothetical cases:

Hypothetical parametrization dependence



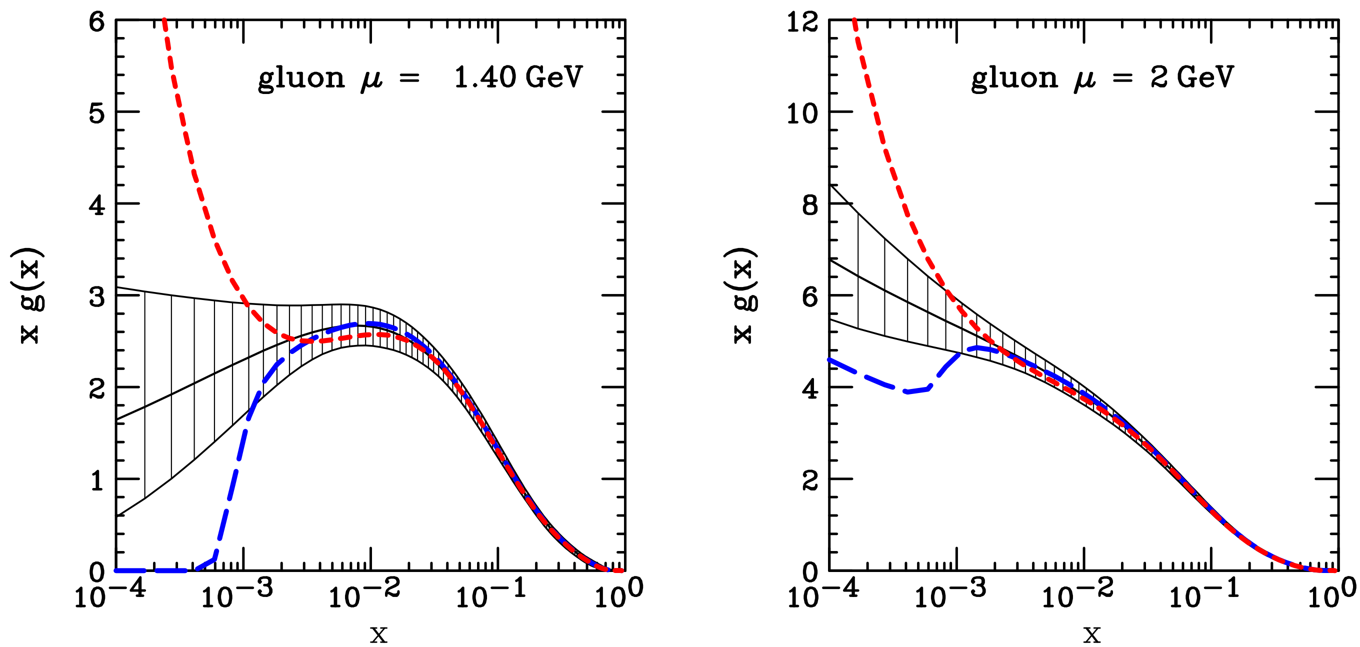
Minimum is at $\chi^2 = 3000$. Contours show $\chi^2 = 3010, 3020, 3030, \ldots, 3110$.

For $y = 0$, the minimum of χ^2 is 3005 at $z = 0$. Error bars show the $\Delta\chi^2 = 10$ error limits along $y = 0$.

Parametrizations are historically considered to be adequate if more elaborate ones only lower χ^2 by a few units. In one of these examples, introducing the parameter y would lower χ^2 by only 5 out of 3000.

The true uncertainty can be much larger than the $\Delta\chi^2 = 10$ error limits calculated with $y = 0$ because of parametrization dependence. Does this happen in practice??

Parametrization dependence at small x



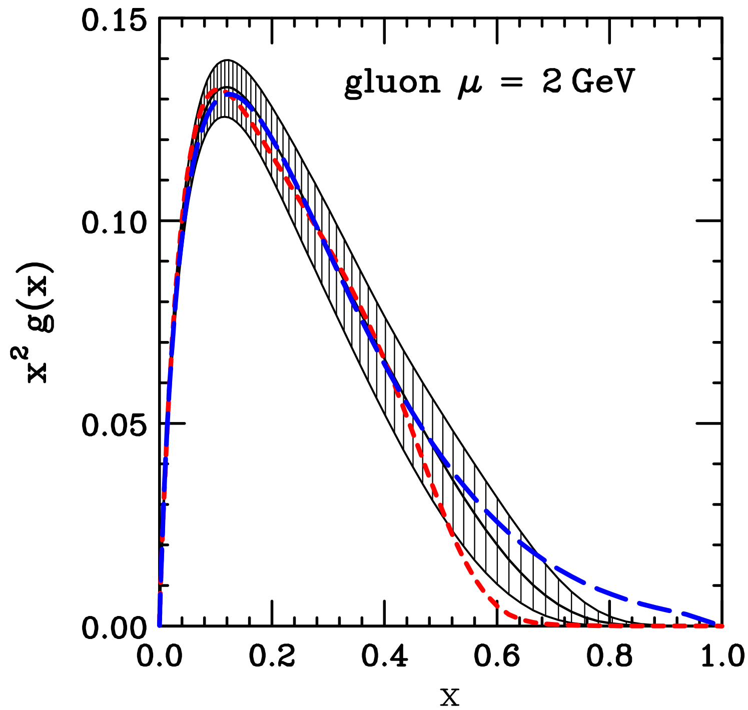
Shaded area is $\Delta\chi^2 = 10$ uncertainty using standard parametrization

$$a_0\, x^{a_1}\, (1 - x)^{a_2} \times (\text{smooth function of x})$$

Curves show results of alternative parametrizations that enhance or suppress the gluon at small x

In a region where the data provide little constraint, the true uncertainty is much larger than is predicted by $\Delta\chi^2 = 10$ — or even $\Delta\chi^2 = 100$ — because of parametrization dependence.

Parametrization dependence at large x



Our standard fitting procedure adds a penalty to χ^2 to force "expected" behavior for the gluon distribution at large x: $1.5 < a_2 < 10$ in

$$x\, g(x, \mu_0) = a_0\, x^{a_1} (1 - x)^{a_2} \exp(a_3 \sqrt{x} + a_4 x + a_5 x^2)$$

Figure shows the $\Delta\chi^2 = 10$ uncertainty range. Curves show $a_2 = 54$ (which produces $\Delta\chi^2 = 10$) and $a_2 = 0$ (which requires almost zero $\Delta\chi^2$)

Non-perturbative theory constraints are important at large x.

Even without the constraints, it is difficult to include the full range of uncertainty at large x using the Hessian method.

Chebyshev Polynomial method

Typical recent gluon parametrization:

$$x\,g(x) = a_0\,x^{a_1}\,(1-x)^{a_2}\,e^{p(x)}$$

where

$$p(x) = -9.53\sqrt{x} + 3.86x + 0.62x^2$$

Attractive features:

- Smooth function
- Power-law dependence at $x \to 0$ — Regge theory ((DGLAP issue?))
- Subleading terms down by $\sim x^{0.5}$ at $x \to 0$ — more Regge theory
- $g(x)$ positive definite ((possibly too strong))

Natural to add more flexibility to remove systematic error caused by the choice of form, by including terms with additional powers of $\sqrt{x}$.

Typical recent gluon parametrization:

$$x\,g(x) = a_0\,x^{a_1}\,(1-x)^{a_2}\,e^{p(x)}$$

where

$$p(x) = -9.53\sqrt{x} + 3.86x + 0.62x^2$$

Problem: When more terms are included, the coefficients quickly become large and the fit becomes unstable.

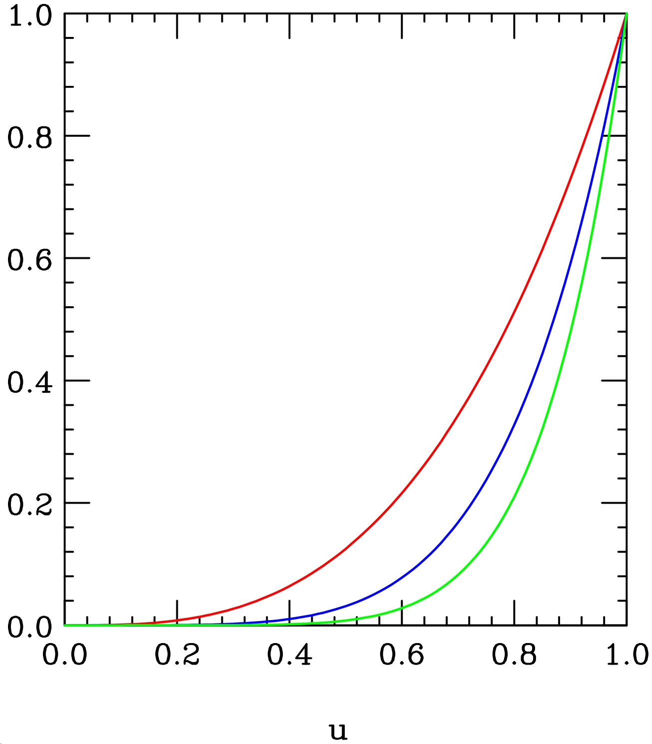
Solution: Rewrite the polynomial as a sum of Chebyshev polynomials: e.g.

$$p = 2.56T_1(y) + 0.62T_2(y) - 0.039T_3(y) + 0.005T_4(y)$$

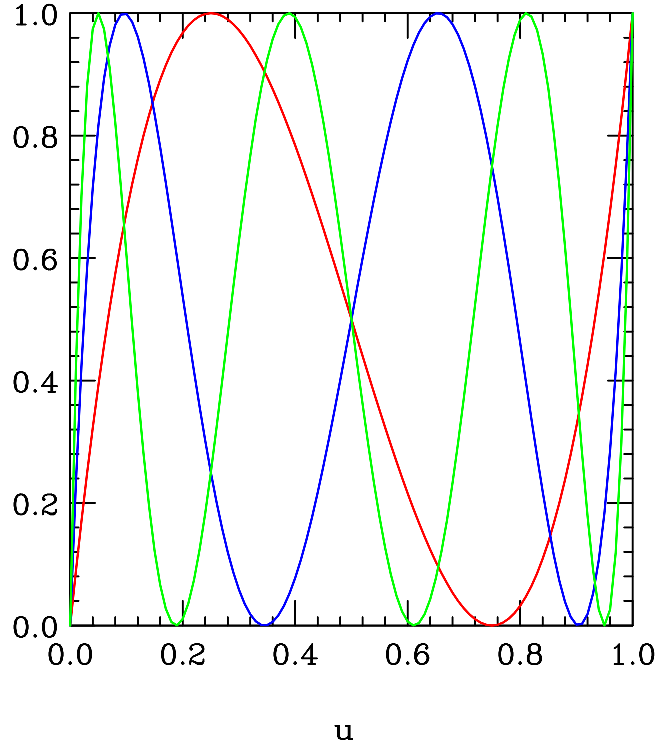
where $y = 1 - 2\sqrt{x}$.

Mathematically equivalent, but Chebyshev expansion is systematic in terms of the scale of structures. Coefficients remain small and decrease with order, allowing many more terms to be included.

Chebyshev vs. power laws



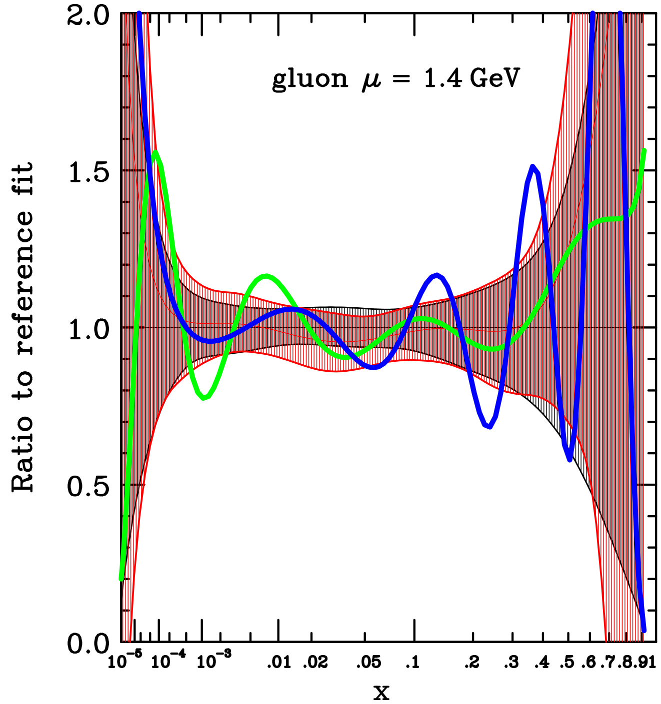
u^3, u^5, u^7



$T_3(u)$, $T_5(u)$, $T_7(u)$

$T_n(u) = \cos(n\theta)$ where $u = \cos(\theta)$.

Chebyshev fits



Gluon uncertainty bands with $\Delta\chi^2 = 10$.

Black: 24 fitting parameters
Red: 37 fitting parameters

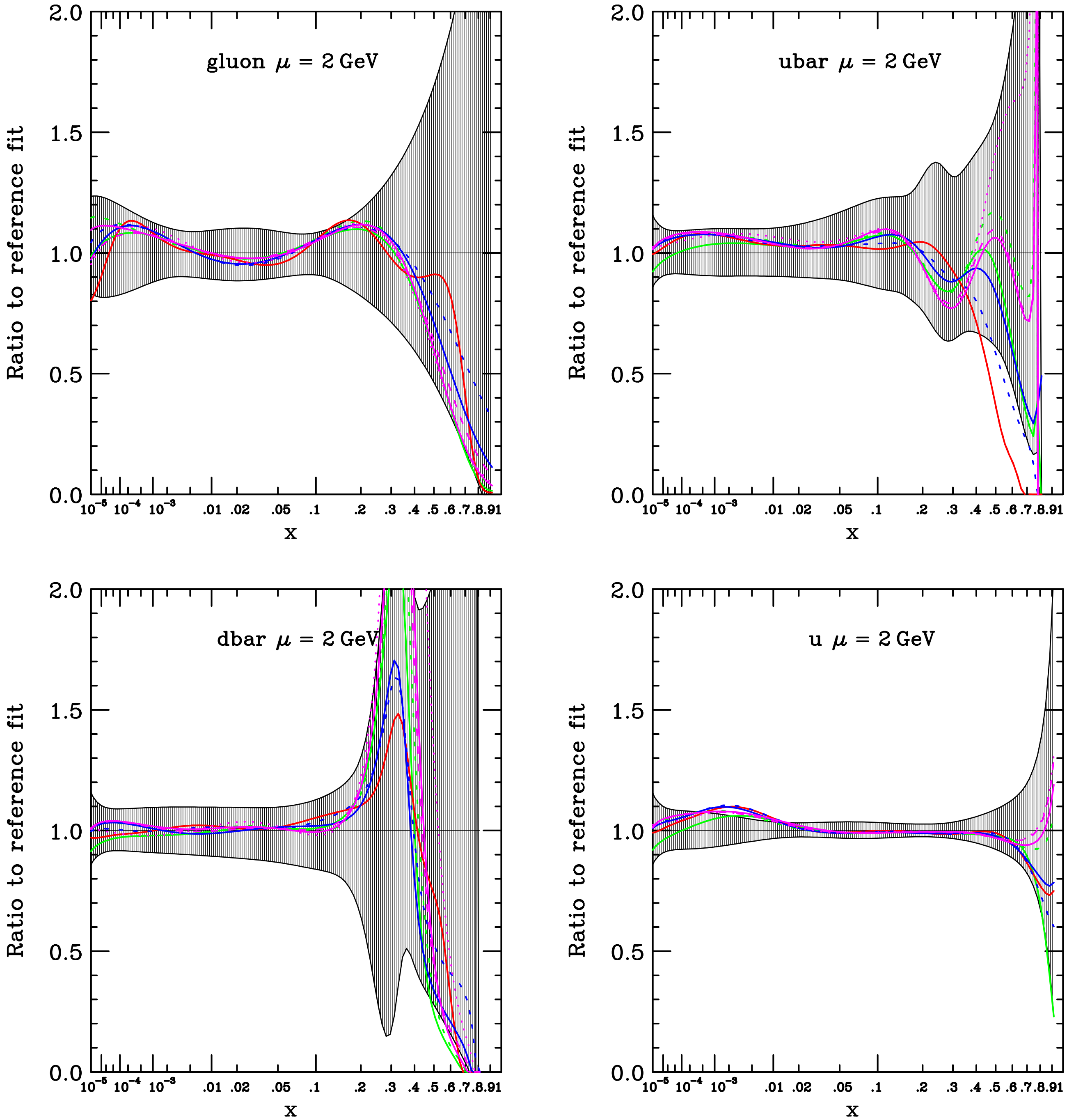
- Small increase in range of uncertainty.

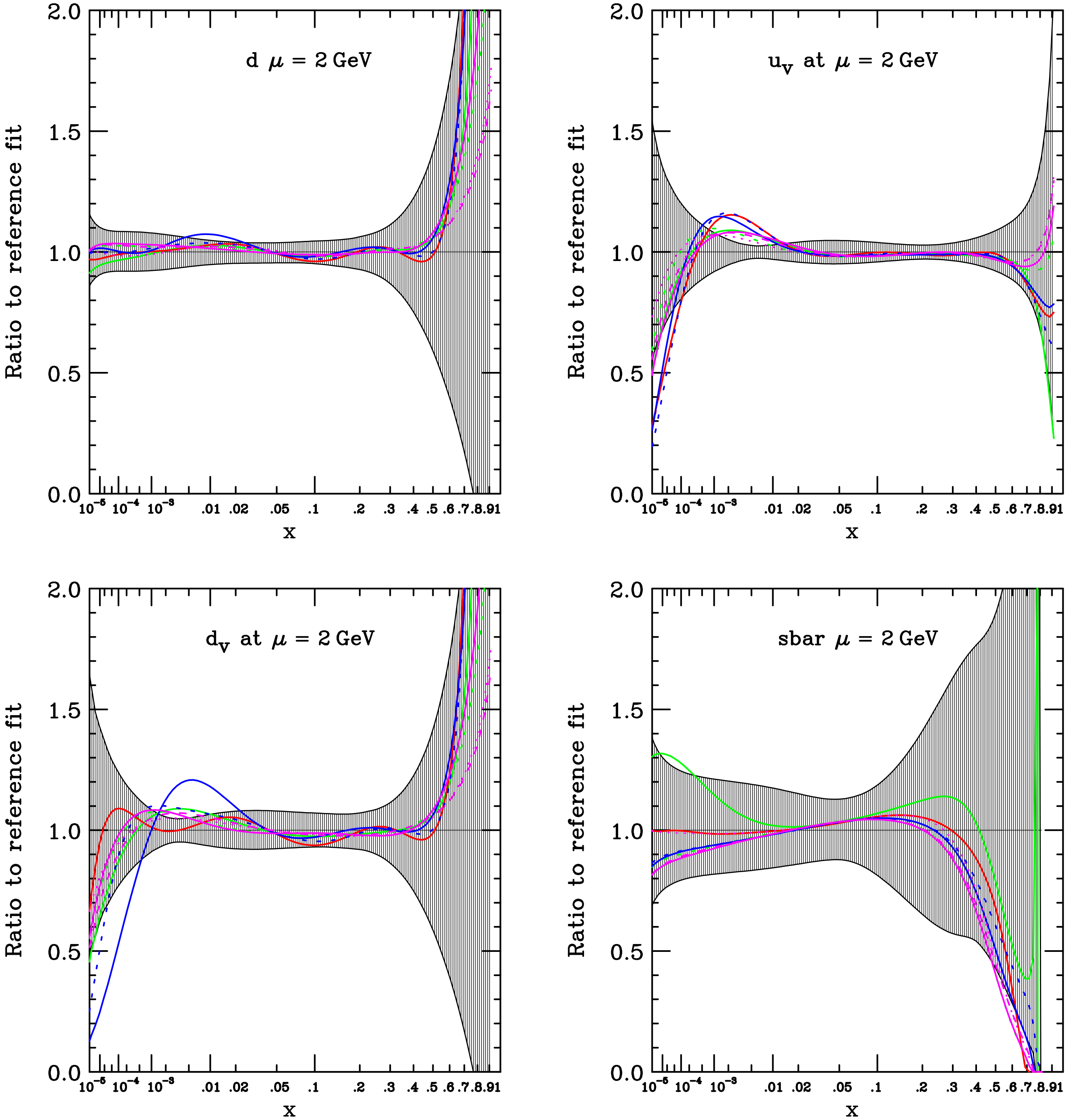
- Change in central fit

Green (58 parameters); blue (68 parameters)
Shows that Parametrization Dependence contributes uncertainty that is larger than $\Delta\chi^2 = 10$.

“Time dependence” of PDFs

CTEQ6.6 uncertainties ($\Delta\chi^2 = 100$) compared to some recent fits.





Conclude: The $\Delta\chi^2 = 100$ uncertainty estimate in CTEQ6.6 was not overly conservative.