Parametrization dependence and Consistency between experiments in parton distribution fitting

Jon Pumplin — DIS 2011 (4/12/2011)

CTEQ/TEA collaborators: J. Huston, H.-L. Lai, P. Nadolsky, C.-P. Yuan

Topic 1: Chebyshev method to reduce parametrization dependence.

Topic 2: Assessing the internal consistency of global PDF fits.

The standard QCD fitting paragigm

1. Parametrize the PDFs $f_a(x, \mu_0)$ at a small μ_0 by smooth functions with lots of free parameters.

2. Calculate $f_a(x, \mu)$ at all $\mu > \mu_0$ by DGLAP evolution.

3. Calculate $\chi^2 = \sum_i [(\text{data}_i - \text{theory}_i)/\text{error}_i]^2$ to measure of the quality of fit to a large variety of experiments.

4. Obtain a Best Fit estimate of the true PDFs by varying the free parameters to minimize χ^2.

5. Estimate the uncertainty range as all PDF sets with χ^2 not more than some $\Delta\chi^2$ (~ 50) above the Best Fit value. (This uncertainty range is characterized by a collection of PDF sets obtained using the eigenvectors of the Hessian matrix.)

Effective χ^2

The χ^2 that is minimized in the global fit includes contributions from experimental systematic errors, as published with the data.

An "Effective χ^2" can be defined by adding penalties to enforce desired features of the global fit, such as that it should agree reasonably well with every experiment.

The Effective χ^2 can also include penalties to enforce desired theoretical features of the fit, such as smoothness of the distributions. This is particularly necessary when a large number of parameters are introduced in order to reduce the dependence on the choice of functional forms.

PDF parametrizations at $Q_0 \sim 1.4\,\mathrm{GeV}$

Typical recent gluon parametrization (CT10)

$$x\,g(x,\mu_0) = a_0\,x^{a_1}\,(1-x)^{a_2}\,e^{p(x)}$$

where

$$p(x) = a_3\sqrt{x} + a_4 x + a_5 x^2$$

- Power-law dependence at $x \to 0$ (Regge theory)
- Spectator counting behavior at $x \to 1$
- Subleading terms down by $\sim x^{0.5}$ at $x \to 0$ (Regge)
- $g(x)$ positive definite (possibly too strong)

New method: very free parametrization

$$p(x) = \sum_{j=1}^{12} b_j\,x^{j/2}$$

Challenge: How to obtain smooth, stable fits with ~ 80 free parameters instead of ~ 25.

Chebyshev Polynomial method

Replace the fitting parameters $\{b_j\}$ by equivalent parameters $\{c_j\}$ where

$$p(x) = \sum_{j=1}^{12} b_j \, x^{j/2} = \sum_{j=1}^{12} c_j \, T_j(y)$$

where $y = 1 - 2\sqrt{x}$ maps the physical region $0 < x < 1$ to $-1 < y < 1$.

Definition of the Chebyshev polynomials:

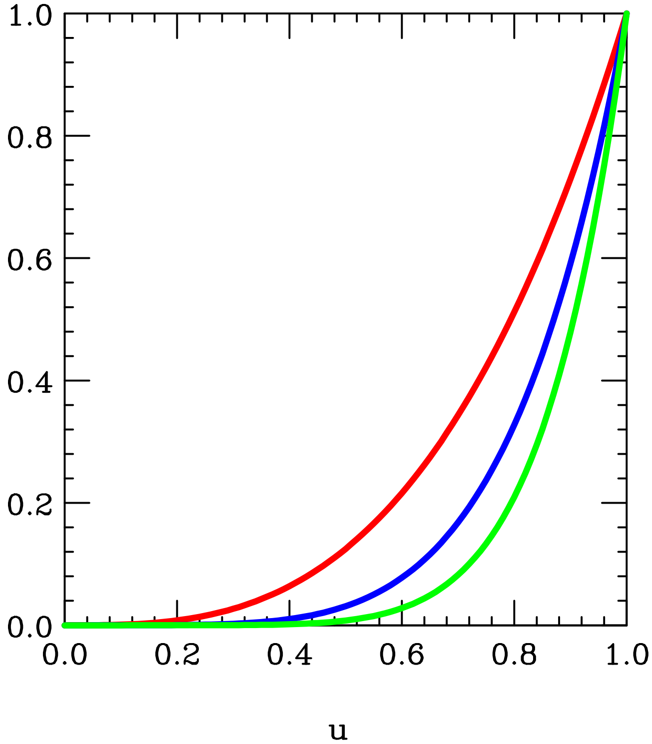
$$T_0(y) = 1$$

$$T_1(y) = y$$

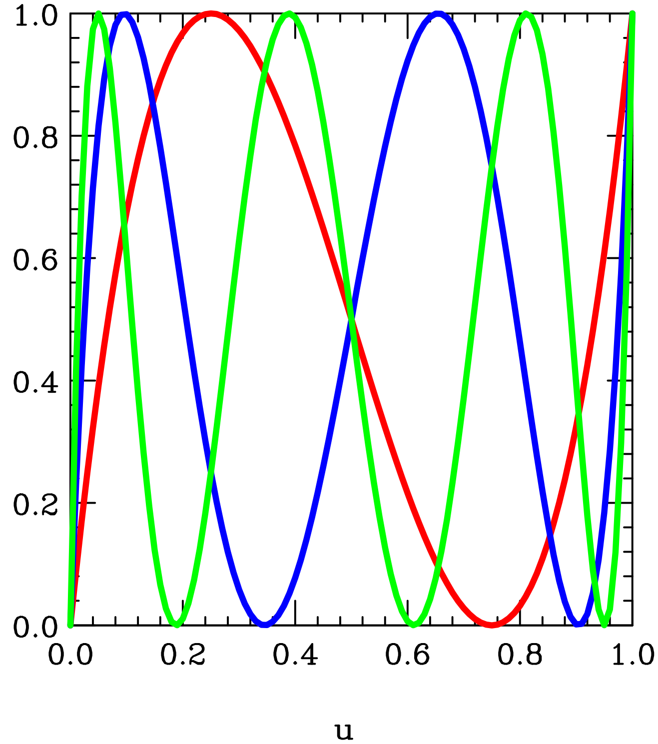
$$T_{n+1}(y) = 2yT_n(y) - T_{n-1}(y)$$

$T_j(y)$ has extreme values of ± 1 at the endpoints and at $j-1$ points in the interior of the physical region $0 < x < 1$. Chebyshev polynomials of increasingly large j thus model structure at an increasingly fine scale in $\sqrt{x}$.

Chebyshev polynomials vs. simple powers



u^3, u^5, u^7



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

Equivalent definition for Chebyshev polynomials:

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

where

$$u = \cos(\theta)$$

Penalty to enforce Smoothness

The Chebyshev parametrizations can easily take on more fine structure in x than is plausible in the nonperturbative physics that is being described.
To avoid this, we add a penalty to χ^2

Observe that the classic form

$$f(x) = a_0\, x^{a_1}\, (1-x)^{a_2}\ ,$$

surely embodies the appropriate smoothness. It has

$$x\,(1-x)\, d(\ln f)/dx \;=\; a_1 - (a_1 + a_2)x$$

which is linear in x. Hence it is natural to define

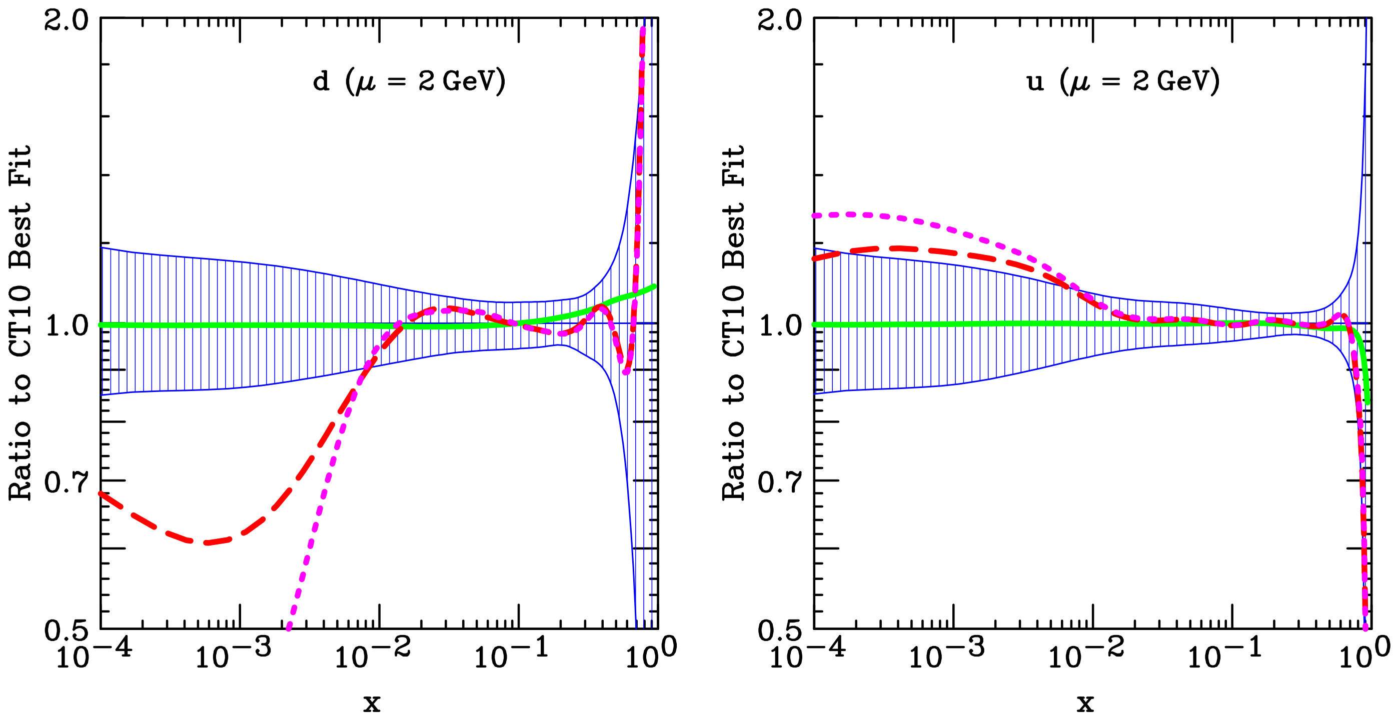
$$\Phi_a(x) \;=\; x\,(1-x)\, d(\ln f_a)/dx$$

$$S_a \;=\; \int_{x_1}^{x_2} \left(\frac{d^2\Phi_a}{dx^2}\right)^2 dx$$

Add $\sum_a C_a\, S_a$ to χ^2, where the sum runs over flavors $a = u$, $\bar{u}$, d, $\bar{d}$, g, and s. The weights C_a are chosen to make the overall penalty ~ 5.

A penalty based on the approximate behavior $f(x) = a_0\, x^{a_1}$ is also included, to enforce smoothness at small x.

Chebyshev fits



Shaded region: fractional uncertainty from CT10 (26 free parameters)

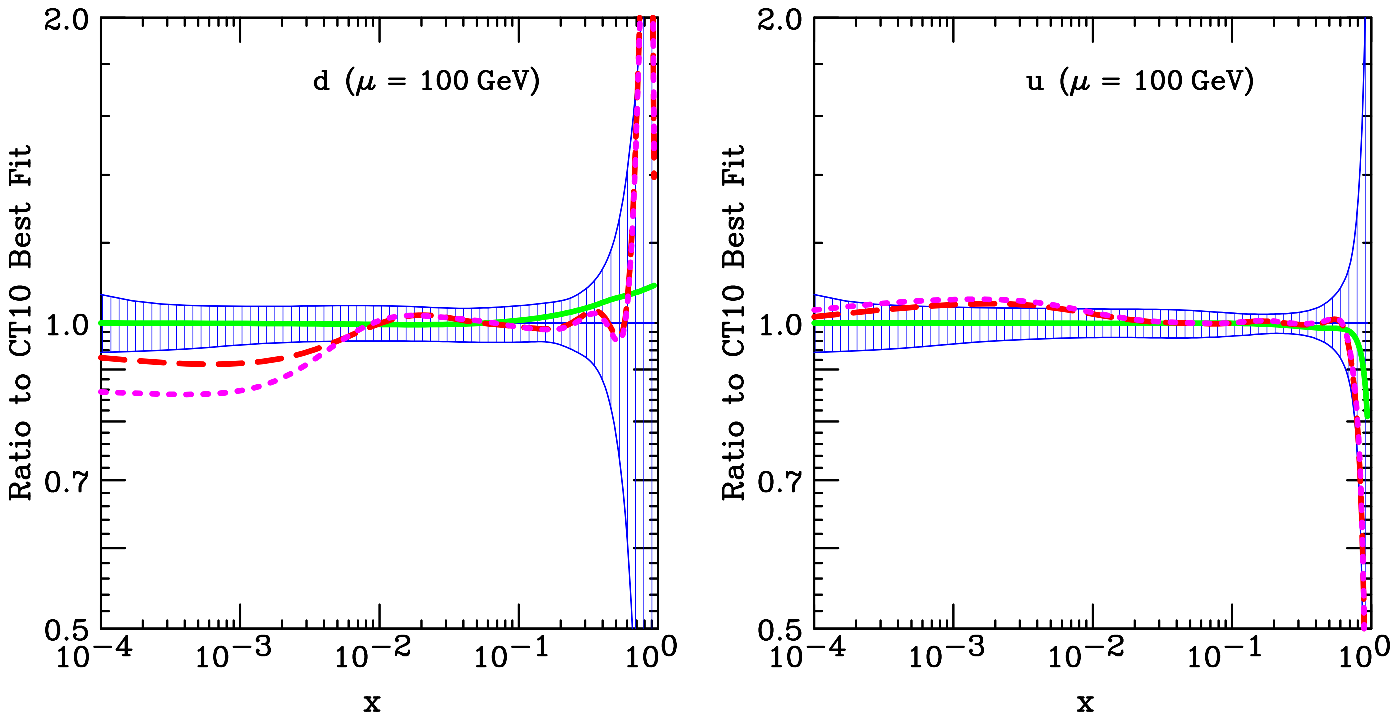
Solid: include Tevatron $W \to e\nu$ asymmetry (wgt 1)

Dashed: Chebyshev fit with $\bar{d}(x)/\bar{u}(x) \to 1$ at $x \to 0$ (83 free parameters, χ^2 lower by ~ 140).

Dotted: Chebyshev fit with $\bar{d}(x)/\bar{u}(x)$ free at $x \to 0$ (84 free parameters, χ^2 lower by ~ 140).

Improvement in χ^2 comes mostly from expts that are sensitive to flavor ratios (Wasy), or to large-x behavior (NMC and BCDMS $\mu p \to \mu X$ and $\mu d \to \mu X$).

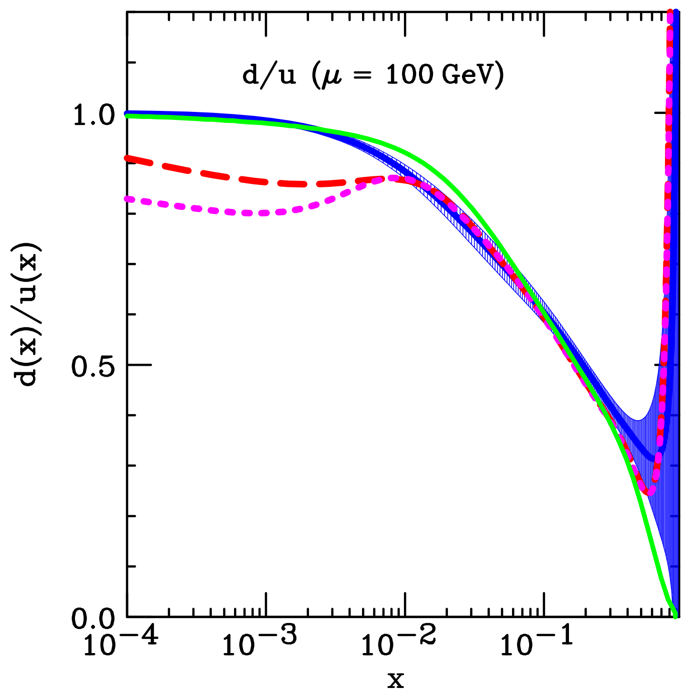
Chebyshev fits at scale = 100 GeV



The change in the central fit that results from increased freedom in the parametrization persists to large scales, as shown here.

The change for $x < 10^{-2}$ is mainly in the $\bar{d}(x)$ and $\bar{u}(x)$ distributions, since $d(x) \approx \bar{d}(x)$ and $u(x) \approx \bar{u}(x)$ in that region.

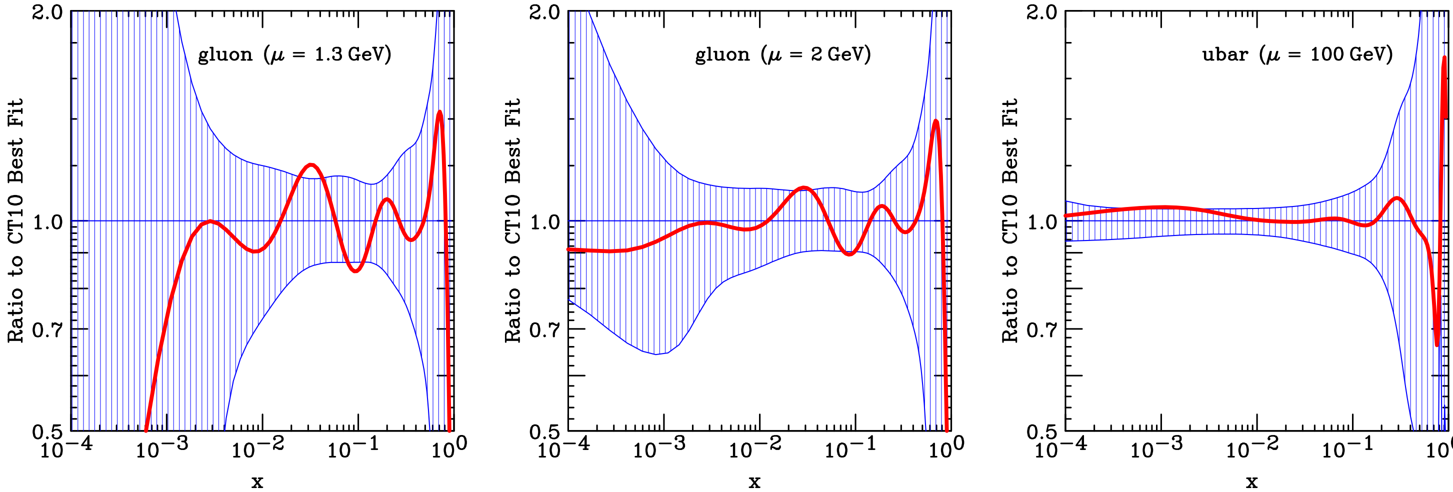
d/u and $\bar{d}/\bar{u}$ at scale = 100 GeV



The increased uncertainty is especially strong in d/u or $\bar{d}/\bar{u}$. It may therefore be important for analyzing the W lepton decay asymmetry at the LHC.

The solid curve is MSTW2008nlo. It lies outside the CT10 uncertainty band in a different way — at least supporting the notion that d/u uncertainty was underestimated in previous fits.

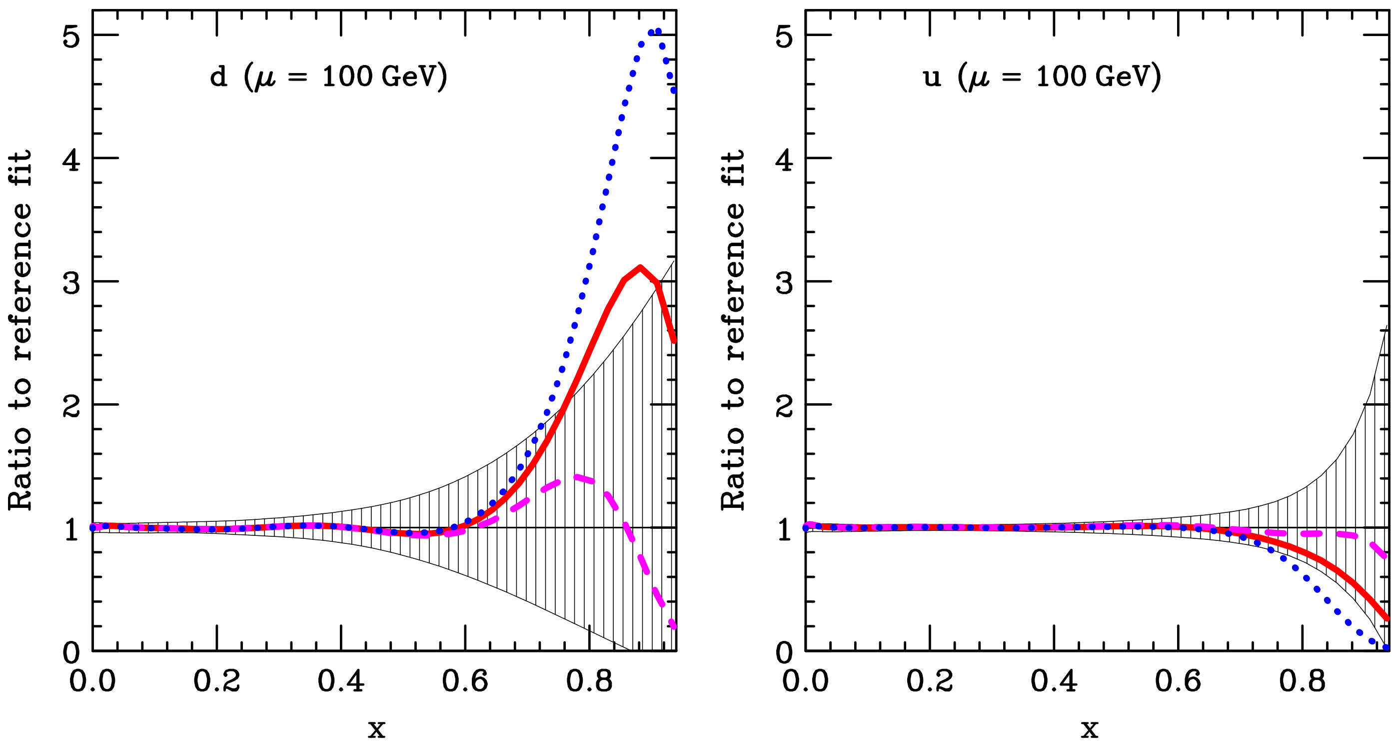
The Smoothing effect of evolution



A fit with reduced smoothness penalty: lots of structure at the input scale $\mu = 1.3\,\text{GeV}$.

Even that much structure is greatly reduced at higher scales due to DGLAP evolution.

Uncertainties at large x



Shaded region: fractional uncertainty according to CT10 (26 free parameters, $\Delta\chi^2 = 100$)

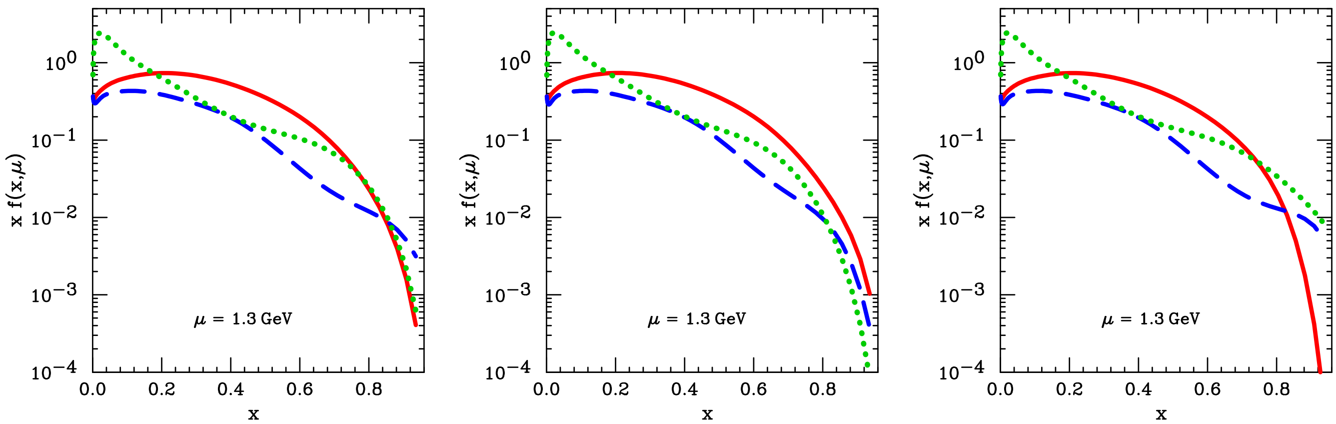
Red curve: Chebyshev Best Fit (84 free parameters)

Blue and Magenta curves: Chebyshev fits with χ^2 higher than Best Fit by only 5.

The true uncertainty at $x \to 1$ is therefore much larger than the estimate from CT10, or other fits based on the Hessian method.

Perhaps some theoretical constraints from Nonperturbative methods can be used to reduce this uncertainty?

Parton distributions at $x \to 1$



Each panel shows results from a Chebyshev fit with very good χ^2.

Solid = up quark
Dashed = down quark
Dotted = gluon

At low scale ($Q = 1.3\,\text{GeV}$) any of the flavors u, d, or g can dominate in the limit $x \to 1$.

Perhaps experts on Nonperturbative models can provide useful restrictions on this uncertainty? (Maybe they can start by explaining why the gluon/quark ratio is not small at $x \to 1$!)

Conclusions on Chebyshev fits

- Chebyshev polynomials combined with smoothness constraints can produce stable PDF fits with ~ 80 free parameters.

- Quality of the fit improves — overall decrease in real χ^2, compared to fits with the traditional ~ 25 parameters is ~ 140.

- Increased flexibility of parametrization expands the uncertainty range for $d(x)/u(x)$ and $\bar{d}(x)/\bar{u}(x)$. This may be important for interpreting the LHC W lepton asymmetry.

Part 2: Internal Consistency of PDF fits

1. Overall χ^2/N
2. χ^2/N for individual experiments
3. DSD method

χ^2 per point

Gaussian statistics predicts $1\,\sigma$ range of $\chi^2 = N \pm \sqrt{2N}$

A typical recent fit (CT10) has $\chi^2 = 3015$ for 2753 data points with 26 fitting parameters. That χ^2 is just a little higher than the expected range:

$$(2753 - 26) \pm \sqrt{2\,(2753 - 26)} = 2650 \text{ to } 2800$$

(Expanding all of the experimental errors uniformly by only 4% would be enough to make the fit appear consistent from this point of view.)

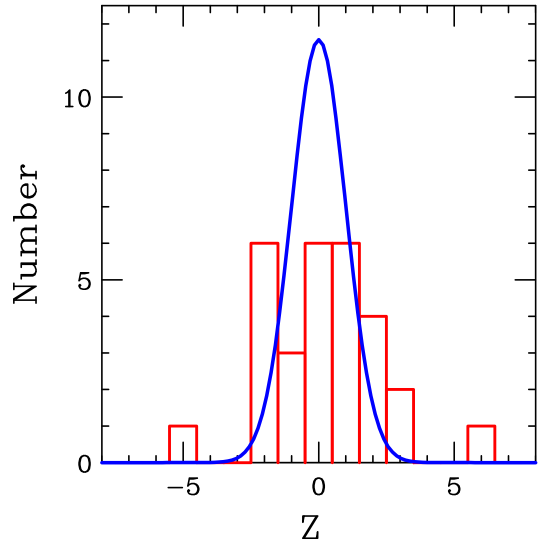
To compare the quality of fit to different experiments with widely different numbers of points, it is convenient to use an Equivalent Normal Distribution variable Z defined such that the cumulative probability for χ^2 according to the statistical χ^2 distribution is equal to the cumulative probability distribution at Z for the Normal distribution.

The distribution of fits to the individual experiments in CT10 is broader than would be expected from Gaussian statistics.

Improvement in Fit quality

Expt	N	Z(CT10)	Z(Cheby)
HERA	579	2.84	2.45
BCDMS F_2p	339	1.51	0.81
BCDMS F_2d	251	0.80	-0.18
NMC F_2p	201	**5.88**	**5.41**
NMC F_2p/d	123	0.48	-0.06
E605 DY	119	-1.60	-2.01
E866 pd/pp	15	-1.36	-2.15
E866 pp	184	2.19	1.59
CDF Wasy 1	11	-0.07	0.27
CDF Jet 1	33	2.27	2.30
D0 Jet 1	90	-1.81	-2.58
CDF Jet 2	72	2.95	0.99
D0 Jet 2	110	1.04	1.83
D0 $d\sigma_Z/dy$	28	-1.90	-1.94
CDF $d\sigma_Z/dy$	29	2.25	2.01
CDF Wasy	11	0.44	0.15
D0 Wasy 2	12	4.16	2.73
H1 F2c	8	0.67	0.75
H1 F2c	10	1.19	1.33
H1 F2b	10	-0.38	-0.45
ZEUS F2c	18	-0.20	0.02
ZEUS F2c	27	-0.85	-0.61
CDHSW F_2	85	-1.98	-1.90
CDHSW F_3	96	-2.11	-2.33
CCFR F_2	69	0.21	0.53
CCFR F_3	86	-5.18	-5.03
NuTeV ν	38	-0.42	-1.62
NuTeV $\bar{\nu}$	33	-0.55	-0.94
CCFR ν	40	1.14	1.09
CCFR $\bar{\nu}$	38	-0.95	-1.07

Distribution of Z for CT10



Smooth curve is gaussian with mean 0 and standard deviation 1.

The observed histogram is broader than this prediction, which demonstrates that the uncertainties are not strictly gaussian.

Data Set Diagonalization

Partition the data into two subsets:

$$\chi^2 = \chi^2_S + \chi^2_{\overline{S}} \, .$$

S can be any one of the experiments, or all experiments of a particular type that might be suspected of an untreated systematic error.

$\overline{S}$ is all the rest of the data.

The DSD method answers the questions

1) What does subset S measure?

2) How consistent is S with the rest of the data?

The essential trick is that in the Hessian method, the linear transformation that leads to

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

is not unique, since any further orthogonal transform of the z_i will preserve it. Such a transformation can be defined using the eigenvectors of the quadratic form corresponding to χ^2_S. Then . . .

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

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

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

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

$$\gamma_i = D_i^2/(B_i^2 + D_i^2)\,.$$

The decomposition also measures the compatibility between S and $\overline{S}$: the disagreement between the two in standard deviations is

$$\sigma_i = |A_i - C_i|/\sqrt{B_i^2 + D_i^2}\,.$$

Experiments that provide at least one measurement with $\gamma_i > 0.1$ ($\sim$ CTEQ6.6)

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.

Consistency tests

Measurements that conflict strongly with the others ($\sigma_i > 3$) are shown in red. There are lots of them!

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

Only measurements that play a significant role ($\gamma_i > 0.1$) are listed.

Outlook

Plan to carry out the DSD procedure for Chebyshev fits.