# Topics in PDF fitting

Jon Pumplin

PDF for LHC (DESY 23 October 2009)

Topics:

1. PDF theory constraints

2. Consistency of data sets used in PDF fitting (arXiv:0909.0268,arXiv:0904.2425)

3. Choice of $\Delta\chi^2$ (arXiv:0909.5176: heavily revised version in progress)

4. Brief comments

Regge behavior of $\bar{u}$ and $\bar{d}$

The Regge behavior $x\,\bar{u}(x,\mu) \propto x^{a_1}$ that we assume for $x \to 0$ at μ_0 is quite well preserved by DGLAP evolution. This can be seen by the nearly straight-line behavior on a log-log plot, with slope nearly independent of μ:
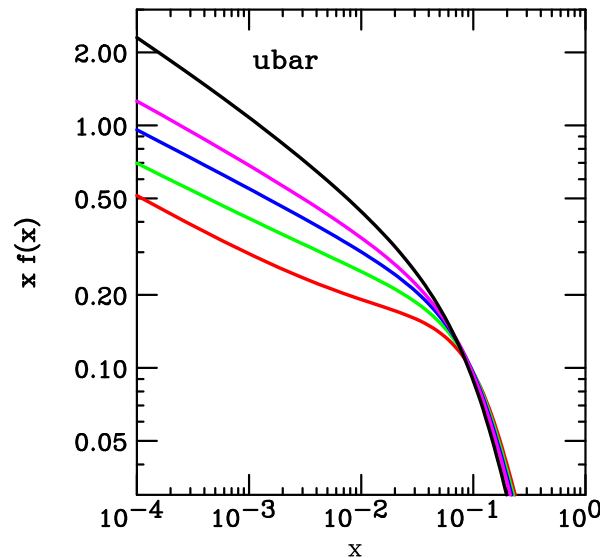
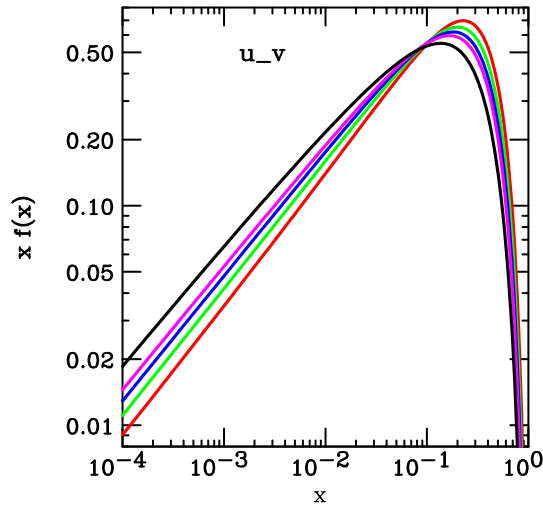$\mu = 1.3/2.0/3.2/5.0/20\,\mathrm{GeV}$

Numerical value of the slope a_1 agrees well with expectations from Regge, which supports the use of the $x\,\bar{u}(x,\mu) \propto x^{a_1}$ ansatz.

Regge theory does not provide a useful constraint on a_1, because the uncertainty from PDF fitting is smaller than the uncertainty of estimates from strong-interaction phenomenology.

These considerations also demand $\bar{u}/\bar{d} \to 1$ at $x \to 0$.

Regge behavior of u_v and d_v

$u_v \equiv u - \bar{u}$, $d_v \equiv d - \bar{d}$. The Regge behavior $x\, u_v(x, \mu) \propto x^{a_1}$ that we assume for $x \to 0$ at μ_0 is also well preserved by DGLAP evolution:

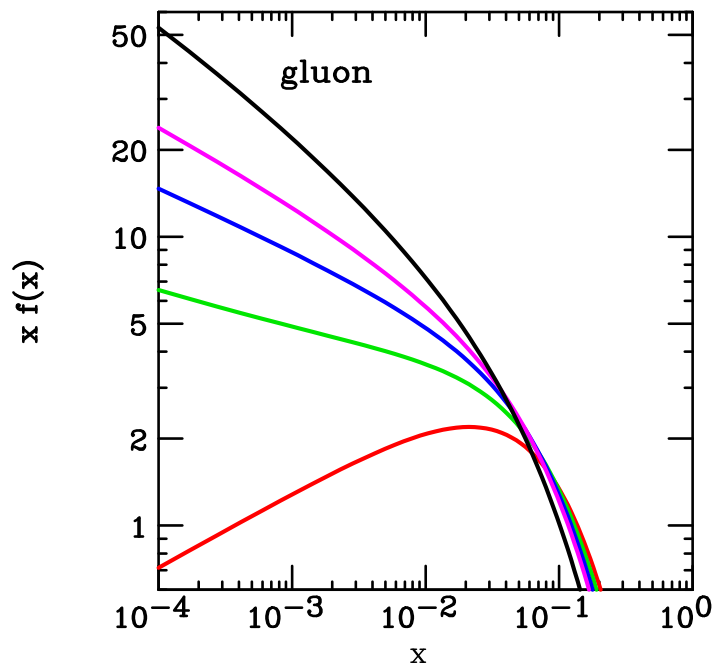$\mu = 1.3/2.0/3.2/5.0/20\,\text{GeV}$

Observed slope value a_1 is again consistent with expectations from Regge theory, which supports using this functional form.

Again the uncertainty in a_1 from PDF fitting is small compared to the uncertainty of its estimate based on Regge theory, so traditional Regge phenomenology does not provide a useful constraint on a_1 to improve the PDF determination.

These considerations also demand $u_v/d_v \to 1$ at $x \to 0$.

Regge behavior of gluon at small x??

In contrast to valence and sea quark distributions, the NLO evolution of the gluon distribution at small x is very rapid. Hence no simple comparison can be made with expectations from Regge theory:

$\mu = 1.3/2.0/3.2/5.0/20\,\text{GeV}$

Rapid change in slope is related to the rapid variation of the observed power $F_2 \sim x^{\lambda(Q^2)}$.

Speculation: perhaps small-x resummation corrections to DGLAP would restore Regge behavior for $g(x, \mu)$?

Measuring internal consistency of the fit

Partition the data into two subsets:

$$\chi^2 = \chi_S^2 + \chi_{\bar{S}}^2$$

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

$$\chi^2 = \chi_S^2 + \chi_{\overline{S}}^2 + \text{const}$$

$$\chi_S^2 = \sum_{i=1}^{N} [(z_i - A_i)/B_i]^2$$

$$\chi_{\overline{S}}^2 = \sum_{i=1}^{N} [(z_i - C_i)/D_i]^2$$

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

$$\mathbf{S} : z_i = A_i \pm B_i$$
$$\overline{S} : z_i = C_i \pm D_i$$

DSD method – continued

$$\chi^2 = \chi_S^2 + \chi_{\overline{S}}^2 + \text{const}$$

$$\chi_S^2 = \sum_{i=1}^{N} [(z_i - A_i)/B_i]^2$$

$$\chi_{\overline{S}}^2 = \sum_{i=1}^{N} [(z_i - C_i)/D_i]^2$$

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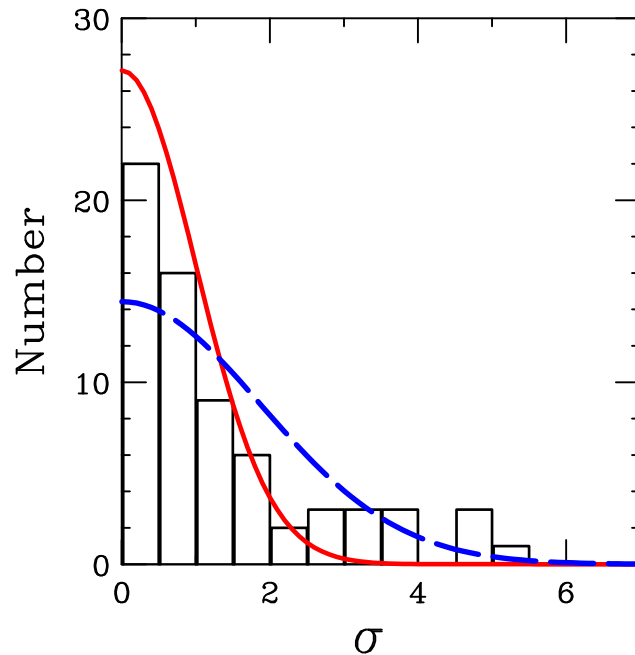
Histogram of the consistency measure σ_i for the 68 significant ($\gamma_i > 0.1$) measurements provided by the 37 experiments in a typical global fit.

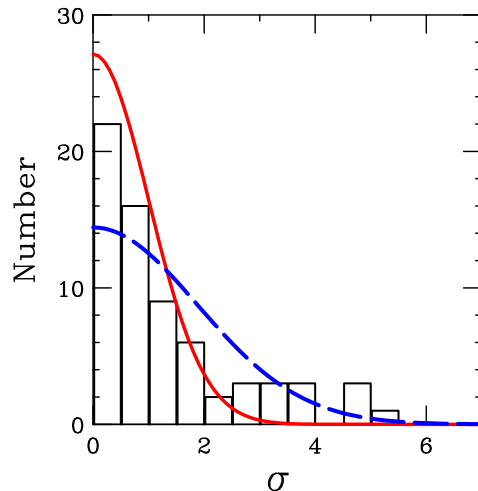Solid curve is the absolute Gaussian prediction

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

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

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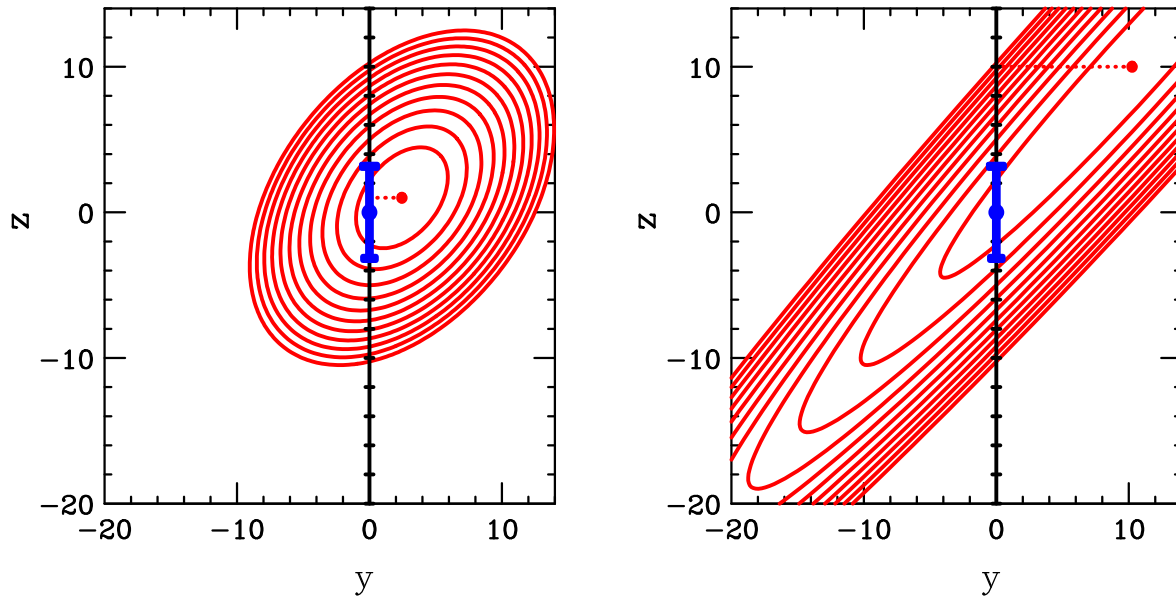
Minimum is at $\chi^2 = 3000$. Contours show $\chi^2 = 3010, 3020, 3030, \ldots, 3110$.

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 these examples, introducing the parameter y would lower χ^2 by only 5 out of 3000.

The true uncertainty is much larger than the $\Delta\chi^2 = 10$ error limits calculated with $y = 0$ in the second example. Does this happen in practice??

Parametrization dependence at small x
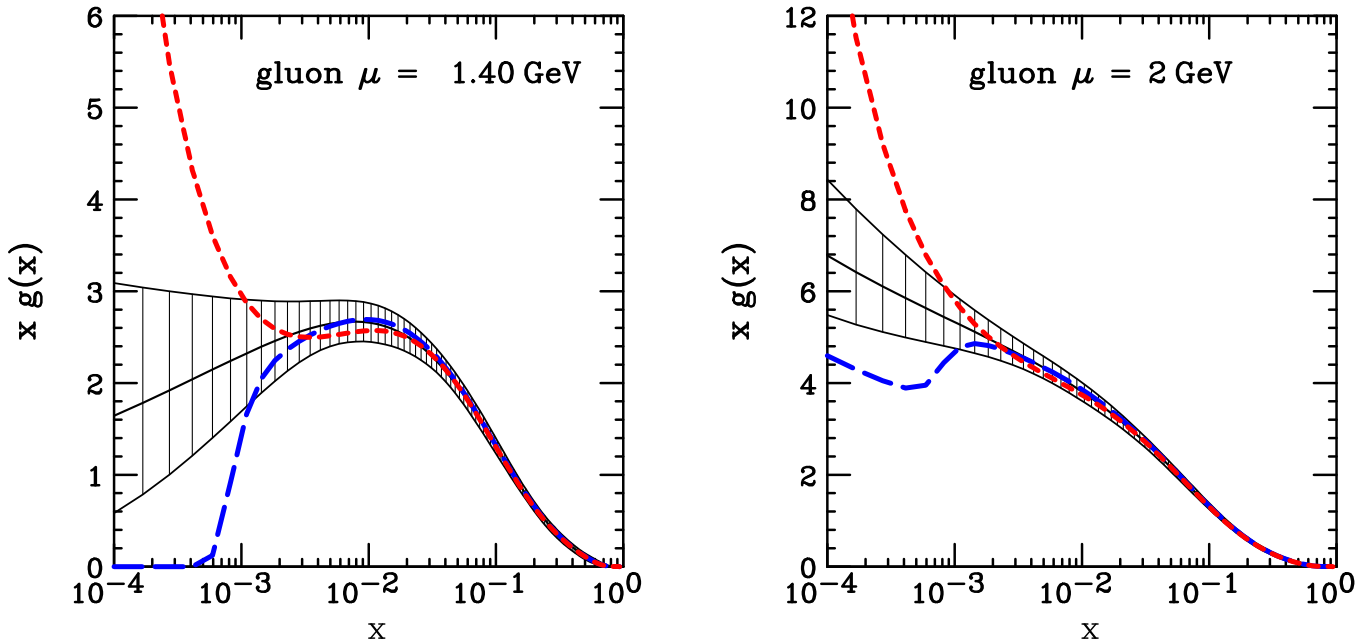
Shaded area is $\Delta\chi^2 = 10$ uncertainty using standard parametrization

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

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

gluon $\mu = 2\,\mathrm{GeV}$
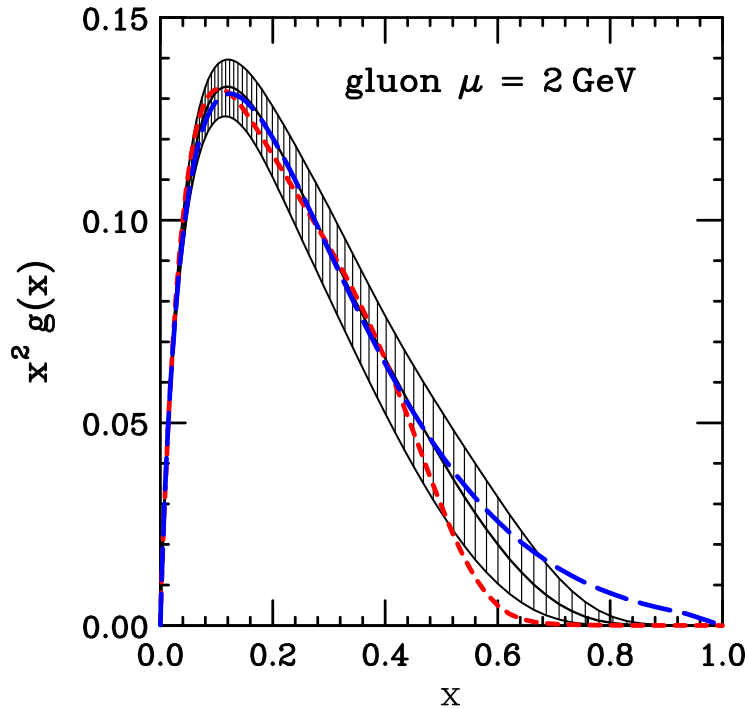
Our standard fitting procedure adds a penalty to χ^2 to force "expected" behavior for the gluon distribution at large x: $1.5 < a_2 < 10$ in

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

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.

Parametrization dependence at intermediate x
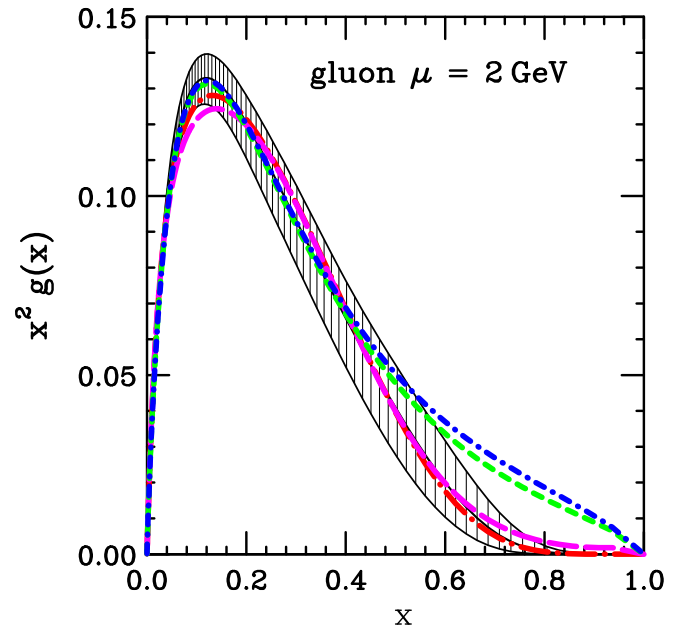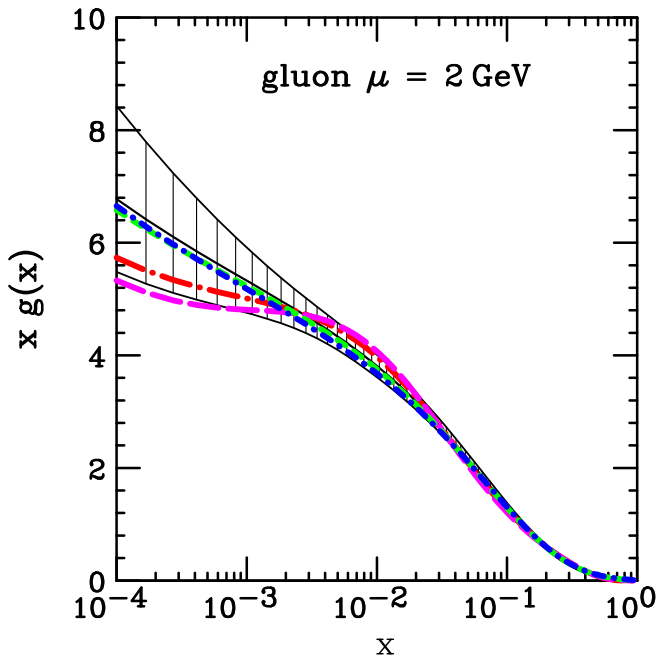
Gluon uncertainty bands with $\Delta\chi^2 = 10$.

Curves show results from alternative parametrizations with up to 8 additional parameters in the full global fit.

Added freedom reduces χ^2 by as much as 10 — 15, but change in the gluon distribution is small in the regions where it is well-determined.

Added freedom only makes larger changes at extreme x — where we already knew there is substantial parametrization dependence.

"Time dependence" of PDFs
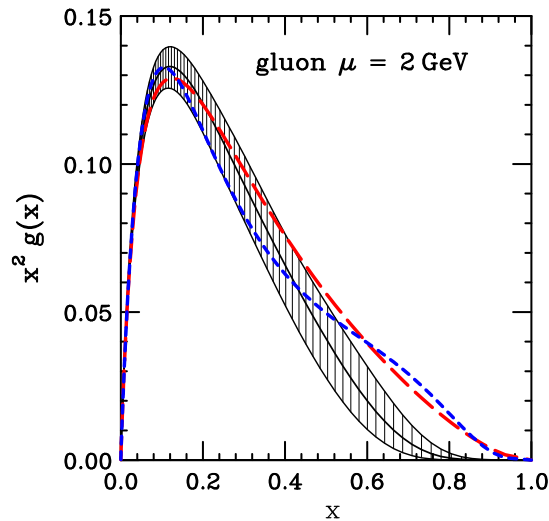
$\Delta\chi^2 = 10$ uncertainties in a recent fit (all weights 1.0; run II jet data only).

CTEQ6.6 central fit: used run I jet data only; different weights for different experiments.

CT09 central fit: used both run I and run II jet data; different weights for different experiments.

It is clear that $\Delta\chi^2 = 1$ for 68% confidence would be overly optimistic.

It appears that $\Delta\chi^2 = 10$ may be (nearly?) large enough, in regions where the data provide substantial constraint.

(Larger time-dependence would be seen for earlier PDFs because of improving treatments, e.g. of heavy quarks after CTEQ6.1.)

"Space dependence" of PDFs
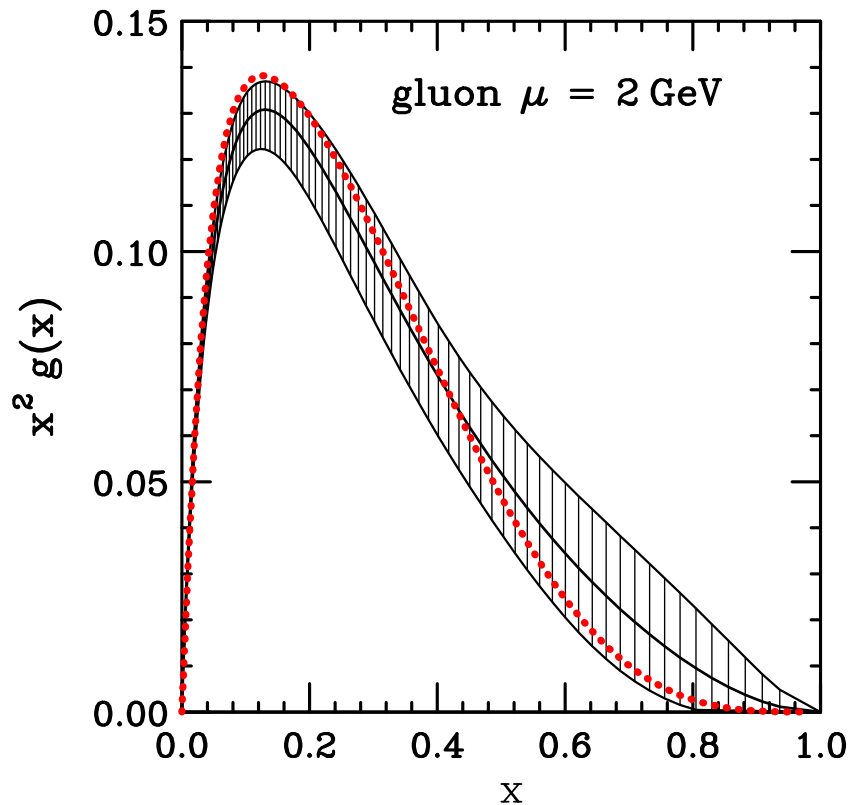
$\Delta\chi^2 = 10$ uncertainties in a recent fit (All weights 1.0; no run I jet data, $\alpha_s(m_Z) = 0.12018$ to match MSTW.)

MSTW2008 central fit

Again it is clear that $\Delta\chi^2 = 1$ for 68% confidence would be overly optimistic.

Again it appears that $\Delta\chi^2 = 10$ may be (nearly?) big enough in regions where the data provide substantial constraint.

Comment # 1: Closure
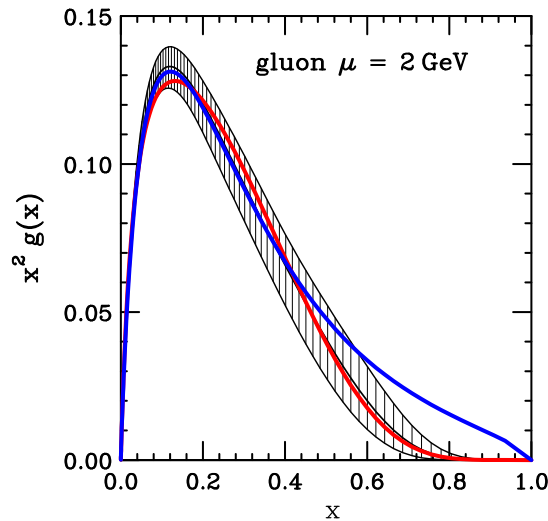
$\Delta\chi^2 = 10$ uncertainties in a recent fit with all weights 1, including run II inclusive jet data only).

Weight 5 for D0 inclusive jet data.
Weight 5 for CDF inclusive jet data.

Old interpretation: these two similar experiments seem to pull in rather different directions, which suggests the need for a large $\Delta\chi^2$ on the basis of disagreements between experiments.

New interpretation: the two experiments are in reasonable agreement with each other; but the large-x behavior is very undefined, so unimportant differences between data sets pick out different but equally-likely large-x behaviors.

# Comment # 2: 90% vs. 68% confidence
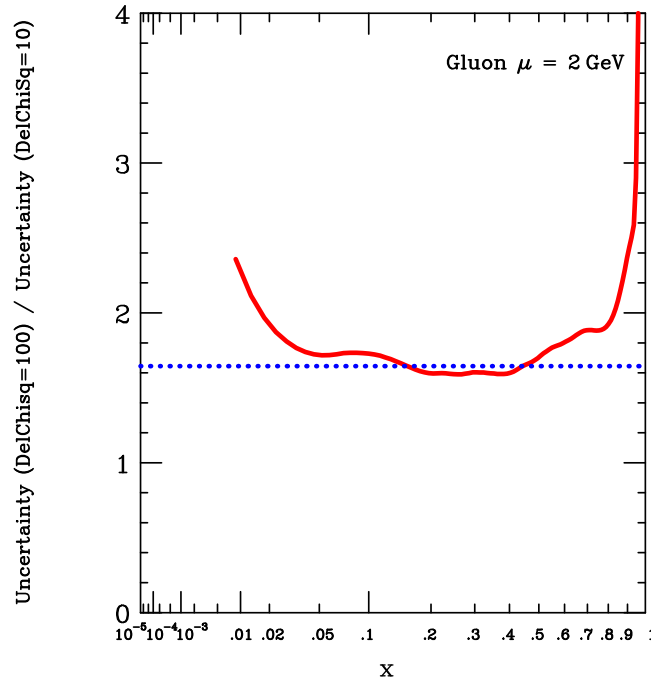
Ratio of uncertainties for gluon calculated with $\Delta\chi^2 = 100$ to uncertainty calculated with $\Delta\chi^2 = 10$.

Gaussian expectation is 1.645 (dotted line)

The $\Delta\chi^2 = 100$ choice has a good feature of expanding the uncertainty more in the regions where it needs to be expanded because of the parametrization dependence (very large and very small x).

Comment # 3: gluon uncertainty at small x

Showed earlier that $g(x)$ is extremely uncertain at low x at low μ: Figures show standard $\Delta\chi^2 = 10$ uncertainties, but curves show acceptable results from more flexible parametrizations that enhance or suppress the gluon at small x.
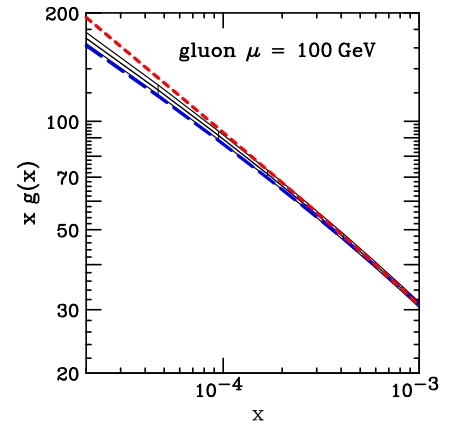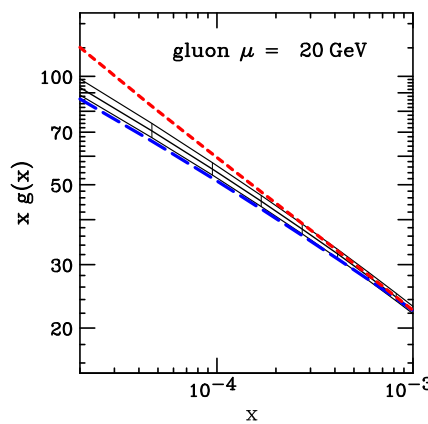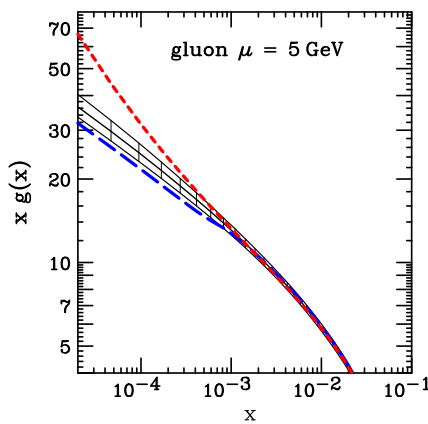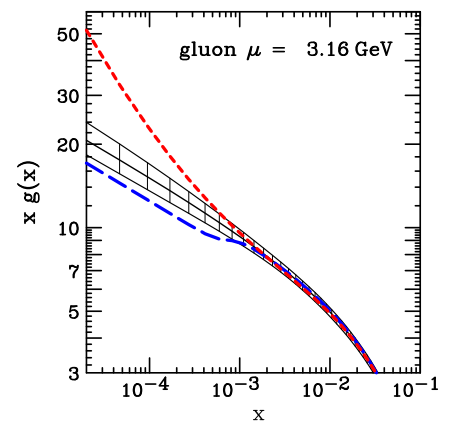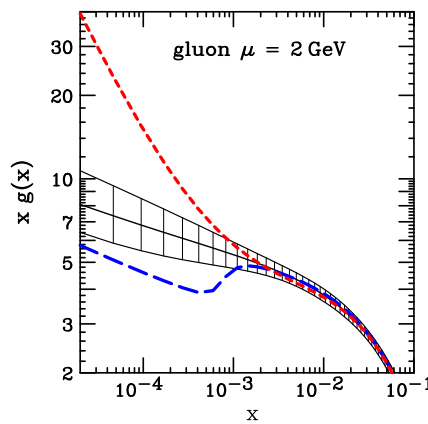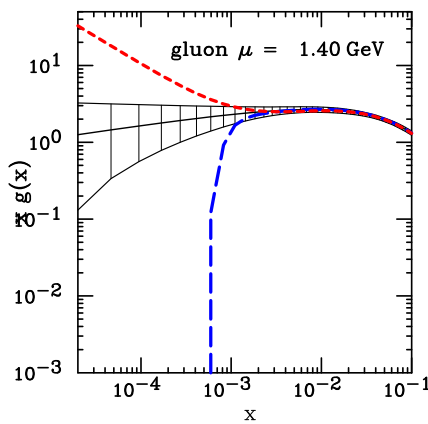
The very large uncertainties disappear at increased μ, where $g(x)$ is dominated by splitting from the known regions at higher x.

These distributions will be probed by $\mu^+\mu^-$ in LHCb.

Thanks

I have enjoyed many constructive discussions and correspondence with

CTEQ/TEA group: Huston, Lai, Nadolsky, Yuan

NNPDF group: Forte, Guffanti

MSTW group: Thorne

H1/ZEUS: Mandy Cooper-Sarkar

Statistics expert: Louis Lyons