Measuring PDFs by QCD fitting

Jon Pumplin

PDF School (DESY 20–23 October 2009)

Hadrons interact at large momentum transfer (=short distance) through their quark and gluon constituents.

Owing to the asymptotic freedom property of QCD, $\alpha_s(\mu)$ is small so most hard pp collisions at the LHC will be described by the interaction of a single quark or gluon from one of the protons with a single quark or gluon from the other.

Hence the subject of this school: we study the PDFs $f_a(x,\mu)$ which describe the "1-body" probability densities for $a = u,\ \bar{u},\ d,\ \bar{d},\ s,\ \bar{s},\ c,\ \bar{c},\ b,\ \bar{b},$ (or γ) with the spin structure and correlations integrated out.

The PDFs $f_a(x, \mu)$ for each flavor a are functions of two variables:

- x = light-cone momentum fraction
- μ = QCD factorization scale ($\approx 1/\text{distance}$), typically Q for DIS; E_T or $E_T/2$ for inclusive jet production.

However, the evolution in μ is computable at NLO or NNLO by the QCD renormalization group DGLAP equations. Hence the problem of determining the PDFs reduces to a problem of determining the x-dependence for each flavor at a chosen small scale μ_0 (e.g. $\sim 1.4\,\text{GeV}$).

The PDFs can be extracted from experiment using the requirement that they must agree with a large body of data that are dependent on them. These PDFs are then available for use in predicting production rates and backgrounds for new measurements.

Two points of view

The PDFs are a Necessary Evil — essential phenomenological tools to make perturbative calculations of signals and backgrounds at hadron colliders. It is of essential practical importance to measure the PDFS in order to make use of data from the Tevatron and LHC. Along with this comes the difficult task of assessing the uncertainty range of the answers obtained.

The PDFs are a Fundamental Measurement — an opportunity to interplay with knowledge from the nonperturbative arenas of QCD, e.g.,

- Regge theory
- Lightcone physics
- Lattice gauge

These connections have been too-much neglected in my opinion.

Even the assumption of independent flavor distributions might be improved upon.

The QCD fitting programme (brief version

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.

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 the best estimate of the true PDFs by varying the free parameters to minimize χ^2.

Theoretical basis for PDF fitting

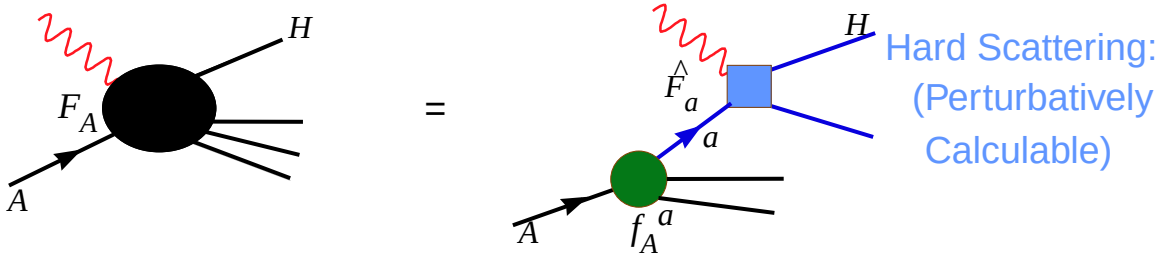
- Factorization Theorem – Short distance and long distance are separable, and PDFs are "universal," i.e., process independent.

- Asymptotic Freedom – Hard scattering is weak at short distance, and hence perturbatively calculable.

- DGLAP Evolution – Evolution in μ is perturbatively calculable, so the functions to be determined depend only on x.

Factorization Theorem

$$F_A^\lambda(x, \frac{m}{Q}, \frac{M}{Q}) = \sum_a f_A^a(x, \frac{m}{\mu}) \otimes \widehat{F}_a^\lambda(x, \frac{Q}{\mu}, \frac{M}{Q}) + \mathcal{O}((\frac{\Lambda}{Q})^2)$$



Experimental Input

Parton Distributions:
Nonperturbative parametrization at Q_0
DGLAP Evolution to Q

The PDF fitting Paradigm

1. Parameterize x-dependence of each flavor at fixed μ_0 ($= 1.4\,\text{GeV}$). Thus $f_a(x,\mu_0)$ depend on "shape parameters" $A_1,\ldots,A_N$ ($N \sim 25-30$). Example: current CTEQ gluon form

   $$x\,g(x,\mu_0) \;=\; a_0\,x^{a_1}\,(1-x)^{a_2}\,\exp(a_3\sqrt{x}+a_4x+a_5x^2)$$

   subject to number sum rule and momentum sum rule constraints.

2. Compute PDFs $f_a(x,\mu)$ at $\mu > \mu_0$ by NLO or NNLO DGLAP.

3. Compute cross sections for DIS(e,μ,ν), Drell-Yan, Inclusive Jets, W-production,... at NLO or NNLO.

4. Compute χ^2 measure of agreement between predictions and measurements:

   $$\chi^2 = \sum_i \left(\frac{\text{data}_i - \text{theory}_i}{\text{error}_i}\right)^2$$

   with appropriate generalizations to include published correlated systematic errors in the experiments, and theoretical "penalties".

PDF fitting Paradigm — continued

5. Minimize χ^2 with respect to the shape parameters $\{A_i\}$ to obtain Best Fit PDFs.

6. The PDF Uncertainty Range is assumed to be the region in $\{A_i\}$ space where χ^2 is sufficiently close to the minimum: $\chi^2 < \chi^2_{\rm min} + \Delta\chi^2$.

   The proper choice for the "tolerance condition" $\Delta\chi^2$ is a perennial hot topic for discussion. Some recent progress on it will be described later, and at PDF4LHC.

   Using the Hessian Method, the uncertainty range can be represented by $2N$ alternative PDF sets which are obtained by displacements from the minimum point in $\{A_i\}$ space along each of the directions that are defined by the eigenvectors of the Hessian matrix, where the size of each displacement is determined by $\Delta\chi^2$.

PDF fitting Paradigm — continued

7. When large values of $\Delta\chi^2$ are assumed, additional conditions are imposed by adding weights or penalties to χ^2 (CTEQ) or adjusting the lengths eigenvector displacements (MSTW) to force acceptable fits to each of the data sets over the entire uncertainty range.

8. The Best Fit, and the Uncertainty Eigenvector Sets which map out the uncertainty range, are made available for applications at http://projects.hepforge.org/lhapdf/

9. Predicted central value for a cross section of interest is obtained by calculating it using the Best Fit. The uncertainty range of the prediction is obtained by the combining predictions made using the uncertainty sets in quadrature.

Handling systematic errors

The simplest definition

$$\chi_0^2 = \sum_{i=1}^{N} \frac{(D_i - T_i)^2}{\sigma_i^2} \qquad \left\{ \begin{array}{l} D_i = \text{ data} \\ T_i = \text{ theory} \\ \sigma_i = \text{ ``expt. error''} \end{array} \right.$$

is optimal for random Gaussian errors:

$$D_i = T_i + \sigma_i r_i \qquad \text{with} \qquad P(r) = \frac{e^{-r^2/2}}{\sqrt{2\pi}}.$$

With systematic errors,

$$D_i = T_i(\mathbf{A}) + \alpha_i r_{\text{stat},i} + \sum_{k=1}^{K} r_k \beta_{ki} \, .$$

The fitting parameters are $\mathbf{A} = \{A_\lambda\}$ (theoretical model) and $\{r_k\}$ (corrections for systematic errors).

Published experimental errors:

- α_i is the 'standard deviation' of the random uncorrelated error.
- β_{ki} is the 'standard deviation' of the k^{th} (completely correlated!) systematic error on D_i.

To take into account the systematic errors, we define

$$\chi'^2(\mathbf{A}, r_k) = \sum_{i=1}^{N} \frac{\left(D_i - \sum_k r_k \beta_{ki} - T_i\right)^2}{\alpha_i^2} + \sum_k r_k^2,$$

and minimize with respect to $\{r_k\}$. The result is

$$\widehat{r}_k = \sum_{k'} \left(a^{-1}\right)_{kk'} b_{k'}, \qquad \text{(systematic shift)}$$

where

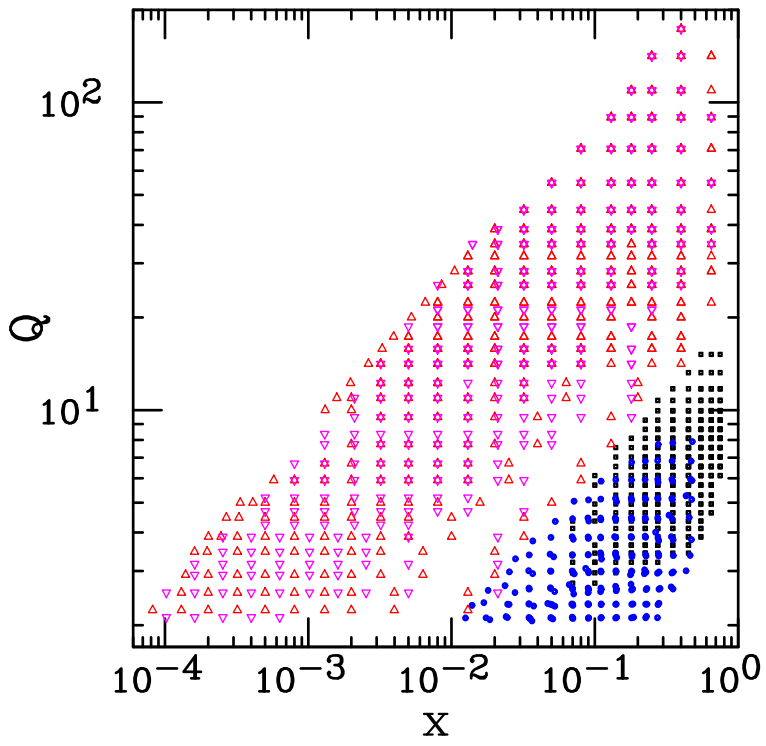
$$\begin{aligned} a_{kk'} &= \delta_{kk'} + \sum_{i=1}^{N} \frac{\beta_{ki}\beta_{k'i}}{\alpha_i^2} \\ b_k &= \sum_{i=1}^{N} \frac{\beta_{ki}\left(D_i - T_i\right)}{\alpha_i^2}. \end{aligned}$$

The $\widehat{r}_k$'s depend on the PDF model parameters $\mathbf{A}$. We can solve for them explicitly since the dependence is quadratic.

We then minimize the remaining $\chi^2(\mathbf{A})$ with respect to the model parameters $\mathbf{A} = \{A_\lambda\}$.

- $\{a_\lambda\}$ determine $f_i(x, Q_0^2)$.
- $\{\widehat{r}_k\}$ are are the optimal "corrections" for systematic errors; i.e., systematic shifts to be applied to the data points to bring the data from different experiments into compatibility, within the framework of the theoretical model.
- A similar treatment could be used for parametrized systematic errors in the theory — e.g. from scale choices.

Kinematic region of ep and μp data



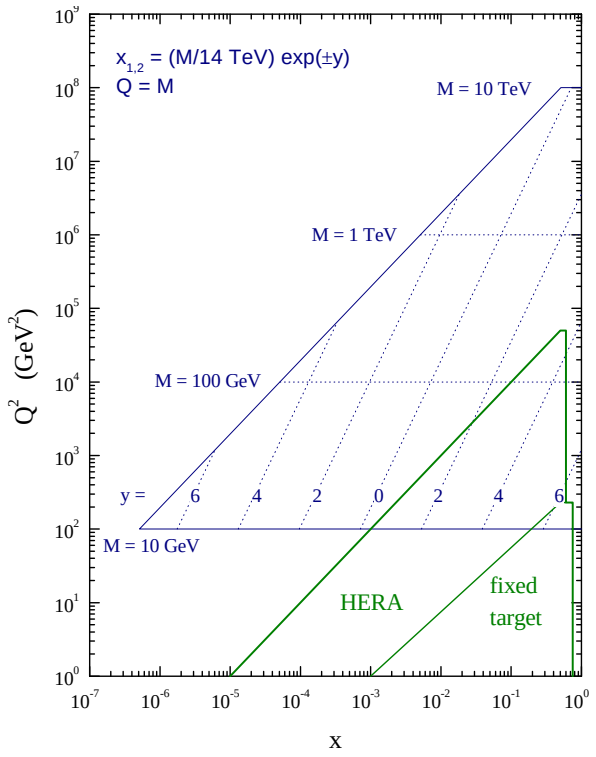
$ep \to eX$ (H1 = Δ, ZEUS = ∇)
$\mu p \to \mu X$ (BCDMS= box, NMC = $\circ$)

Drell-Yan data, neutrino DIS, and Tevatron W and Z data are also very important for differentiating among different flavors.

Tevatron inclusive jet data are very important for constraining the gluon distribution.

HERA II (not yet included in CTEQ fits), more Tevatron run II, and eventually the LHC will dramatically extend the range and accuracy.

Kinematic Map for LHC



LHC will explore new territory in x and μ $(= Q)$.

DGLAP evolution at large μ should be very reliable, so the PDFs needed to calculate the production of new heavy objects are in pretty good shape.

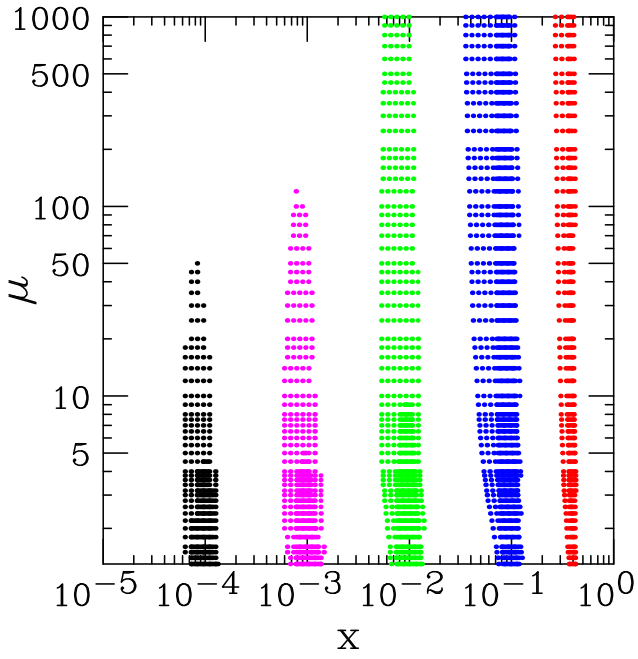
Significant new territory will come into play at small x when forward Z^0 or lower-mass $\ell^+\ell^-$ pairs are measured.

Large x is important: the difference between central collisions at $x = 0.20$ vs. $x = 0.28$ is the same as the difference between running LHC at $\sqrt{s} = 7$ vs. $\sqrt{s} = 14$ TeV!

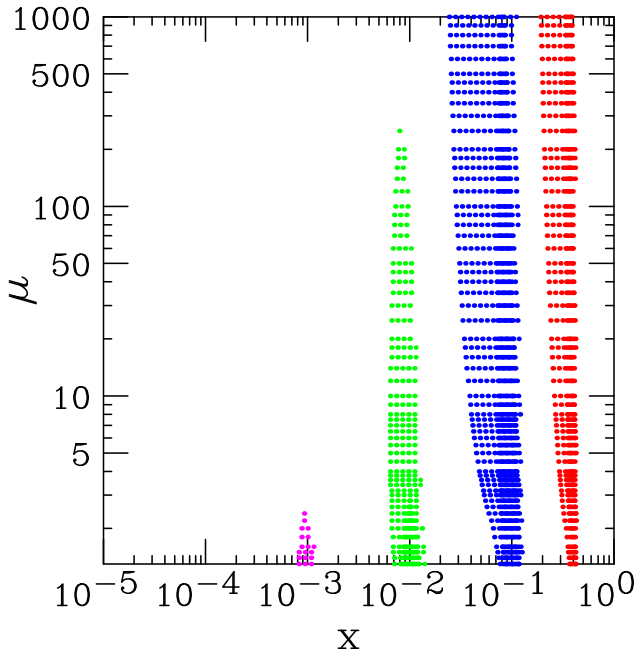
At the same time, one of the delights at the LHC is that it will the allow the study of PDFs at very small x — where interesting effects like BFKL are predicted — since the large s allows x_1 or x_2 to be very small while M is large enough for a perturbative calculation to be reliable, in accord with $s = x_1\, x_2\, M^2$.

Evolutionary influences of quarks

Regions of PDF change $>0.2\%$ (solid) or $>0.05\%$ (dotted) produced by a 1% change at $Q_0 = 1.3\,\mathrm{GeV}$ in a narrow band of x:



$\bar{d} + \bar{u}$

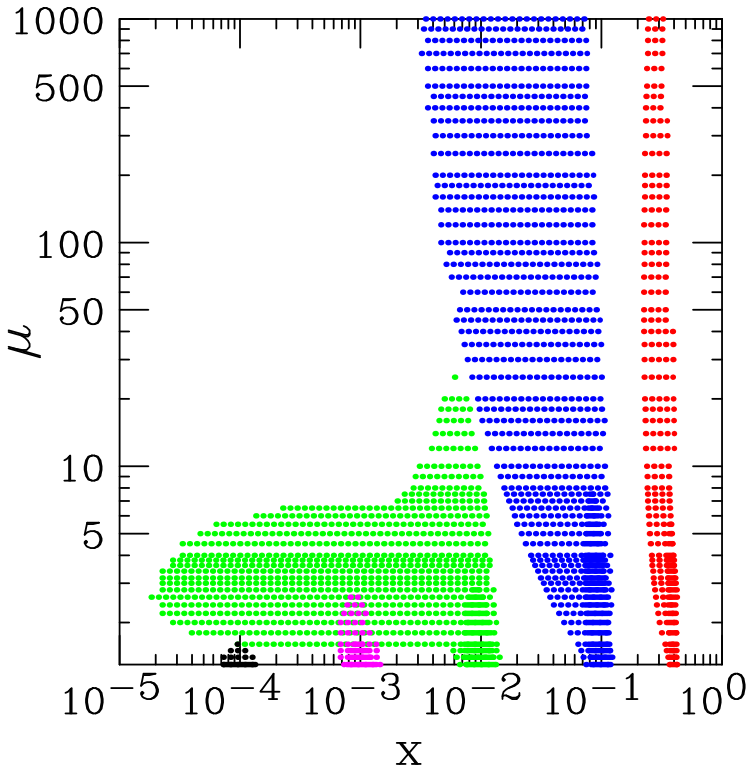


u_v

- Valence quarks are unimportant at small x as expected.
- Quark evolution is mostly at constant x, with a bit of feed-down toward smaller x.

Evolutionary influences of gluon

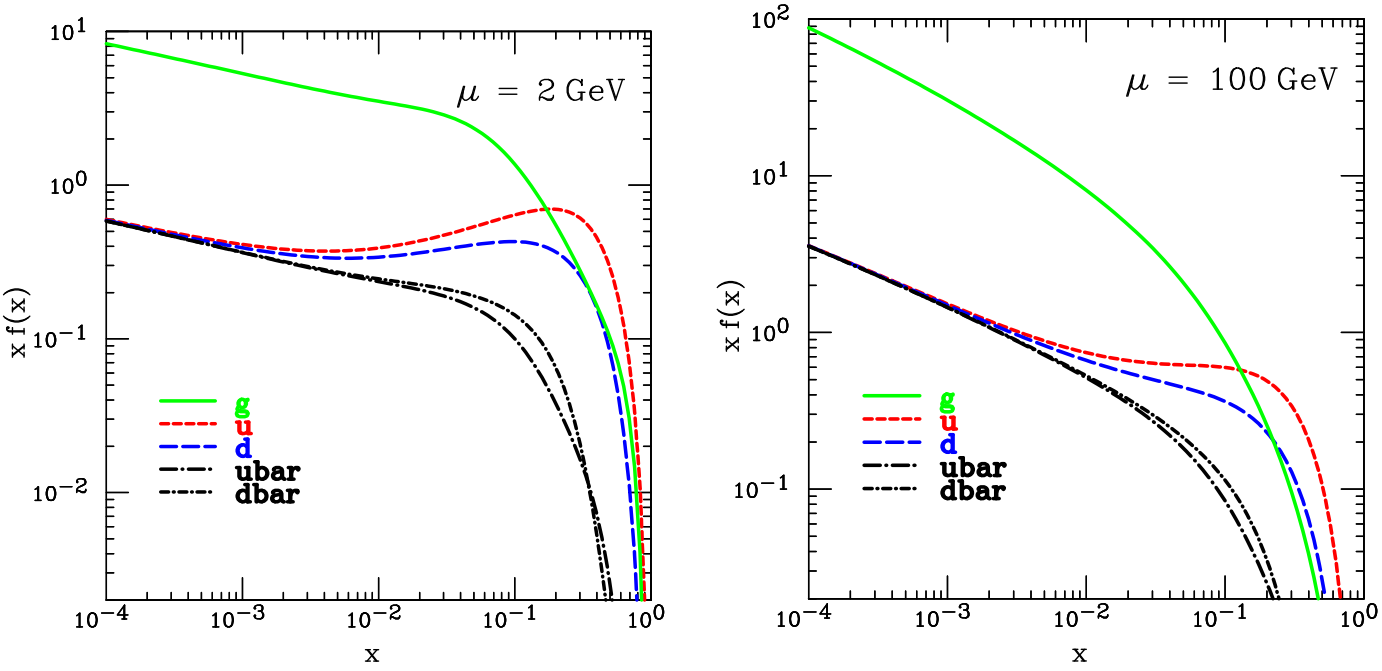
Regions of PDF change $>0.2\%$ (solid) or $>0.05\%$ (dotted) caused by a 1% change in gluon at $Q_0 = 1.3\,\mathrm{GeV}$ in a narrow band of x:



- Influence of input $g(x)$ spreads in x much more than quarks.

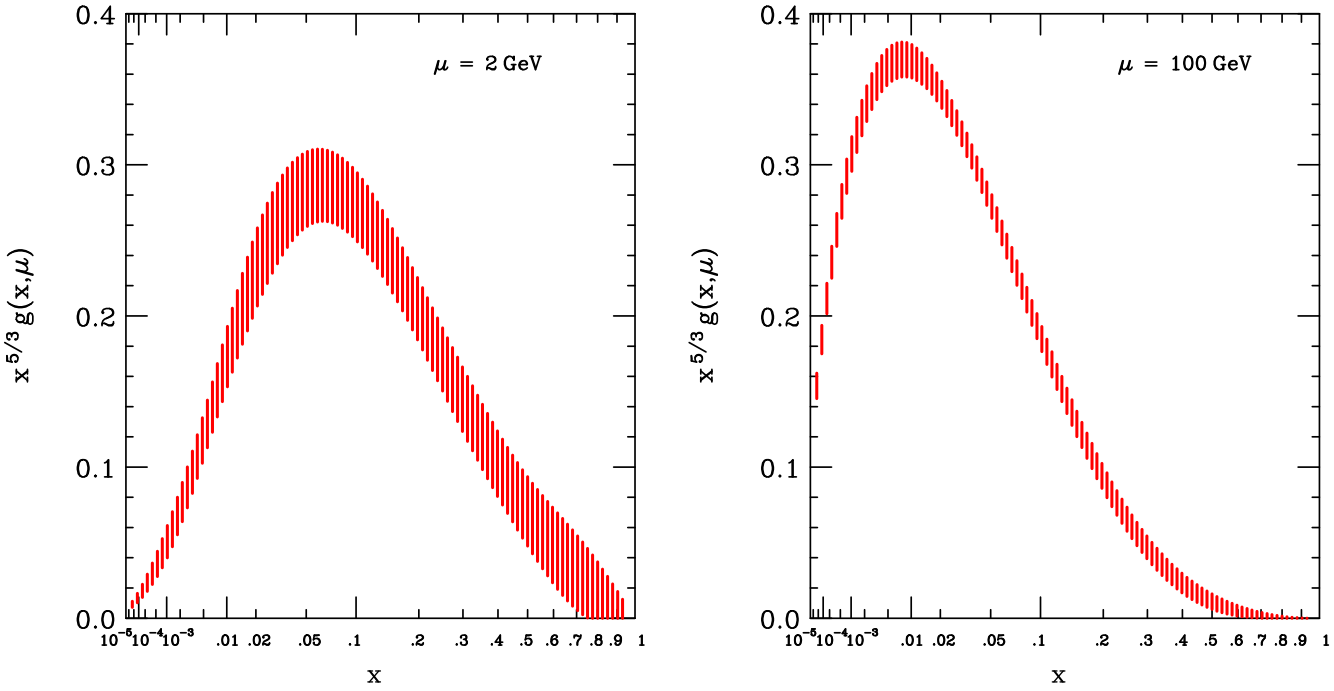
- Small-x gluon at $Q_0 = 1.3\,\mathrm{GeV}$ has little direct influence
  $\Rightarrow$ gluons at moderate and high Q are radiatively generated.

PDF results at $\mu = 2\,\text{GeV}$ and $100\,\text{GeV}$



- Valence quarks dominate for $x \to 1$
- $u > d$ because $N_u = 2,\ N_d = 1$.
- Gluon dominates for $x \to 0$, especially at large μ.
- $\bar{u}$ and $\bar{d}$ are different — they even cross over.
- $u = \bar{u} = d = \bar{d}$ at $x \to 0$ is imposed in the parametrization, but is consistent with the data: dropping this condition allows very little reduction in χ^2.

Uncertainty Results (Gluon)

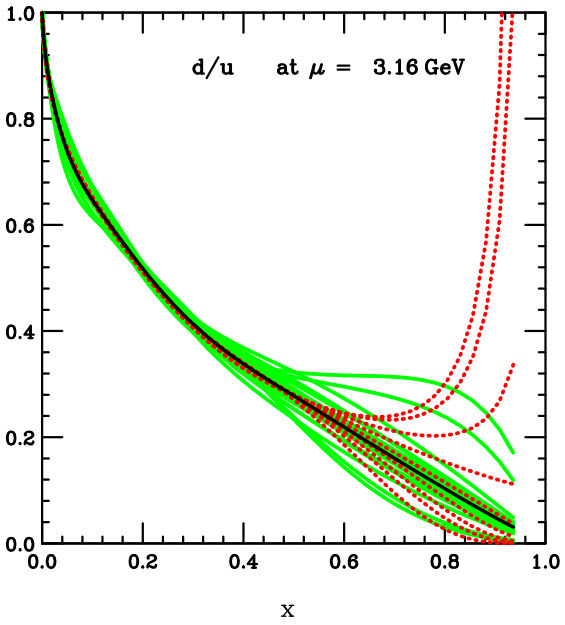


$\Delta\chi^2 = 100$ uncertainty bands. Horizontal axis is $x^{1/3}$ in order to show a wide range of x.

Vertical axis is weighted by $x^{5/3}$ to make the area under the curve proportional to the momentum fraction carried by gluon. That momentum fraction is strongly constrained by DIS data, so the envelope itself is not an allowed PDF — e.g., if $g(x)$ is larger than the central value at $x \approx 0.5$, it will be smaller than the central value at $x \approx 0.05$.

"Convergent evolution": the uncertainty is much smaller at $\mu = 100\,\text{GeV}$.

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



Black: CTEQ6.5 central fit

Green: 40 CTEQ6.5 eigenvector uncertainty sets

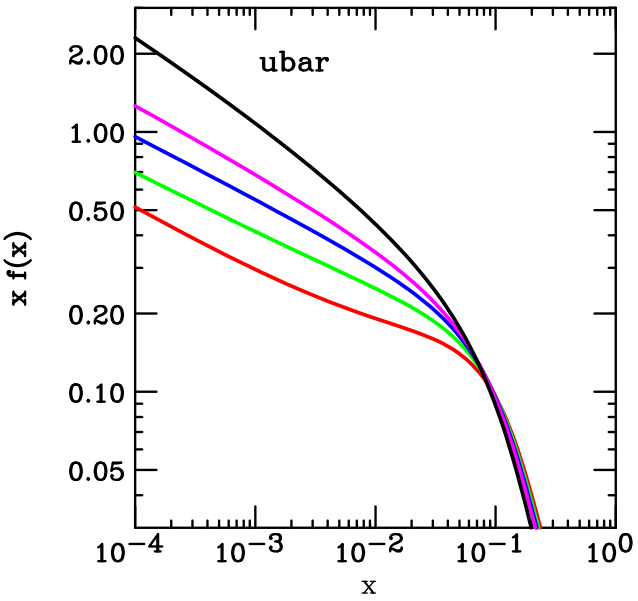
Red: results from equally-acceptable alternative parametrizations

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 $x > 0.8$.

Regge behavior of $\bar{u}$

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:



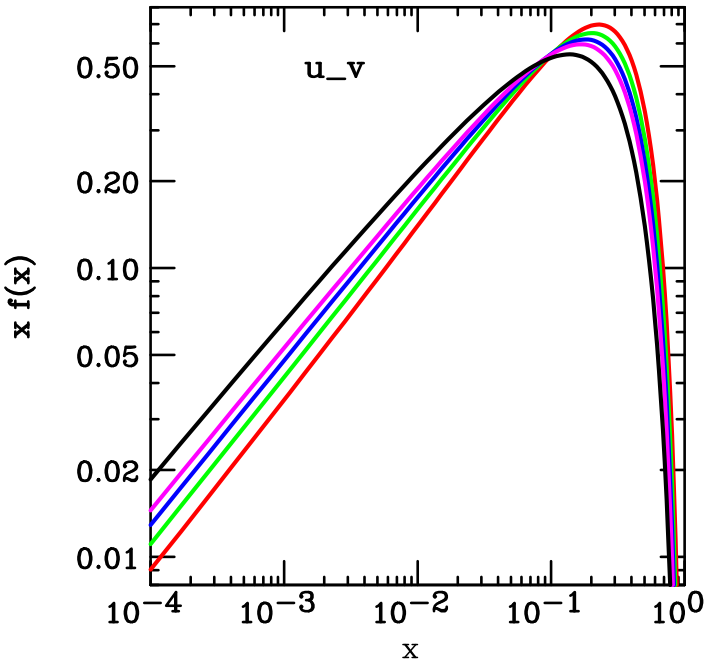
Red/Green/Blue/Magenta/Black:
$\mu = 1.3/2.0/3.2/5.0/20\,\text{GeV}$.

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 small compared to the uncertainty of estimates from strong-interaction phenomenology.

Regge behavior of $u_v \equiv u - \bar{u}$

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:



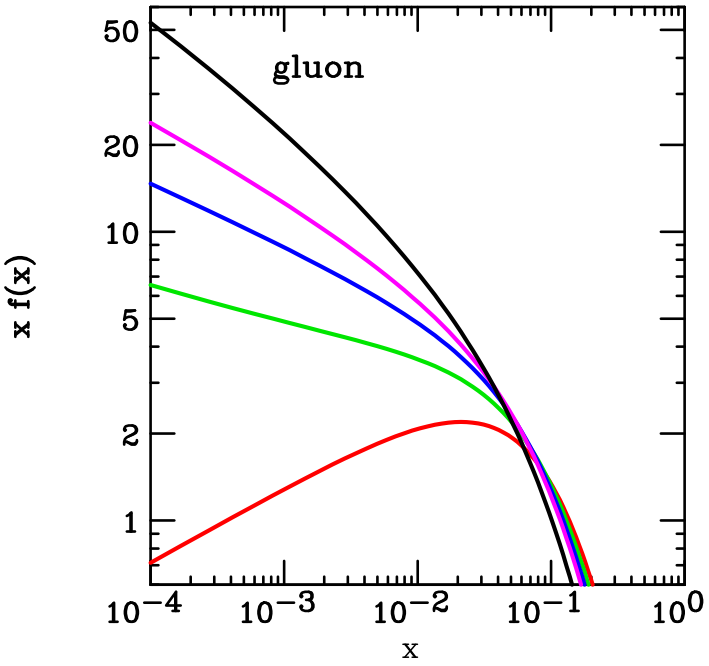
where Red/Green/Blue/Magenta/Black:
$\mu = 1.3/2.0/3.2/5.0/20\,\text{GeV}$.

Again the observed slope value a_1 is consistent with expectations from Regge theory, which supports the choice of 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.

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:

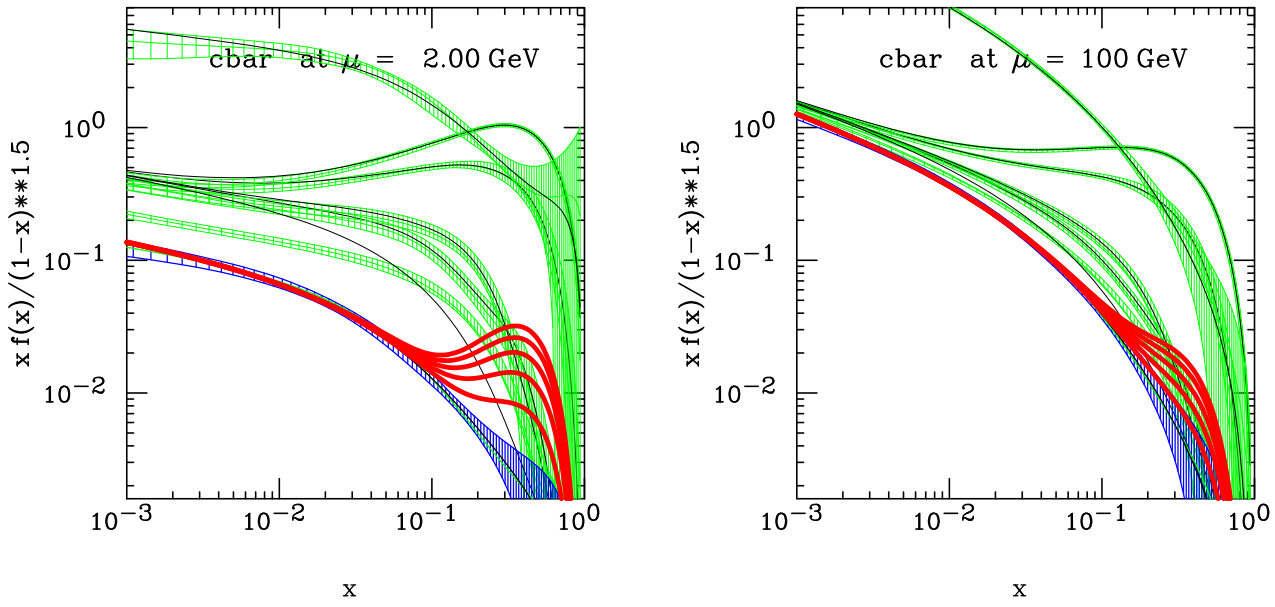


where Red/Green/Blue/Magenta/Black:
$\mu = 1.3/2.0/3.2/5.0/20\,\mathrm{GeV}$.

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

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

PDFs with Intrinsic Charm



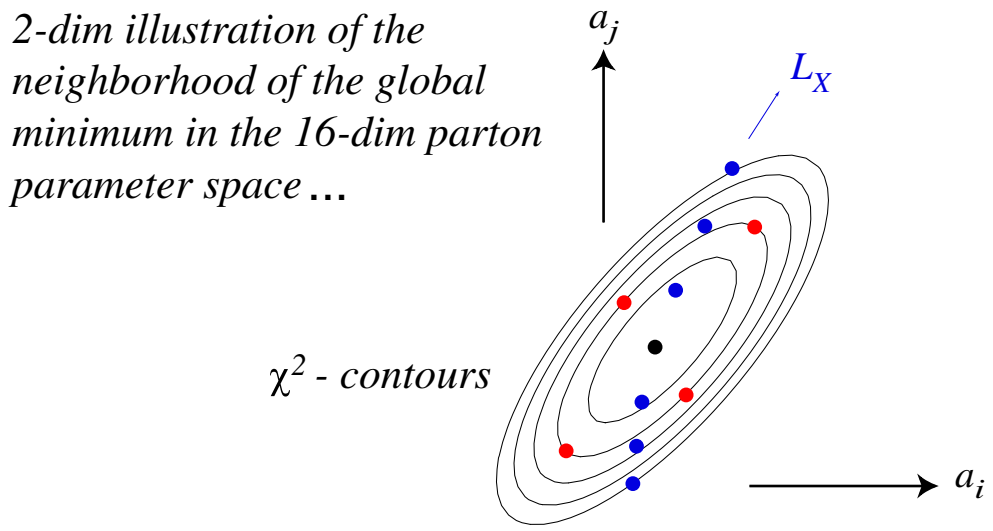
Green: g, u, d, $\bar{u}$, $\bar{d}$, $s = \bar{s}$ CTEQ6.5

Blue: Charm from gluon splitting

Red: Intrinsic Charm using form of Brodsky et al. at $\mu_0 = 1.3\,\text{GeV}$, normalized to probability 0.5%, 1.0%, 1.5%, 2.0%, 2.5% for $c\bar{c}$.

- Typical estimate 1.0% according to fans of intrinsic charm; $> 2.5\%$ ruled out by Global Fit.
- IC could be "large" $(\bar{c} > \bar{u}, \bar{d})$ for $x > 0.2$.

Uncertainties: Lagrange Multiplier method



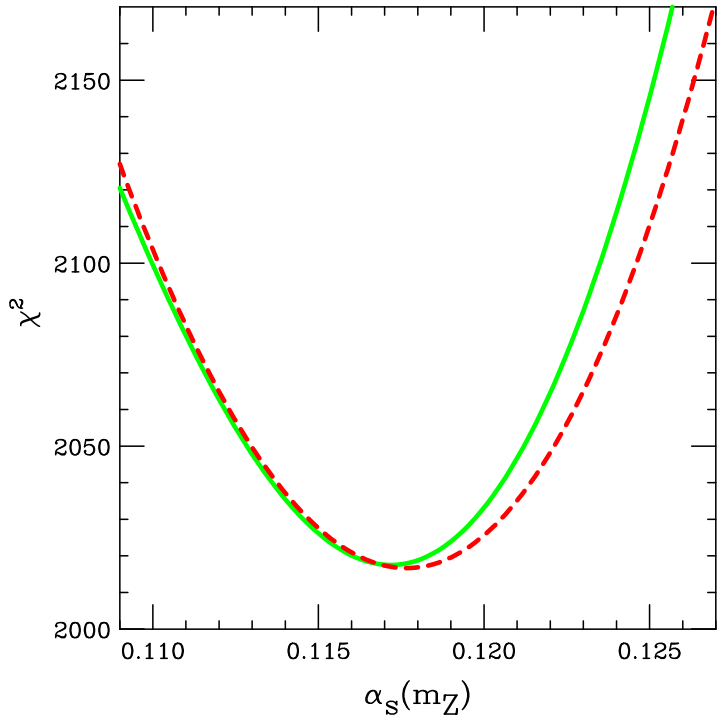
Track χ^2 as function of a physical quantity of interest.

- Easy example: plot χ^2 for the global fit as a function of $\alpha_s(m_Z)$, to get a measurement of the strong coupling based on PDF fits.

- LM example: plot χ^2 for the global fit as a function of the predicted $t\bar{t}$ cross section at LHC, to get a clean estimate of the PDF uncertainty in that quantity. The plot can be generated by minimizing $\chi^2 + \lambda\,\sigma_{t\bar{t}}$ for a series of different values of the LM parameter λ.

- This method can be used to generate PDF sets that predict the extreme values of $\sigma_{t\bar{t}}$, or σ_W, or $\langle y \rangle$ for rapidity distribution of W; or . . .

Dependence of fit χ^2 on $\alpha_s(m_Z)$



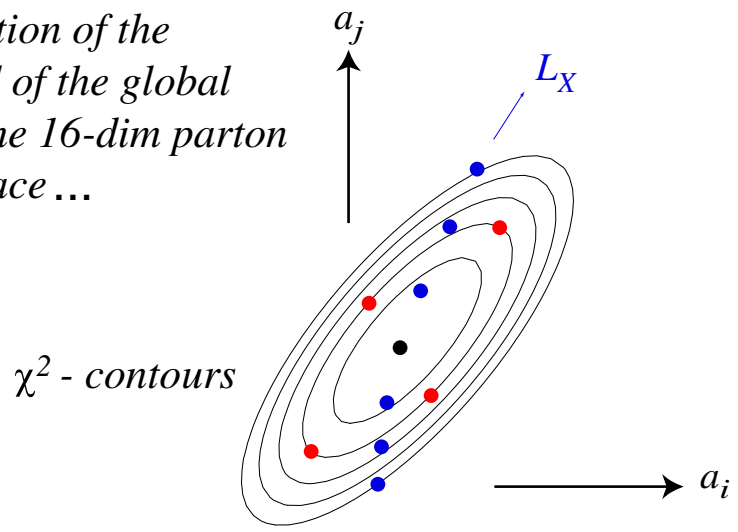
Have two different curves because have tried two different functional forms for $\alpha_s(\mu)$, which are identical at NLO. The difference between them is an unavoidable systematic error.

Minima of χ^2 at $\alpha_s(m_Z) = 0.1172$ and 0.1176 are close to the world average.

Choosing an appropriate $\Delta\chi^2$ tolerance range turns this into a measurement of $\alpha_s(m_Z)$. But the uncertainty of that measurement is larger than the uncertainty of the world average – which is dominated by LEP data.

Uncertainty methods: Hessian

2-dim illustration of the neighborhood of the global minimum in the 16-dim parton parameter space ...



In the neighborhood of the minimum, χ^2 has a quadratic form

$$\chi^2 = \chi^2_{\rm min} + \sum_{ij} H_{ij} \, (A_i - A_i^{(0)}) \, (A_j - A_j^{(0)}) \, .$$

It is convenient to put this in a diagonal form by using the eigenvectors of H:

$$\chi^2 = \chi^2_{\rm min} + \sum_i z_i^2$$

where

$$A_i = A_i^{(0)} + \sum_j w_{ij} \, z_j \, .$$

Hessian method – continued

In the Hessian method, χ^2 is diagonalized in the neighborhood of the minimum:

$$\chi^2 = \chi^2_{\text{min}} + \sum_i z_i^2 \, .$$

The uncertainty range is then described by PDF eigenvector sets defined by

$$(z_1, z_2, z_2, \ldots) = \begin{cases} (+T, 0, 0, \ldots) \\ (-T, 0, 0, \ldots) \\ (0, +T, 0, \ldots) \\ (0, -T, 0, \ldots) \\ \ldots \end{cases} .$$

According to the quadratic approximation, $T = \sqrt{\Delta\chi^2}$. In practice, T is adjusted separately for each eigenvector set to produce the desired $\Delta\chi^2$.

Hessian method – continued

The $2N$ PDF eigenvector sets can be used to compute the PDF uncertainty for any prediction F. A symmetric form for the uncertainty is given by

$$\Delta F = \frac{1}{2}\sqrt{\sum_i \left(F(S_i^{(+)}) - F(S_i^{(-)})\right)^2}$$

where $S_i^{(+)}$ and $S_i^{(-)}$ are the PDF sets that are displaced along the eigenvector direction i. A more accurate method is to compute asymmetric limits:

$$\Delta F = \sqrt{\sum_{i,\pm} \left(F(S_i^{(\pm)}) - F(\text{BestFit})\right)^2}$$

where the sum includes only positive displacements to calculate the upper limit of F and only negative displacements for the lower limit.

Parton distributions published by the CTEQ group have $\Delta\chi^2$ chosen to estimate 90% confidence limits. This is done because non-quadratic behavior of χ^2 associated with "flat" directions causes the 90% confidence limits to not be as broad as one would estimated by multiplying the traditional 68.2% ("1σ") limits by the factor 1.64 that would be predicted by standard Gaussian statistics.

Iterative technique in the Hessian method

In the quadratic approximation (Taylor series to second order),

$$\chi^2 = \chi^2_{\mathsf{min}} + \sum_{ij} H_{ij}\,(A_i - A_i^{(0)})\,(A_j - A_j^{(0)})$$

where

$$H_{ij} = \frac{\partial^2 \chi^2}{\partial A_i\,\partial A_j}\,.$$

Formally, this can be put into a diagonal form

$$\chi^2 = \chi^2_{\mathsf{min}} + \sum_i z_i^2$$

by a linear transformation

$$A_i = A_i^{(0)} + \sum_j w_{ij}\,z_j\,,$$

where the transformation matrix $\mathbf{w}$ is constructed from the eigenvectors of $\mathbf{H}$.

Iterative technique – continued

In practice, it is not so simple to obtain

$$\chi^2 = \chi^2_{\mathrm{min}} + \sum_i z_i^2 \,,$$

because the curvature of χ^2 as a function of the displacement from the minimum in the space of fitting parameters

$$D = \sqrt{\sum_i (A_i - A_i^{(0)})^2}$$

varies over orders of magnitude among different directions. This causes non-quadratic behavior of χ^2 to spoil the simple calculation of H by finite differences.

This difficulty is overcome by an iterative method in which new coordinates obtained using the eigenvectors of the Hessian are treated as old ones and the method is repeated until it converges. By the end of the iteration, χ^2 is probed in all directions at the appropriate scale of $\Delta\chi^2$.

Measuring internal consistency of the fit

Partition the data into two subsets:

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

where subset S can, for example, be chosen as

- any single experiment
- 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 (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^2_S and $\chi^2_{\overline{S}}$ 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

$$\chi^2 = \chi^2_0 + \sum_{i=1}^{N} z_i^2$$

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

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

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 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	$\sum_i \gamma_i$	(γ_1, σ_1), (γ_2, σ_2), ...
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)

Measurements in a recent PDF fit

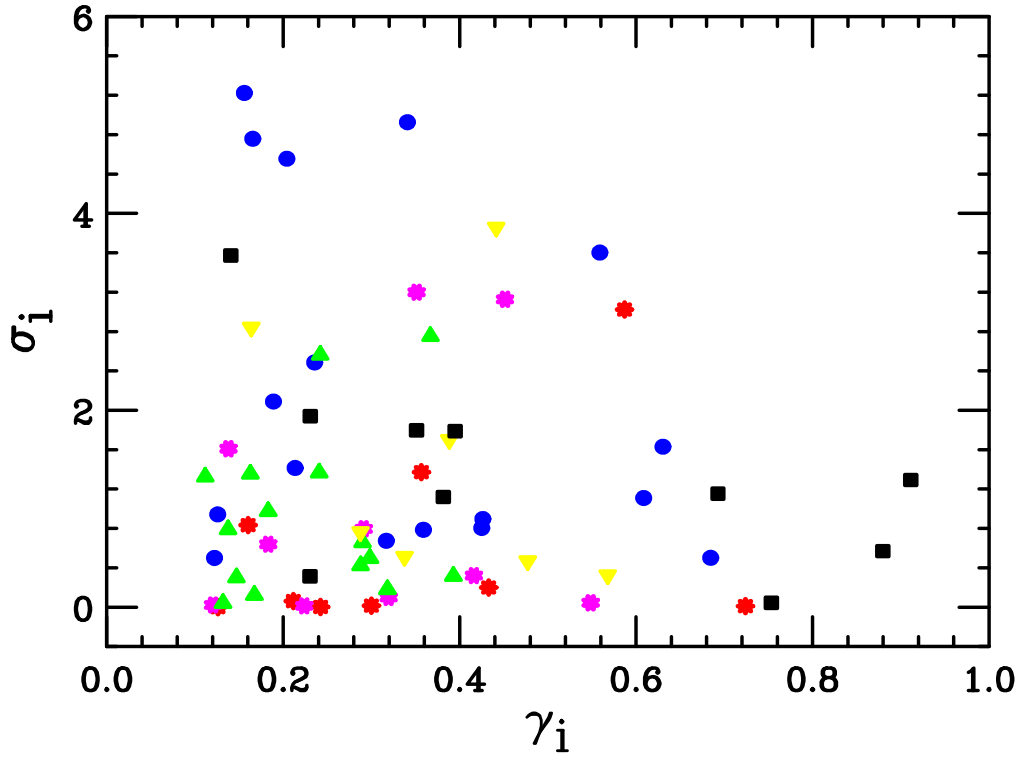


Figure showing the results in the table.

ep (daisy);

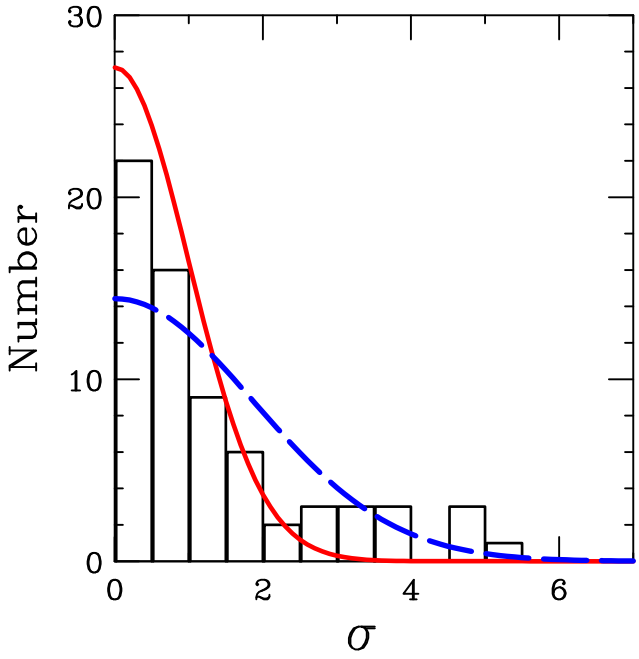
$\mu p, \mu d$ (○);

pp, pd, pCu (box);

$\bar{p}p$ (∇);

νA (Δ).

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

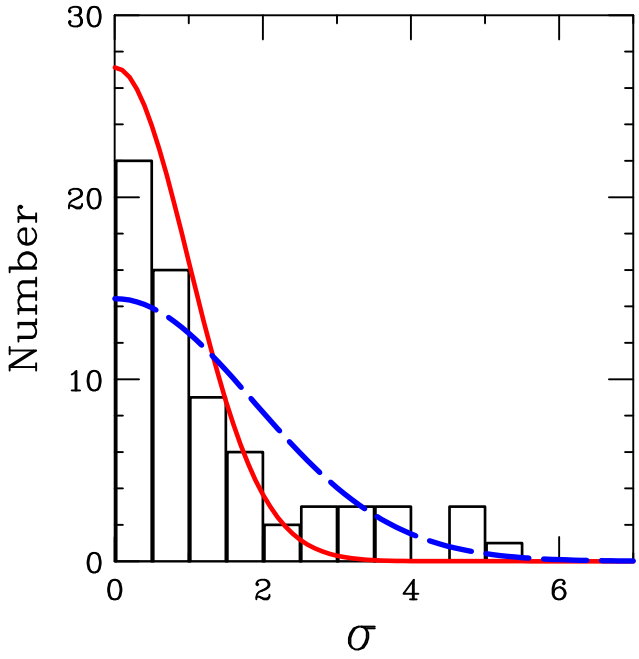
$$\frac{dP}{d\sigma} = \sqrt{\frac{2}{\pi}}\,\exp(-\sigma^2/2)\,.$$

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" $\Delta\chi^2 \approx 10$ 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.

Studies relating to the choice of $\Delta\chi^2$

It is important to know if we are underestimating or overestimating the PDF uncertainties.

For properties that we have little information, the Hessian method generally underestimates uncertainties, because completely unknown behavior requires parametrizations assumptions for convergence. However, fortunately, this is generally not too important because the properties that present-day PDF data are insensitive to are also generally unimportant for LHC phenomenology. Example: $u(x) - \bar{u}(x)$ at small x is poorly known, and also unimportant.

Will discuss this further in the PDF4LHC workshop.

Sum rule tests

A direct test of the treatment of uncertainties can be made by treating the valence quark numbers and/or the total partonic momentum as free parameters in the fit, since for these cases we know the true answer exactly:

$$N_u = \int_0^1 [u(x) - \bar{u}(x)]\, dx \qquad \text{SM value} = 2$$

$$N_d = \int_0^1 [d(x) - \bar{d}(x)]\, dx \qquad \text{SM value} = 1$$

$$m = \sum_a \int_0^1 f_a(x)\, x\, dx \qquad \text{SM value} = 1$$

(These are scale-independent under DGLAP.)

If m only is set free, it moves to 1.025 with a reduction of 5 in χ^2.

If N_u and N_d are set free, they run to 2.6 and 1.3 with a reduction of 10 in χ^2.
(N_u and N_d are not well determined in the global fit, because the data are insensitive to $u(x) - \bar{u}(x)$ and $d(x) - \bar{d}(x)$ at small x, where these quantities are much smaller than $\bar{u}(x)$ and $\bar{d}(x)$.)

Sum rule tests – continued

If all three are set free, the fit prefers

$$N_u = 2.8 \qquad N_d = 1.5 \qquad m = 1.03$$

with χ^2 lower by 15.

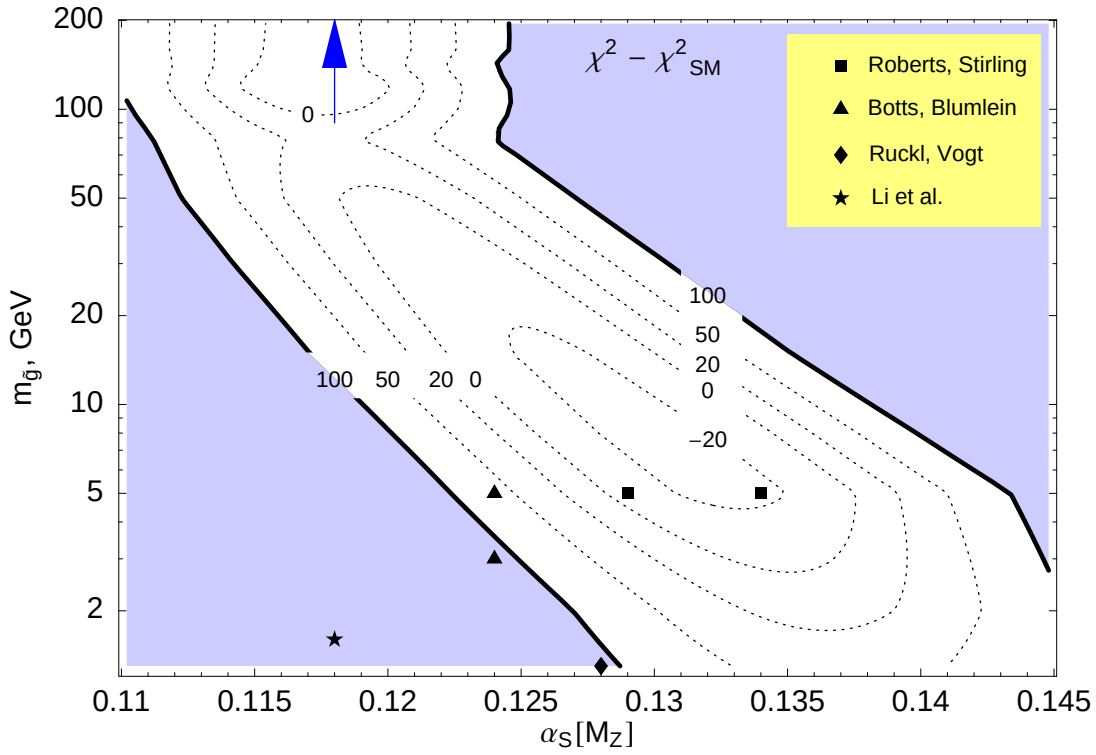
Hence we do not want to think of $\Delta\chi^2 = 15$ as a significant improvement — at least for the prediction of quantities that are poorly determined.

Uncertainty example: Light Gluino

(E. Berger, P. Nadolsky, F. Olness and J. P., Phys. Rev. D **71**, 014007 (2005)

Hypothesizing a gluino of mass $\sim 10\,\mathrm{GeV}$ improved a previous global fit by ~ 25 units in χ^2.

We took this an intriguing possible hint for plausible New Physics. But you would be crazy to consult a statistical table of χ^2 probabilities and declare it inescapable.

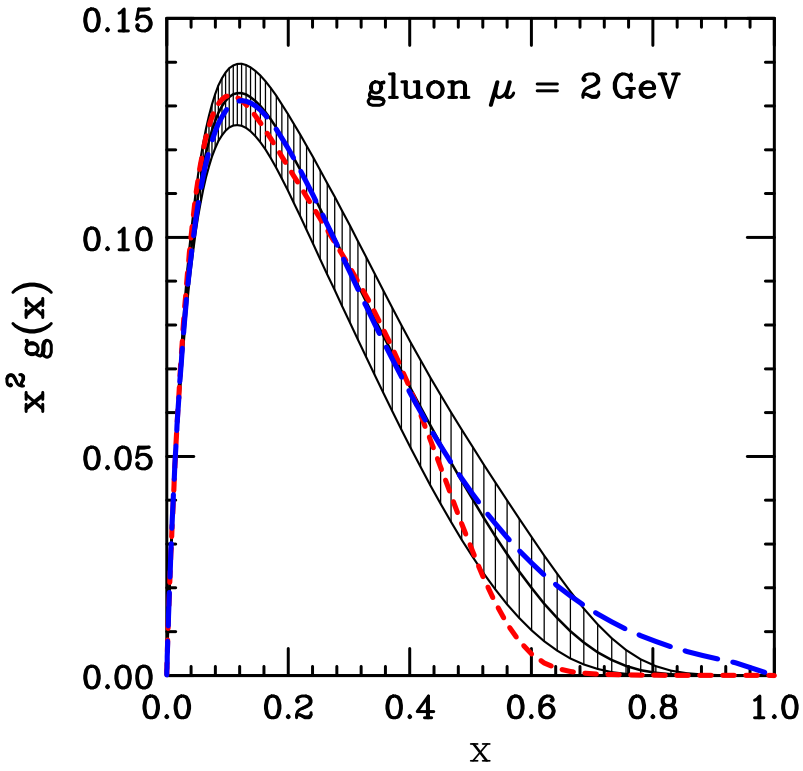


Parametrization dependence at large x

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.

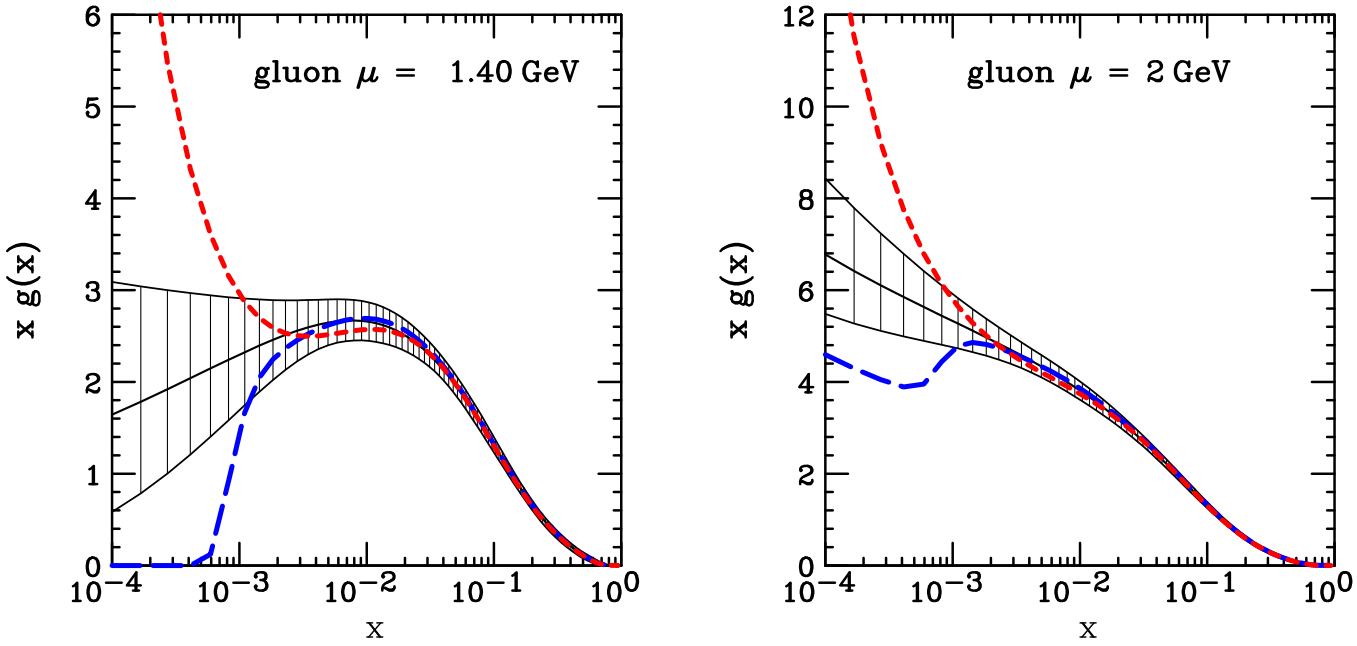


Curves show $a_2 = 54$ (which produces $\Delta\chi^2 = 10$) and $a_2 = 0$ (which requires almost zero $\Delta\chi^2$)

Non-perturbative theory constraints are important at large x.

Parametrization dependence at small x

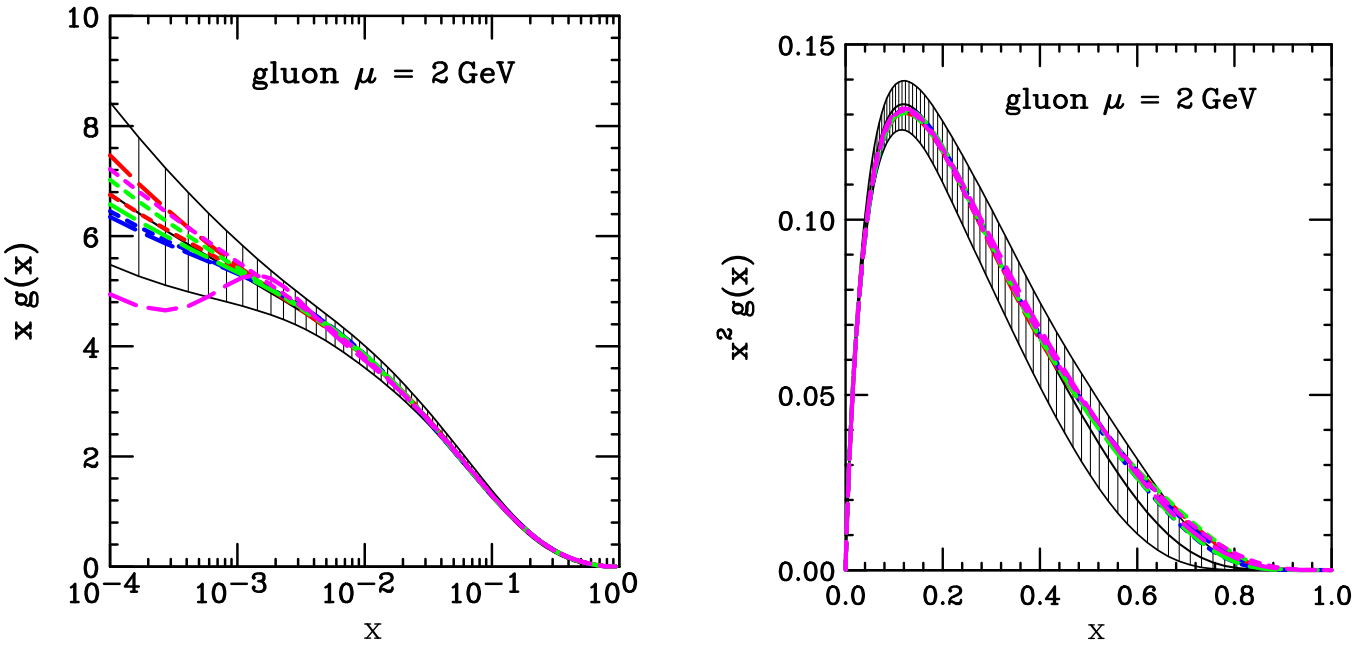
Figure shows $\Delta\chi^2 = 10$ uncertainties. Curves show results of alternative parametrizations that enhance or suppress the gluon at small x



In a region where the data provide little constraint, the true uncertainty is much larger than $\Delta\chi^2$ shows because of parametrization dependence.

There is very little constraint on gluon at small x for low scale μ; but at higher scales, the small-x gluon is generated mainly by DGLAP evolution down from higher x, so the uncertainties – e.g. for heavy objects created from gluons at LHC – are not so large.

Parametrization dep. at intermediate x

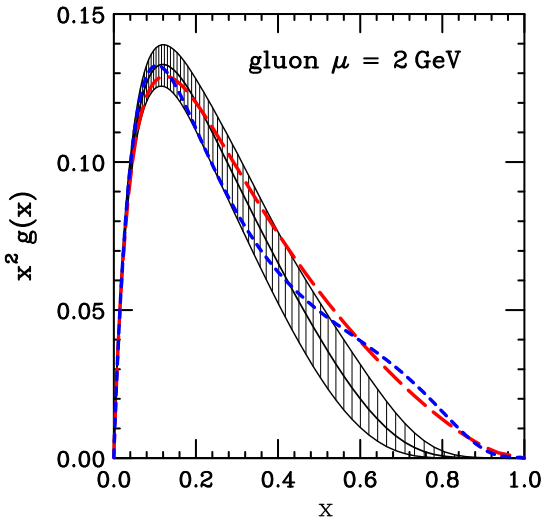


Figures show gluon uncertainty at $\Delta\chi^2 = 10$.

Curves show results from alternative parametrizations with up to 8 more parameters added.

The added freedom reduces χ^2 by as much as 10 – 15, but the change in the gluon distribution is small except at extreme x — where we already knew there was 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.

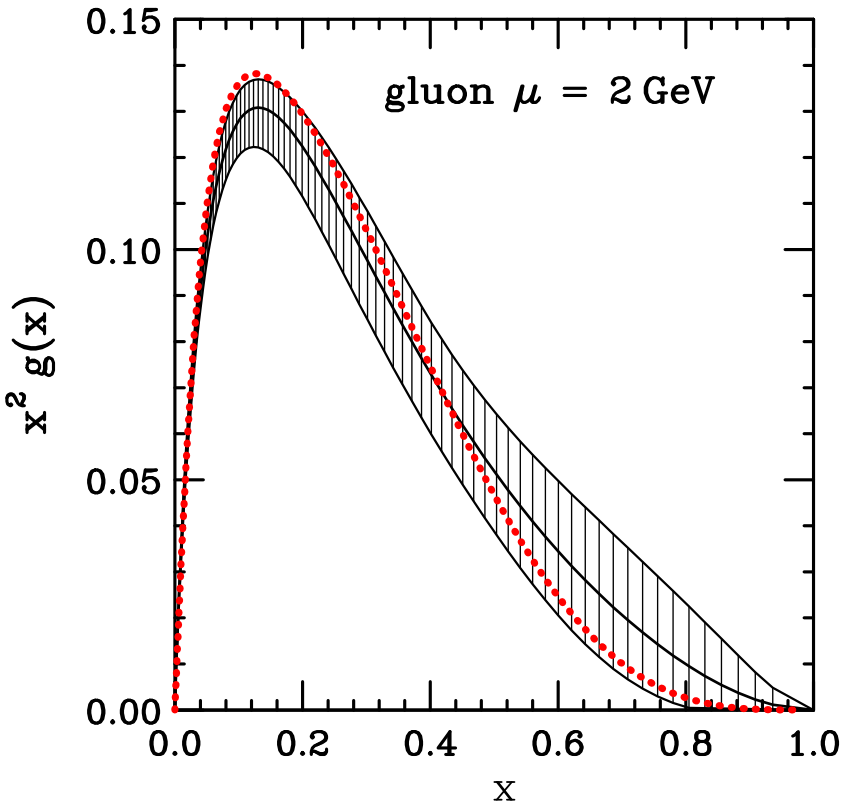
CT09 central fit: used both run I and run II jet data; different weights for different experiments.

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

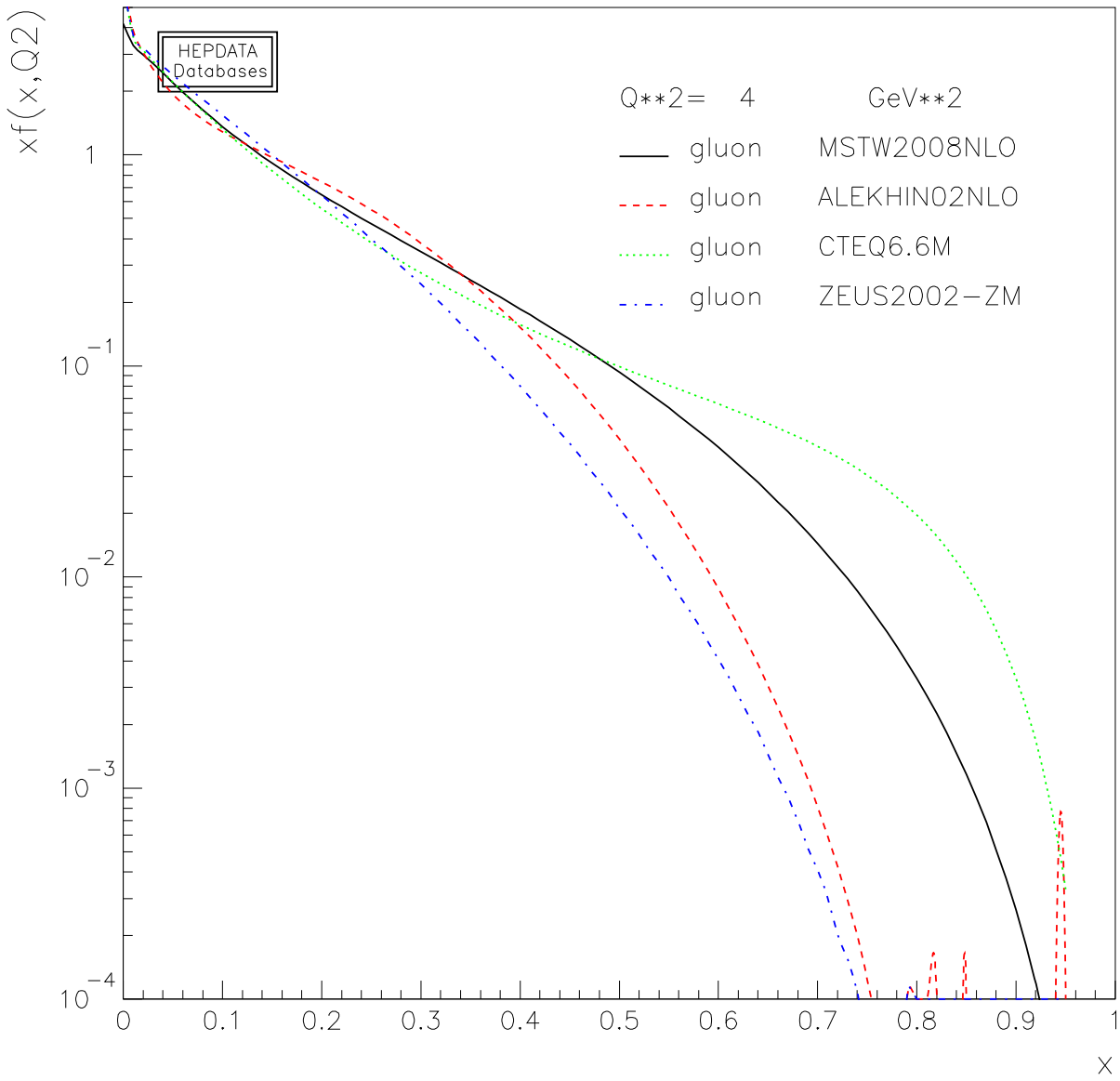
Again it is clear that $\Delta\chi^2 = 1$ for 68% confidence would be overly optimistic.

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

Conclusion

- There is an active ongoing program to determine the PDFs that are needed for LHC.
- As befits a critical mission component, there are several groups working independently on the problem.
- Estimating the size of the uncertainties caused by systematic errors in the theory is a current hot topic in which further progress can be expected.

To illustrate how easy it is to access the PDFs, a final figure was obtained by a few clicks on `http://durpdg.dur.ac.uk/hepdata/pdf3.html`


HEPDATA
Databases
Q**2= 4 GeV**2
gluon MSTW2008NLO
gluon ALEKHIN02NLO
gluon CTEQ6.6M
gluon ZEUS2002−ZM
xf(x,Q2)
1
10^−1
10^−2
10^−3
10^−4
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
X