MSU Approach to Parton Distributions and Uncertainties

Jon Pumplin

Les Houches Workshop – May 2001

Outline:

1. CTEQ/MSU Global Analysis
2. Overview of MSU uncertainty studies
3. Lagrange Multiplier method
4. Hessian Matrix method
5. Measures of consistency
6. Results on gluon on quark distributions
7. Results on W and Z production
8. Interface to Monte Carlos

References:

- Pumplin, Stump, Tung, Huston, Brock, et al. hep-ph/0008191, 0101032, 0101051
- J. Collins & JP, "Tests of goodness of fit to multiple data sets" hep-ph/AnyDayNow
- JP, "Non-Gaussian Statistics and Effective Chi-Squared" hep-ph/AnyDayNow

CTEQ Global analysis

Experimental Input:

- Include all relevant data on equal footing: ≈ 1400 points with $Q > 2\,\text{GeV}$ from e, μ, ν DIS; lepton pair production (DY); lepton asymmetry in W production; high p_T inclusive jets; $\alpha_s(M_Z)$ from LEP

Theoretical Input:

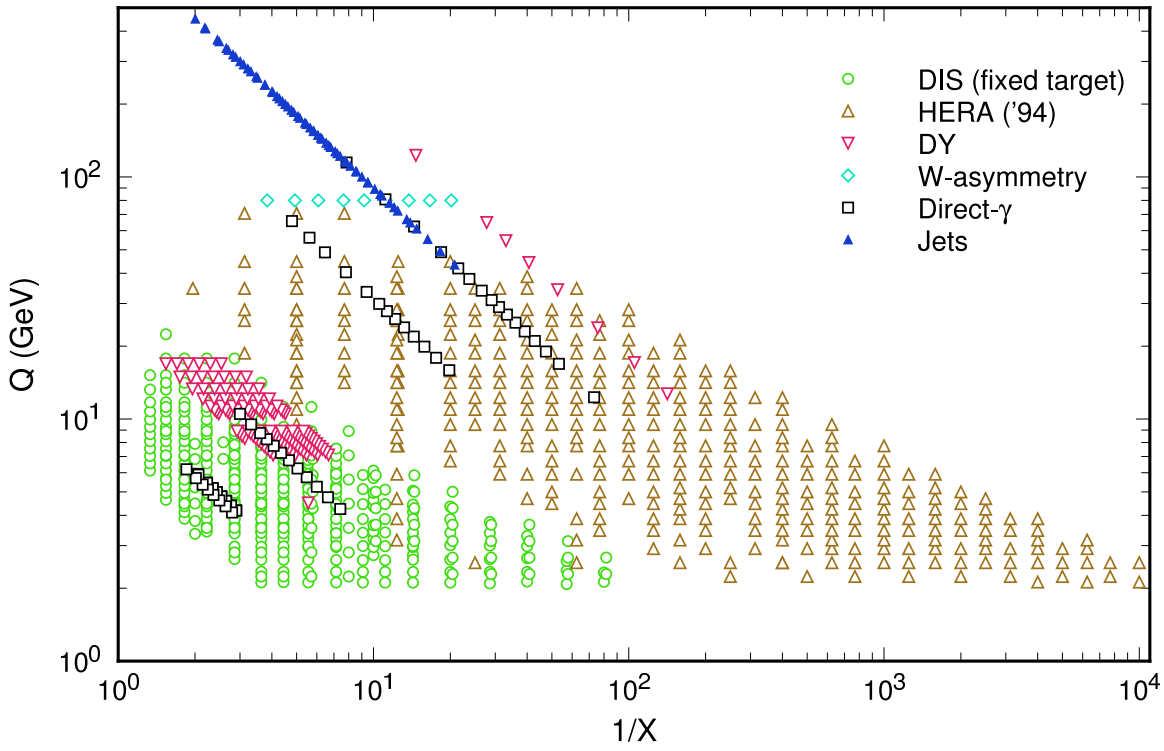
- NLO QCD evolution and hard scattering
- Parametrize: $A_0\, x^{A_1}\,(1-x)^{A_2}\,(1+A_3 x^{A_4})$ at Q_0
- $s = \bar{s} = 0.4\,(\bar{u}+\bar{d})/2$ at Q_0; no intrinsic b or c

Effective $\chi^2_{\text{global}} = \sum \chi^2_n$ summed over experiments:

$$\chi^2_n = \left(\frac{1-\mathcal{N}_n}{\sigma^N_n}\right)^2 + w_n \sum_I \left(\frac{\mathcal{N}_n\, D_{nI} - T_{nI}}{\mathcal{N}_n\, \sigma^D_{nI}}\right)^2$$

- Normalization factor $\mathcal{N}_n$ is prototype for including correlated systematic errors – moving toward full error correlation matrix where available
- Find Best Fit PDFs by minimizing with respect to the parameters.
- Estimate uncertainty as region of parameter space where $\chi^2 < \chi^2(\text{BestFit}) + T^2$ with $T \approx 10$

Map of kinematic region covered by data

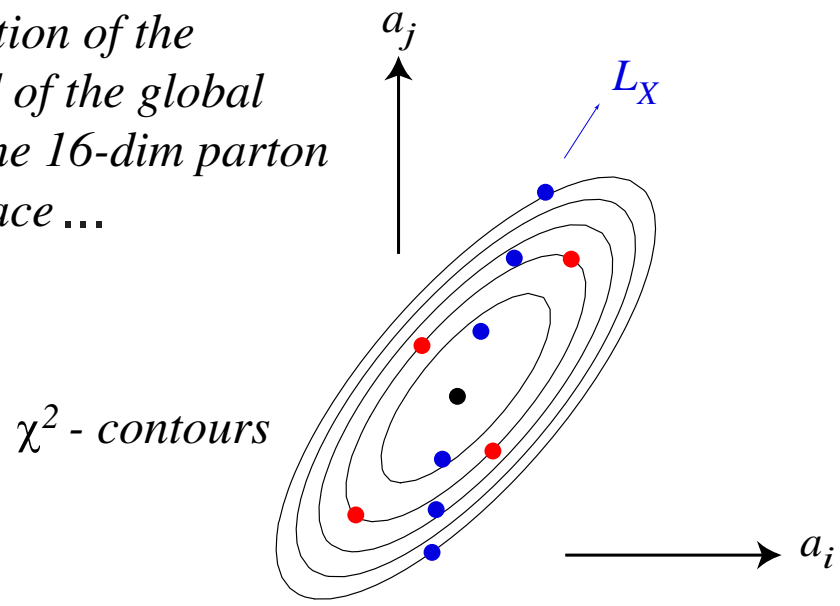


A wide variety of data is tied together by the Theory of Evolution, namely DGLAP.

Consistency, or lack thereof, between experiments can be observed only in the context of a global fit.

Overview of MSU uncertainty studies

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



- Lagrange Multiplier Method: Trace χ^2 as function of F (e.g. σ_W) by minimizing $\chi^2 + \lambda F$. Yields special-purpose PDFs that give extremes of F; e.g. extremes of σ_W, or $\langle y \rangle$ for rapidity distribution of W, or σ for $t\bar{t}$ production; or $\sigma_{t\bar{t}}(\sqrt{s} = 14\,\text{TeV})/\sigma_{t\bar{t}}(\sqrt{s} = 2\,\text{TeV})$, or M_W mass measurement error,. . .

- Hessian Matrix Method: use eigenvectors of error matrix. Yields ≈ 32 sets $\{S_i^{\pm}\}$ that are displaced "up" or "down" by $\Delta\chi^2 = 100$ from the best fit. Get error by sum of squares and construct extreme PDFs for any problem of interest. More simply, can just look at extremes from the 32 sets – Big improvement over just looking at extremes from obsolete PDFs!

Hessian (Error Matrix) method

Classical error formulae

$$\Delta\chi^2 = \sum_{ij}(a_i - a_i^{(0)})H_{ij}(a_j - a_j^{(0)})$$

$$(\Delta F)^2 = \Delta\chi^2 \sum_{ij} \frac{\partial F}{\partial a_i}(H^{-1})_{ij}\frac{\partial F}{\partial a_j}$$

where the Hessian matrix H is inverse of error matrix. Direct application of this formula fails because of extreme differences in variation of χ^2 for different directions in the space of fitting parameters ("steep" and "flat" directions), as revealed by the eigenvalues of H. (It is well known that the error matrix computed by Minuit is not useful in complex multiparameter applications.)

This problem is solved by an iterative procedure that finds and rescales the eigenvectors of H, leading to a diagonal form

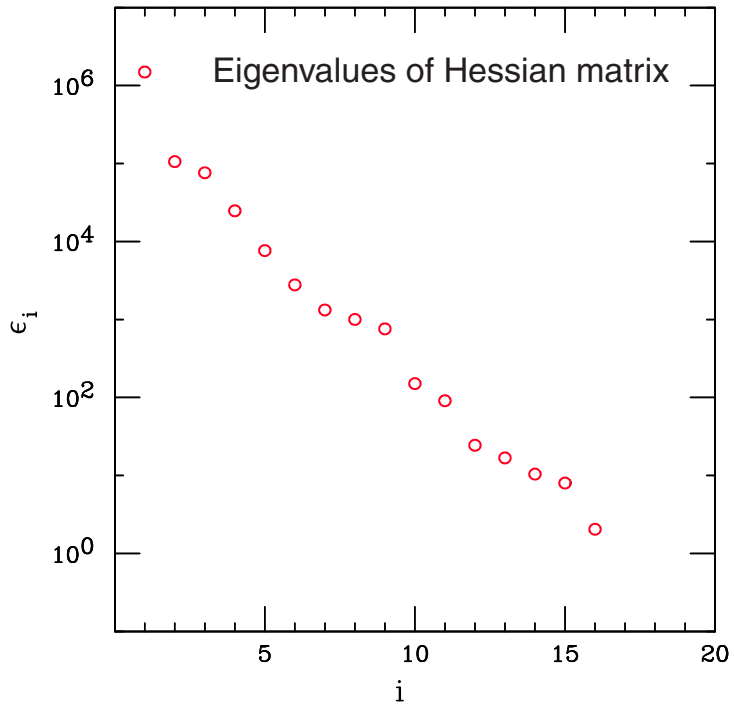
$$\Delta\chi^2 = \sum_i z_i^2$$

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

where $S_i^{(+)}$ and $S_i^{(+)}$ are PDF sets that are displaced along the eigenvector directions.

The iterative procedure is available in FORTRAN at http://www.pa.msu.edu/~pumplin/iterate/

It is under discussion to become a new option in Minuit.



Region of acceptable global fits

$\chi^2 - \chi^2(\text{BestFit}) < T^2$ with $T \approx 10 - 15$

i.e., $\Delta\chi^2 < 100 - 200$. Would have $T = 3$ for "3σ limit", if Gaussian error treatment were OK, which it is NOT because of unknown correlated errors in theory and experiments.

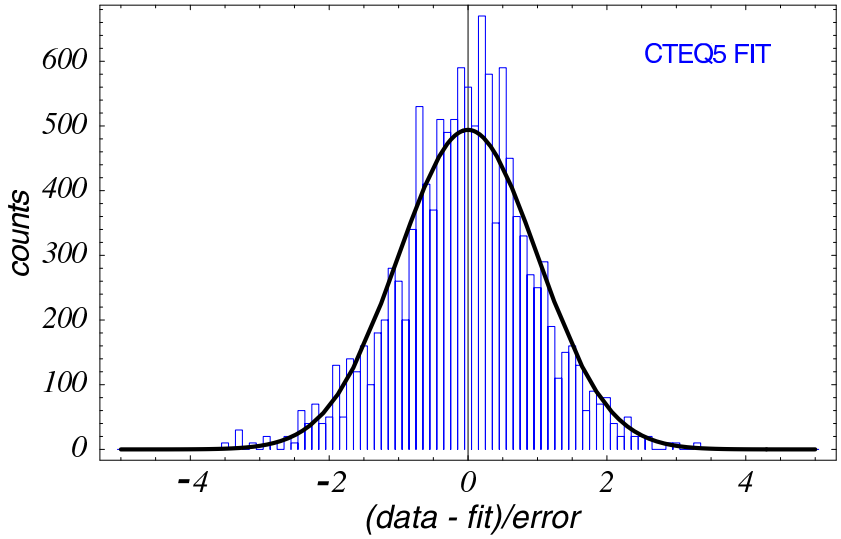
T is determined by consistency requirements: the allowed variations from the Best Fit must include variations as large as those created when each data set is added to the analysis. (Observe that when several of the data sets are added to the fit, the χ^2 for those already included increases by ≈ 20.)

Systematic method (Collins & Pumplin): explore variation of χ^2 for Expt i vs. χ^2 for all others as function of weight assigned to Expt i.

Another way to estimate T: Look at quality of fit to each experiment as a function of physical quantities of interest such as σ_W. $\Delta\chi^2 \sim 100$ may be almost invisible in eyeball comparisons to data — e.g., increasing all discrepancies by 5% above their values in the best fit makes $\Delta\chi^2 = 125$. Or it may be concentrated in one or two experiments and be quite noticeable.

Measures of consistency in CTEQ5

Overall fit looks "Normal": curve is Gaussian $dP/dx \propto \exp(-x^2/2)$ with no adjustable parameters.



But individual data sets are not so Gaussian...

Experiment	N	χ^2_n/N	$(\chi^2_n - N)/\sqrt{2N}$
BCDMS H2	168	0.87	−1.2
BCDMS D2	156	1.42	3.7
H1 F2 96	172	0.63	−3.4
Zeus F2 94	186	1.34	3.3
NMC H2	104	1.04	0.3
NMC D2/H2	123	0.90	−0.8
NMC D2/H2	13	0.99	0.0
CCFR F2	87	0.85	−1.0
CCFR F3	87	0.38	−4.1
E605 $\mu^+\mu^-$	119	0.77	1.8
NA51	1	0.44	−0.4
CDF W asym	11	0.78	−0.5
E866	11	0.45	−1.3
D0 jets	24	0.95	0.2
CDF jets	33	1.65	2.6
Total	1295	0.96	−0.9

New ways to measure consistency of fit

(Work in progress with John Collins)

Key idea: In addition to the

Hypothesis-testing criterion $\Delta\chi^2 \sim \sqrt{2N}$

we use the stronger

Parameter-fitting criterion $\Delta\chi^2 \sim 1$

The parameters here are relative weights assigned to various experiments, or to results obtained using various experimental methods. Examples:

- Plot minimum χ^2_i vs. $\chi^2_{\rm tot} - \chi^2_i$, where χ^2_i is one of the experiments, or all data on nuclei, or all data at low Q^2,...

or

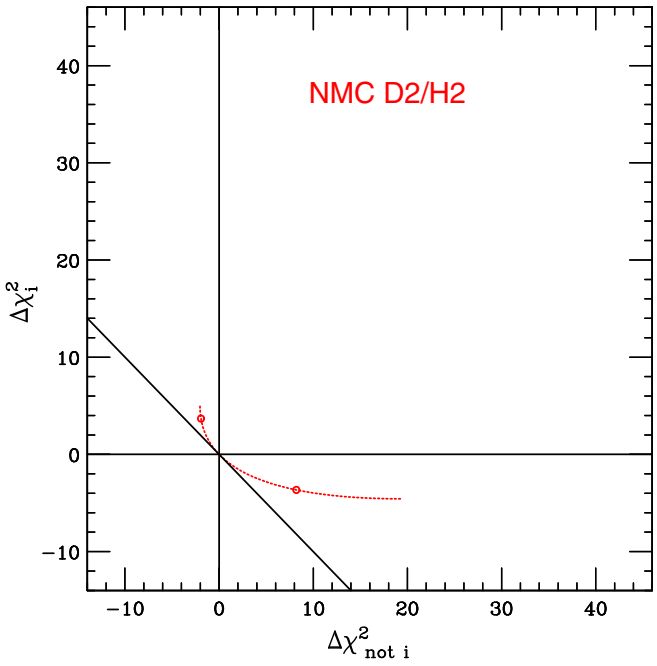
- Plot both as function of Lagrange multiplier u where $(1-u)\chi^2_i + (1+u)(\chi^2_{\rm tot} - \chi^2_i)$ is the quantity minimized.

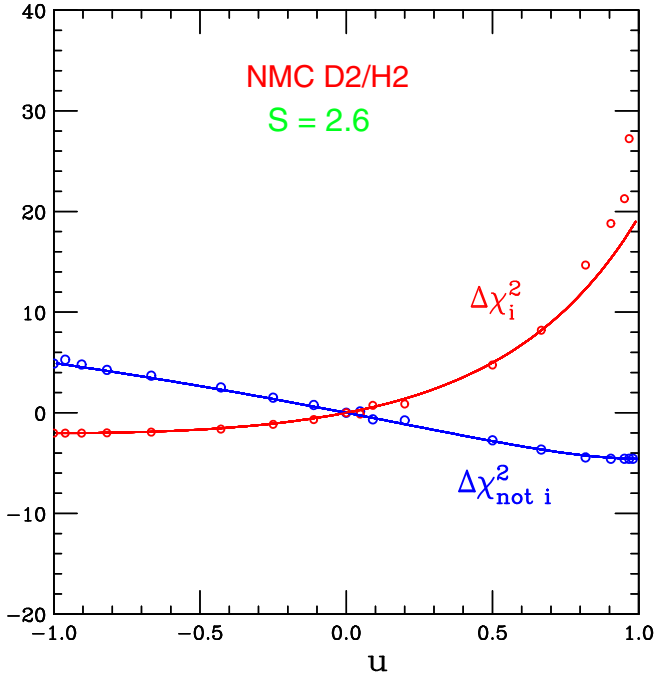
Can obtain quantitative results by fitting to a model with a single common parameter p:

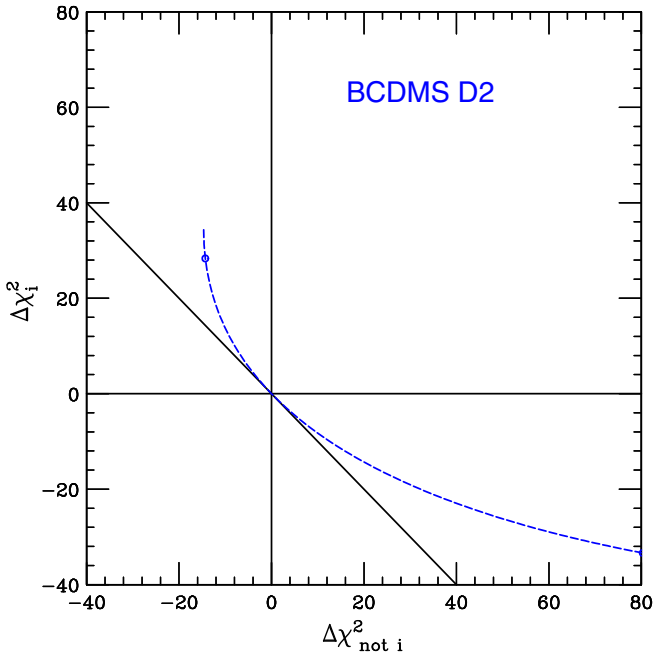
$$\chi^2_i = A + \left(\frac{p}{\sin\theta}\right)^2 \quad \Rightarrow \quad p = 0 \pm \sin\theta$$

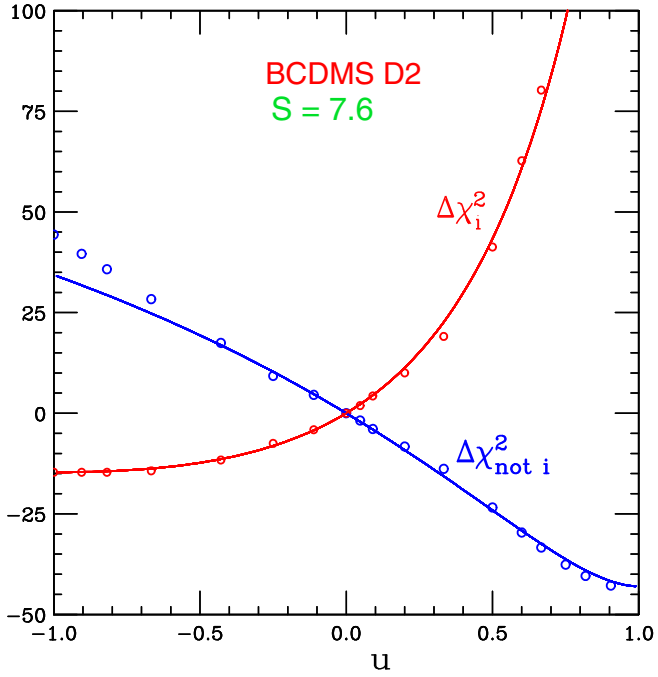
$$\chi^2_{{\rm not}\ i} = B + \left(\frac{p-S}{\cos\theta}\right)^2 \quad \Rightarrow \quad p = S \pm \cos\theta$$

These differ by $S \pm 1$, i.e., by S "standard deviations"





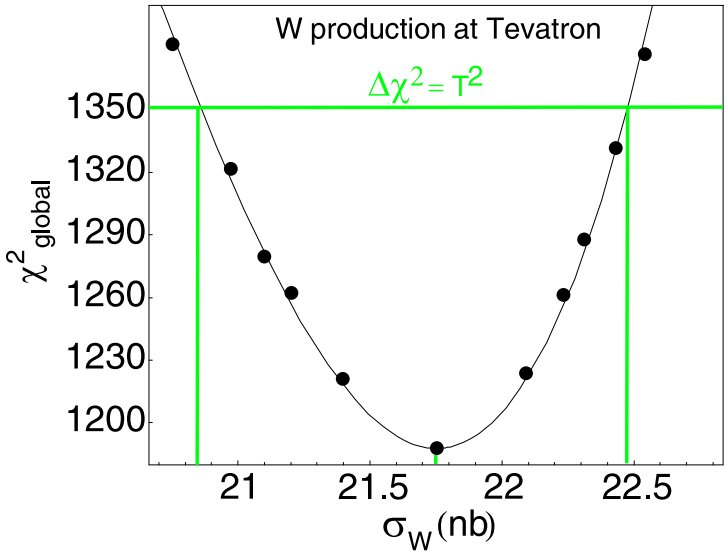


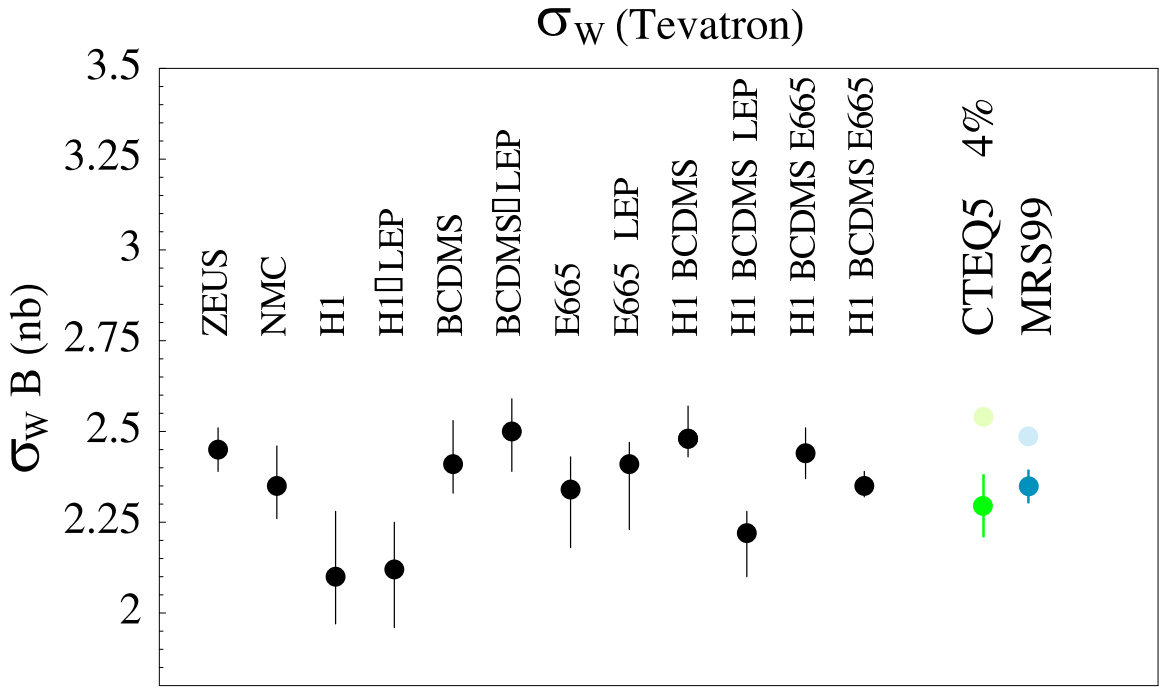


Fits to 8 of the experiments in the CTEQ5 analysis

Expt	1	2	3	4	5	6	7	8
S	2.7	3.3	3.3	4.2	5.3	7.6	7.4	8.3
$\tan\phi$	0.56	0.54	0.99	0.86	0.71	1.14	0.65	0.39

Lagrange Multiplier results





(Assumes leptonic branching fraction 0.1056)

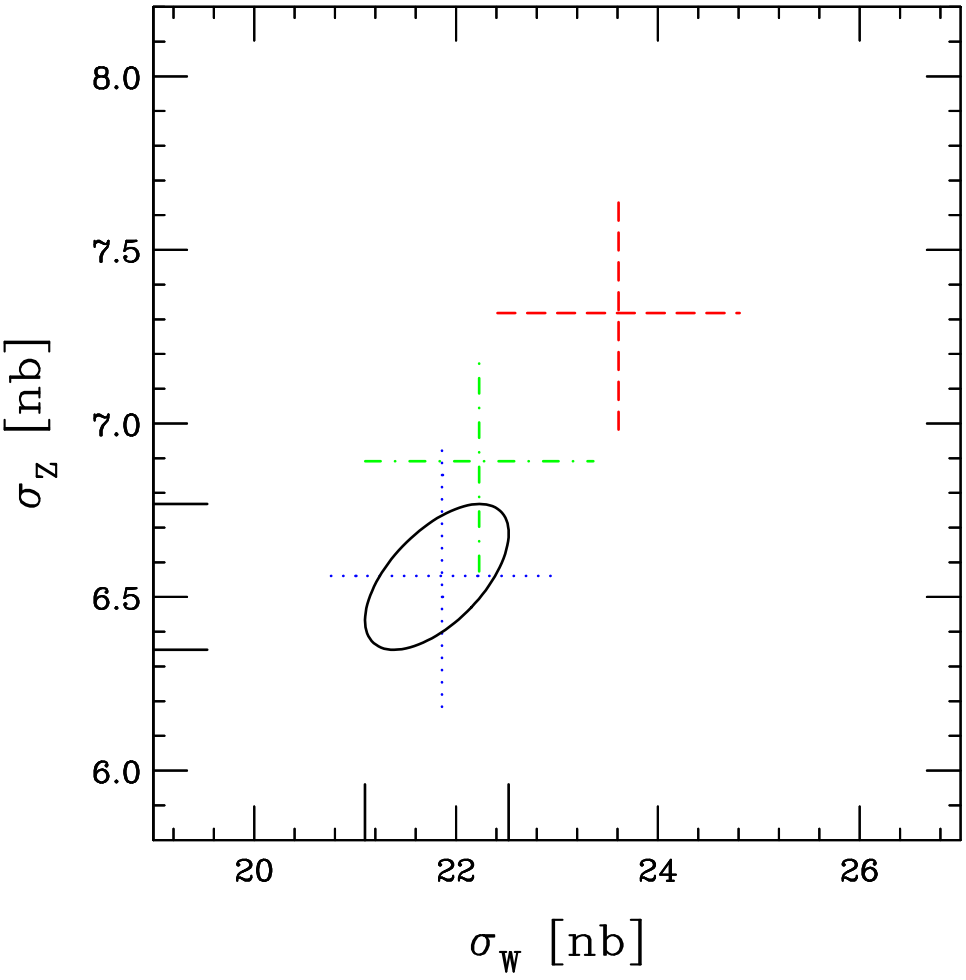
Black points from Giele et al. hep-ph/0104053

CTEQ5 point from hep-ph/0101032

MRST point from Thorne's talk at FNAL

(resolve disagreement with Giele?)

σ_W vs. σ_Z correlation at 1.8 TeV



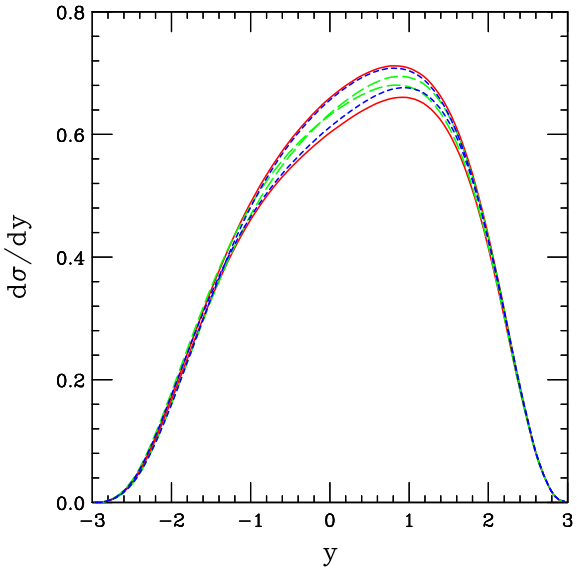
CTEQ5 prediction is the ellipse, obtained using two Lagrange multipliers or by the Hessian method.

Data points (which would also be represented better by ellipses because of strong experimental correlation!) are DØ, CDF, and CDF using same experimental luminosity estimate as DØ.

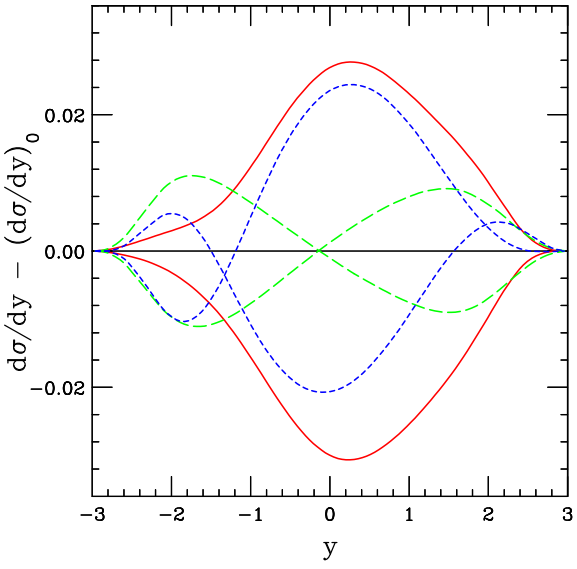
W rapidity distributions

Our methods allow us to calculate the extreme predictions due to PDF uncertainty for whatever quantity is of experimental interest.

For example, extremes of σ_W, $\langle y \rangle$, $\langle y^2 \rangle$ for W production at FNAL:

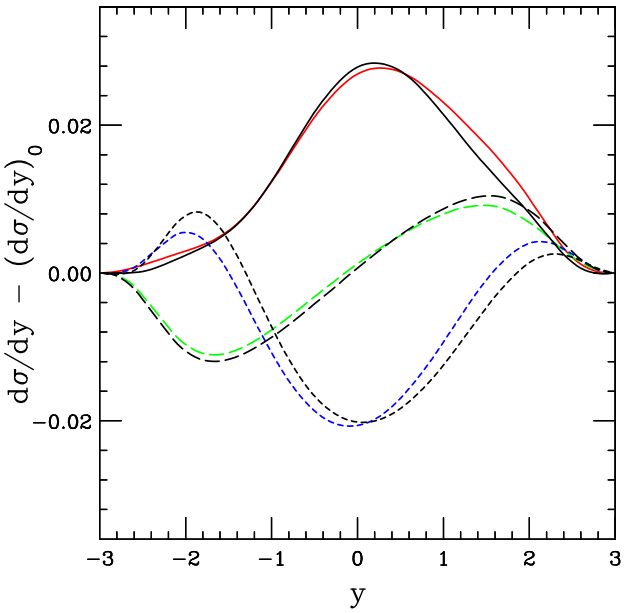


Same curves after subtracting central values. . .

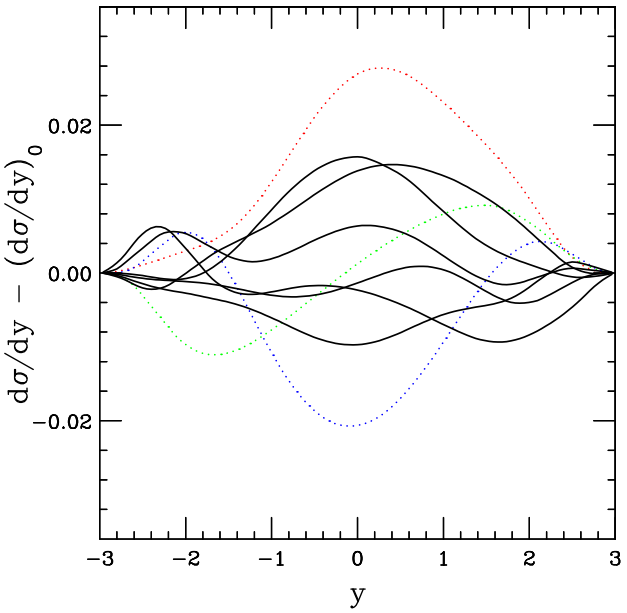


Agreement between Lagrange and Hessian

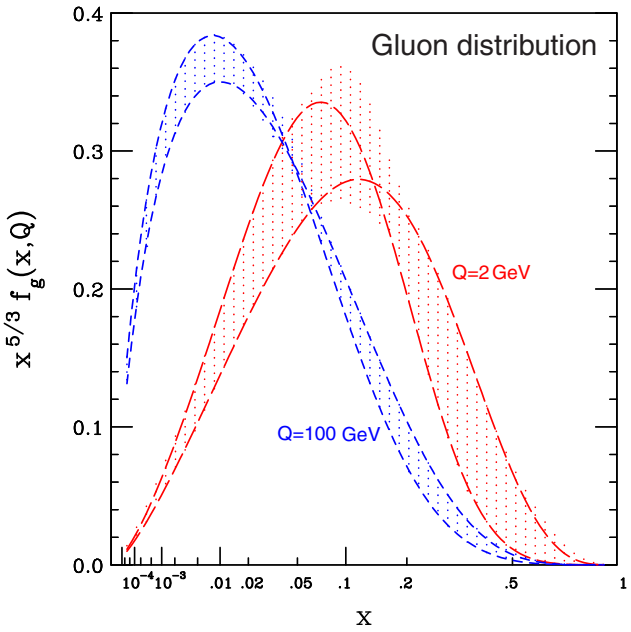
Results for maximum σ_W, $\langle y \rangle$, $\langle y^2 \rangle$ calculated using both methods demonstrate that the approximations made in the Hessian method are OK.



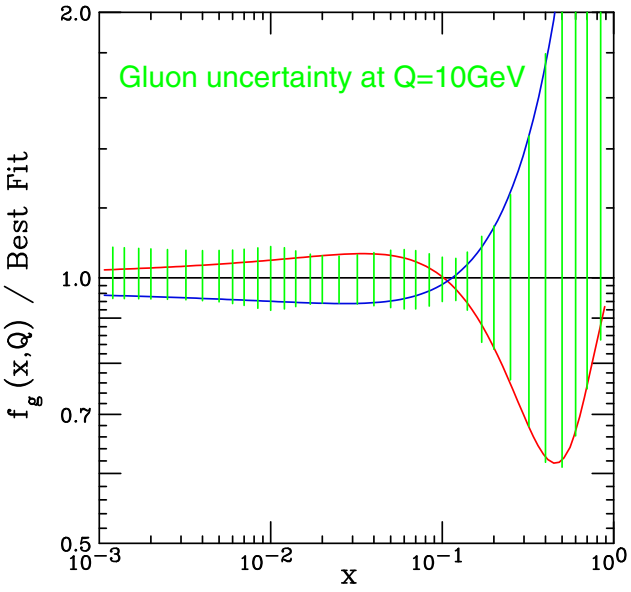
Random PDFs with $\Delta\chi^2 = 100$ (black curves) do not efficiently generate the extreme distributions. . .



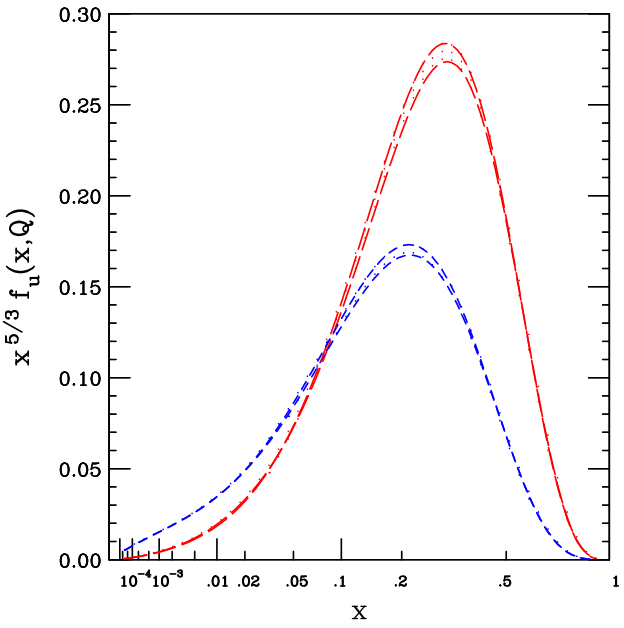
Hessian results: Uncertainty of gluon



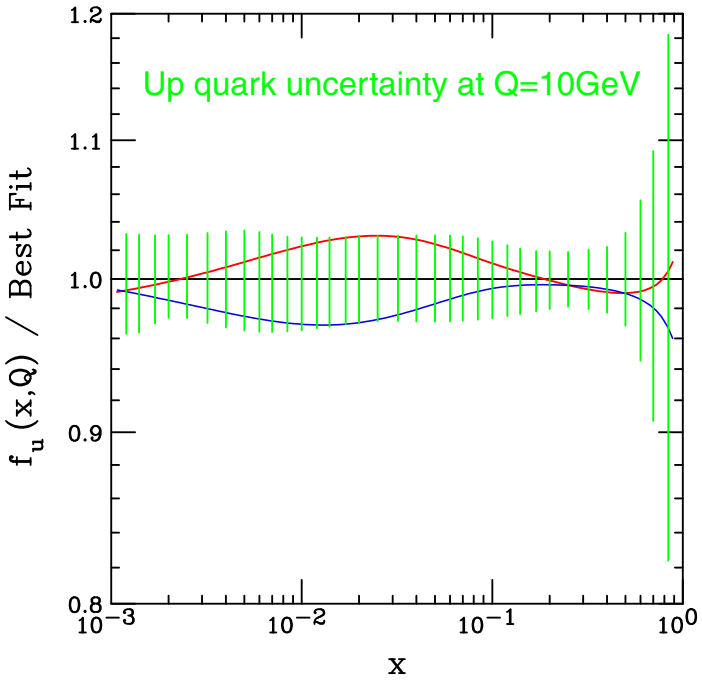
Shaded region shows the range of uncertainties for the gluon distribution in CTEQ5. It is the envelope of distributions like the red and blue curves that minimize or maximize $G(x)$ at a specific value of x. The envelope itself is NOT a possible PDF!



Uncertainty of u-quark

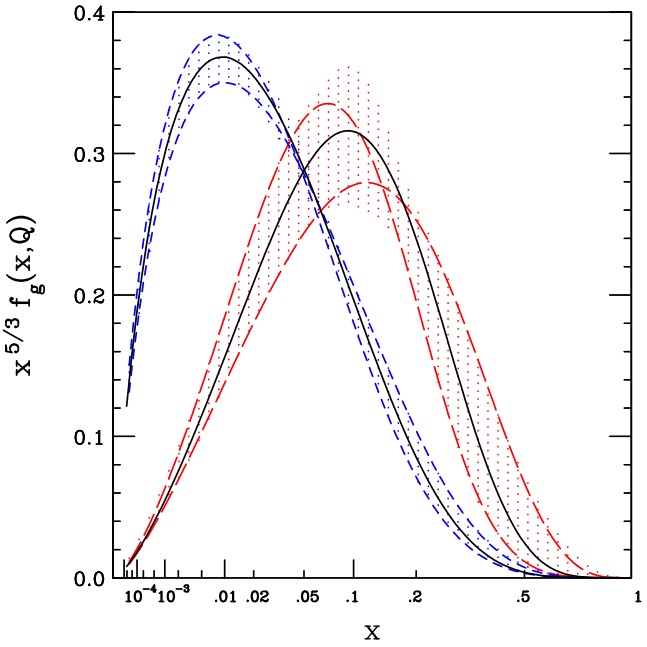


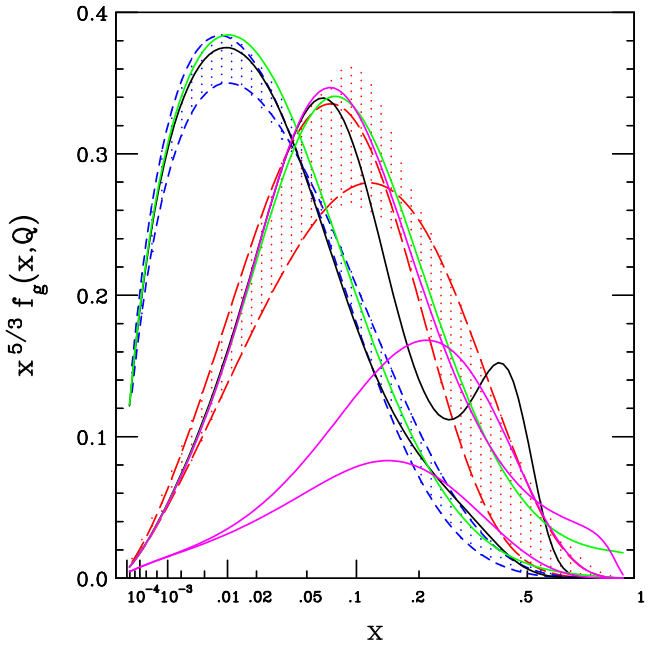
Fractional uncertainty is much smaller than for gluon (note scale is different from gluon plot)



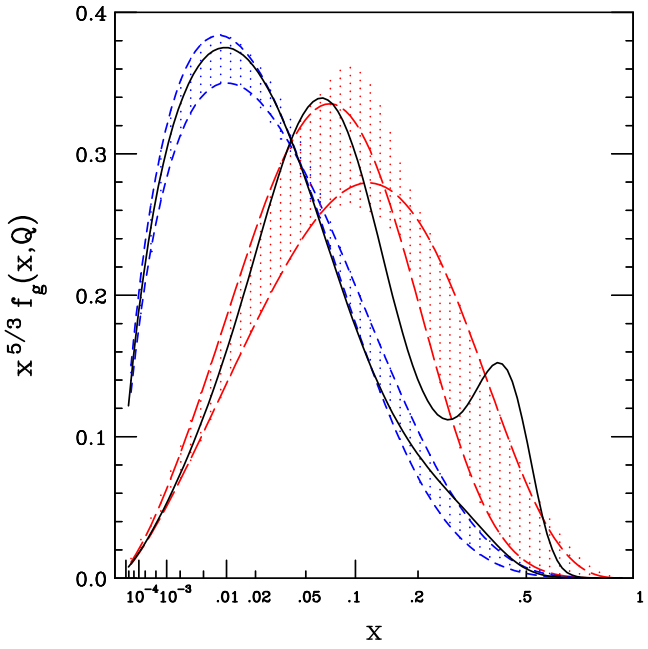
New gluon distributions

Plot $x^{5/3}\,G(x,Q)$ vs. $x^{1/3}$ so area shows contribution to momentum sum rule. $Q = 2\,\mathrm{GeV}$, $Q = 100\,\mathrm{GeV}$





With new H1 and D0 data

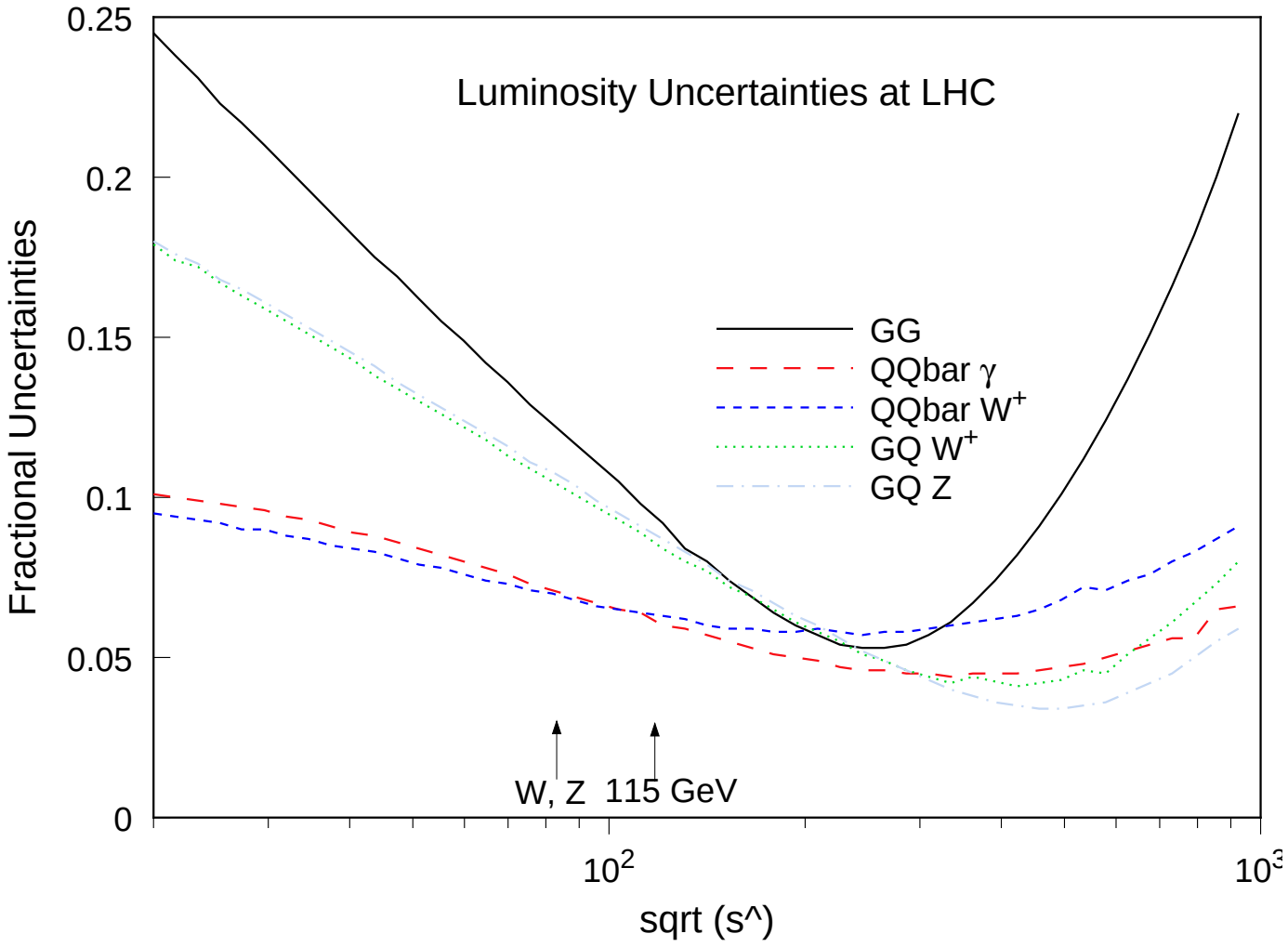


"HJ solution"

Notes: $G(x)$ has become somewhat larger, but within old errors. "HJ" has χ^2 lower by 75 – parametrization dependence or interesting physics?

New Zeus data and correlation matrix errors not yet included.

Uncertainty in parton Luminosities



Interface to Monte Carlos

Methods described here demand increased flexibility from the Monte Carlo simulations, which must be made to easily accept changes in the PDF set.

Have now – or will generate – PDF sets that are tailored to give the extremes of specific quantities, e.g., minimum and maximum cross sections or extreme rapidity distributions for W, or Z, or Higgs, or high-pt jets.

Also have ≈ 32 PDF sets from the Hessian Method, which have $\Delta\chi^2 = 100$ in the directions of the eigenvectors of the error matrix.

Action items for WORKshop

- The most natural interface to the Monte Carlos would be for them to call a user-supplied function that returns the PDFs as a function of x and flavor at the non-perturbative scale $Q_0 = 1\,\mathrm{GeV}$, which is the value currently used by PDF analysis groups.
- If this is done, the DGLAP evolution code must be carefully standardized, because choices of power corrections and grid points can potentially affect the evolution over the very low Q region, and hence affect the PDFs at all Q.
- The Hessian method requires calculations with currently 32 different PDF sets (up and down along 16 eigenvector directions). Giele's method requires still more PDFs. It should be possible to run a Monte Carlo simulation just once, keeping track of the PDF values that lead to each simulated event. Then results for different-but-similar PDFs could be found by reweighting the generated events.