The Art of Global Fitting for Parton Distributions*

Jon Pumplin

Department of Physics and Astronomy,

Michigan State University, East Lansing, MI 48824 USA

I. INTRODUCTION

High energy hadrons interact through their quark and gluon constituents. The interactions become weak at short distances thanks to the asymptotic freedom property of Quantum Chromodynamics, which allows perturbation theory to be applied to a rich variety of experiments. The nonperturbative nature of the proton for single hard interactions is thus characterized by Parton Distribution Functions (PDFs) $f_a(Q,x)$ of momentum scale Q and light-cone momentum fraction x. The evolution in Q is determined perturbatively by QCD renormalization group equations, so the non-perturbative physics can be characterized by functions of x alone at a fixed small Q_0. The ongoing project to extract those functions from experiment is the subject of this talk.

The goal of the CTEQ global fitting group (like that of our friendly rivals MRST) is to extract the universal non-perturbative parton densities of the proton—or potentially of the neutron or nucleus—from a large variety of hard-scattering experiments, based on three pillars:

- **Factorization** $\Rightarrow$ Short distance and long distance are separable (see Fig. **??**);
- **Asymptotic Freedom** $\Rightarrow$ Hard scattering processes are perturbatively calculable;
- **DGLAP Evolution** $\Rightarrow$ PDFs are characterized by functions of x at a fixed small Q_0, with the PDFs at all higher Q being determined from these by renormalization group evolution.

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

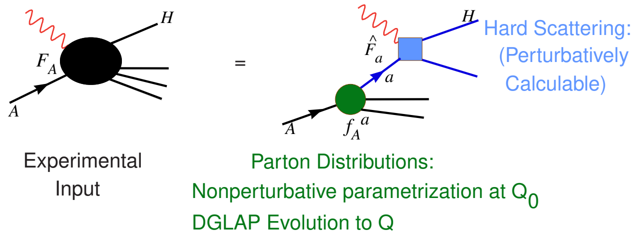

FIG. 1: The factorization theorem

The goals of this work are to test the consistency of QCD theory with the collective body of experiments; to test the consistency of individual experiments with QCD and the others;

* Presented at the Ringberg Workshop "New Trends in HERA Physics," Sept 29,2003

to explore the range of uncertainties of the estimated PDFs; and to make state-of-the-art extracted PDFs widely available, since they are needed to interpret all experiments that use hadron beams or targets—HERA, RHIC, Tevatron, LHC, and non-accelerator.

The kinematic range of available data, after cuts to suppress effects at low Q and low W, are shown in Fig. **??**. The data cover a wide range of scales. They are tied together by the DGLAP evolution equation and by the fact that the PDFs are universal. Consistency or inconsistency between different processes, and between different data points for the same process can be observed only by applying QCD to tie them together in a global fit. Future data from HERA, Tevatron run II (W, Z production), HERA II, and LHC will dramatically extend the range and accuracy of this global fit.

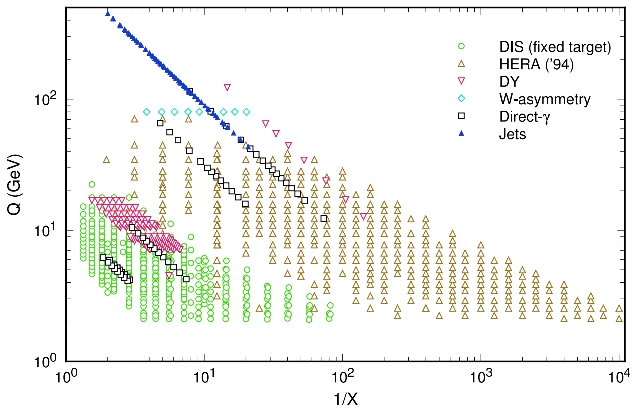

FIG. 2: Kinematic region covered by data

II. SOME DETAILS OF THE CTEQ6 GLOBAL ANALYSIS

Input from Experiment: ~ 2000 data points with $Q > 2\,\mathrm{GeV}$, $W > 3.5\,\mathrm{GeV}$ 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.

Input from Theory: NLO QCD evolution and hard scattering.

Parametrization at Q_0: Use the form $A_0\,x^{A_1}\,(1-x)^{A_2}\,e^{A_3 x}(1+A_4 x)^{A_5}$ for $u_v = u - \bar{u}$, $d_v = d - \bar{d}$, $\bar{u} + \bar{d}$, and g.

Assumptions based on lack of information: $s = \bar{s} = 0.4\,(\bar{u} + \bar{d})/2$ at Q_0; no intrinsic b or c.

Procedure: Construct effective $\chi^2_{\mathrm{global}} = \sum_{\mathrm{expts}} \chi^2_n$, including published systematic error correlations. Minimize χ^2_{global} to obtain "Best Fit" PDFs.

Uncertainty estimates: Use the variation of χ^2_{global} in neighborhood of the minimum to estimate uncertainty limits as the region of parameter space where $\chi^2 < \chi^2(\mathrm{BestFit}) + T^2$ with $T \approx 10$. This "Tolerance Factor" $T \sim 10$ is quite different from the traditional value 1 from Gaussian statistics, because of unknown systematic errors in theory and experiments. It can be estimated from the apparent inconsistencies between experiments when they are combined in the global fit.

A note on the details of parametrization at Q_0: For $d_{\rm val}$, $u_{\rm val}$, $\bar{u}+\bar{d}$ or g, we use the general form

$$xf(x,Q_0) = A_0\, x^{A_1}\,(1-x)^{A_2}\, e^{A_3 x}(1+e^{A_4}x)^{A_5}$$

This corresponds to

$$\frac{d}{dx}\ln(xf) = \frac{A_1}{x} - \frac{A_2}{1-x} + \frac{c_3+c_4 x}{1+c_5 x}$$

i.e., we add a 1:1 Padé approximation term to the singular terms of the time-honored $A_0\, x^{A_1}\,(1-x)^{A_2}$ parametrization, which is based on Regge theory and quark counting rules. A sufficiently flexible parametrization is important; but for convergence, there must not be too many "flat directions." For that reason, some of the parameters are frozen at arbitrary values for some flavors.

To measure a set of continuous PDF functions at Q_0 on the basis of a finite set of data points would appear to be an ill-posed mathematical problem. However, this difficulty is not so severe as might be expected since the actual predictions of interest that are based on the PDFs are discrete quantities. In particular, fine-scale structure in x in the PDFs at Q_0 tend to be smoothed out by evolution in Q. They correspond to flat directions in χ^2 space, so they are not accurately measured; but they have little effect on the applications of interest.

Some representative Best Fit parton distributions from the analysis are shown in Figs **??**, **??**. Ones sees that valence quarks dominate for $x \to 1$, and the gluon dominates for $x \to 0$, especially at large Q.

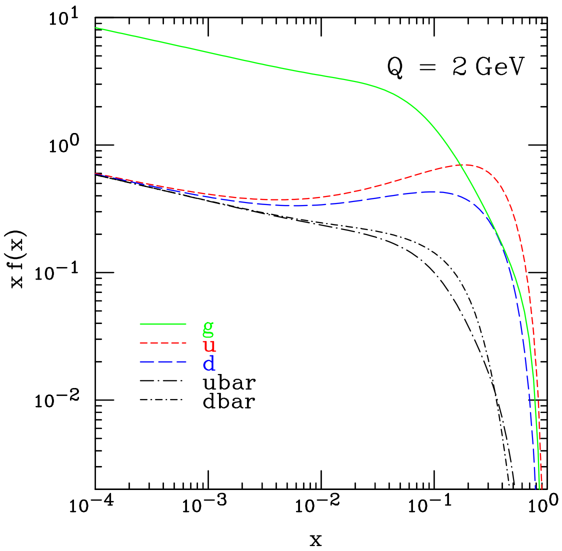

FIG. 3: Parton distributions at $Q = 2$ GeV

III. INCLUDING EXPERIMENTAL SYSTEMATIC ERRORS

The simplest definition of chi-squared,

$$\chi_0^2 = \sum_{i=1}^{N} \frac{(D_i - T_i)^2}{\sigma_i^2} \qquad \begin{cases} D_i = \text{ data} \\ T_i = \text{ theory} \\ \sigma_i = \text{ "expt. error"} \end{cases}$$

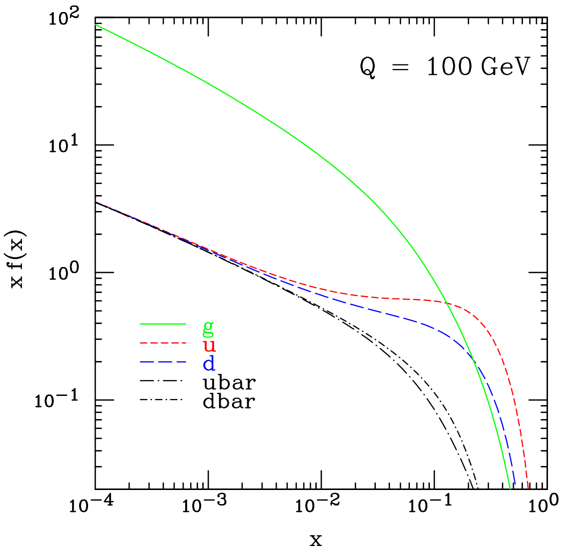

FIG. 4: Parton distributions at $Q = 100$ GeV

works for random Gaussian errors,

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

When correlated experimental systematic errors have been estimated, the errors become

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

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

In that case the parameters $\{r_k\}$ that control the corrections for systematic errors join the original PDF shape parameters $\{a_\lambda\}$ as parameters for which the global χ^2 is to be minimized. In effect, the other experiments in the global fit are used to determine the unknown systematic error coefficients. (This is an extension of a familiar situation in which the overall normalization of an experiment may be determined by fitting to others if it is not well measured directly.)

To take into account the systematic errors, we define

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

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

$$A_{kk'} \; = \; \delta_{kk'} + \sum_{i=1}^{N} \frac{\beta_{ki}\beta_{k'i}}{\alpha_i^2}$$

$$B_k = \sum_{i=1}^{N} \frac{\beta_{ki}\,(D_i - T_i)}{\alpha_i^2}.$$

Since the dependence of χ^2 on the $\widehat{r}_k$'s is quadratic, we are able to minimize with respect to those parameters explicitly for any set of PDF shape parameters $\{a_\lambda\}$. We then minimize the remaining $\chi^2(a)$ with respect to the PDF shape parameters $\{a_\lambda\}$. Then the $\{a_\lambda\}$ determine $f_i(x, Q_0^2)$. Meanwhile, $\{\widehat{r}_k\}$ 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.

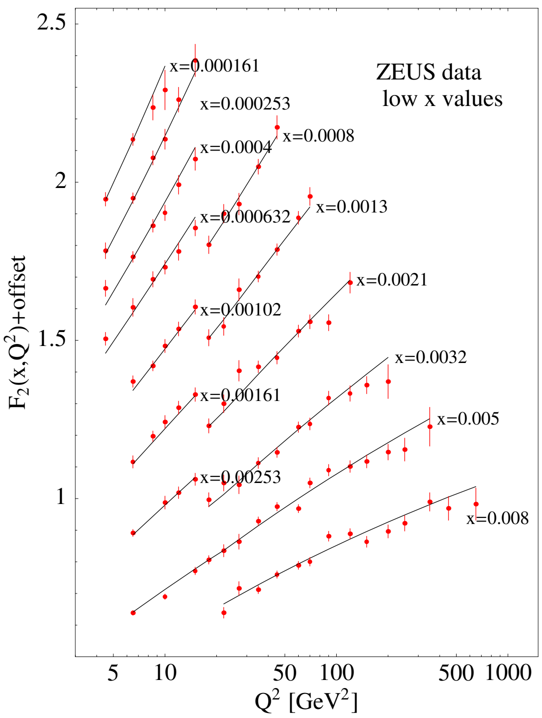

FIG. 5: CTEQ6M fit to ZEUS data at low x

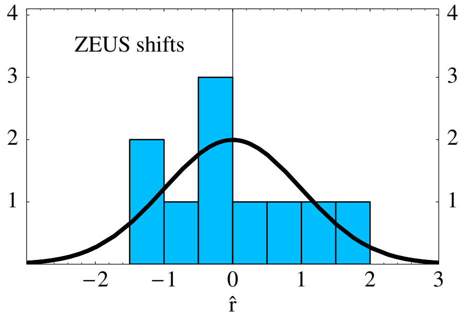

FIG. 6: The 10 systematic error parameters for the ZEUS data. Curve is a Gaussian of width 1.

An example of how this works is shown in Fig. **??**: The data points include the estimated corrections for systematic errors, i.e., the central values plotted have been shifted by an amount that is consistent with the estimated systematic errors, where the systematic error parameters are determined using other experiments via the global fit. The error bars are statistical errors only. Fig. **??** shows that the systematic error parameters come out of order 1 as they should. Fig. **??** shows that the residuals defined as (data-theory)/error also come out with the expected distribution. In contrast, Fig. **??** shows that if the systematic shifts are not included, the fit becomes bad.

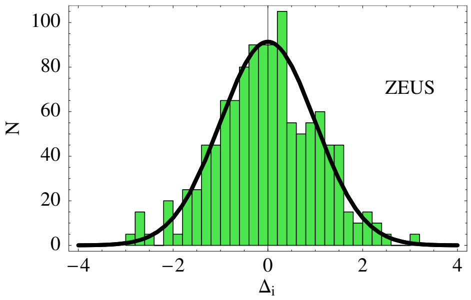

FIG. 7: Residuals for the ZEUS data. Curve is a Gaussian of width 1.

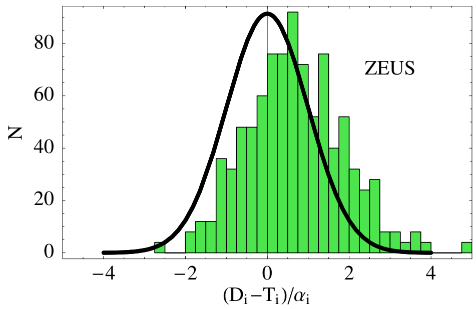

FIG. 8: Residuals for the ZEUS data without including systematic shifts.

IV. UNCERTAINTIES IN PDFS

There are several **Sources of uncertainty:**

- Experimental errors included in χ^2
- Unknown experimental errors
- Parametrization dependence
- Higher-order corrections & Large Logarithms
- Power Law corrections ("higher twist")

There are som **Fundamental difficulties:**

- Good experiments run until systematic errors dominate, so the magnitude of remaining systematic errors involves guesswork.

- Systematic errors of the theory and their correlations are even harder to guess.
- Quasi–ill-posed problem: we must determine continuous functions from a discrete data set. (Because of the smoothing effect of DGLAP evolution, this is not as impossible as it sounds.)
- Some combinations of variables are unconstrained, e.g., $s - \bar{s}$ before NuTeV data.

There are several **Approachs to estimating the uncertainty.** In all of the uncertainty methods, we continue to use χ^2 as a measure of the quality of the fit; but vary weights assigned to the experiments to estimate the range of acceptable fits, rather that relying on the classical $\Delta\chi^2 = 1$.

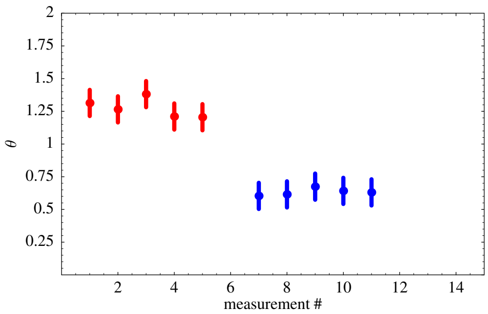

FIG. 9: Hypothetical measurements of hypothetical parameter θ.

The essence of the Uncertainty Problem can be seen in the Fig. **??**, which shows some hypothetical measurements of a single parameter θ. Suppose the quantity θ has been measured by two different experiments, or extracted using two different approximations to the True Theory. What would you quote as the Best Fit and the Uncertainty? The disagreements are of course not so obvious in the many-parameter global fit. However, the disagreements can be probed in one dimension by, for example, studying the variations of the Best Fit that result from assigning different weights to different experiments, or to different kinematic regions etc. Much of this is discussed in our papers; and more is work in progress.

V. OUTLOOK

Parton Distribution Functions are a necessary infrastructure for precision Standard Model studies and New Physics searches at hadron colliders and at experiments using hadron targets. Some issues that were discussed, but not necessarily included in this writeup due to space limitations:

- PDFs of the proton are increasingly well measured.
- Useful tools are in place to estimate the uncertainty of PDFs and to propagate those uncertainties to physical predictions. There is adequate agreement between various methods for estimating the uncertainty:
  - "Hessian Method" based on the eigenvectors of the error matrix

  - "Lagrange Multiplier Method" based on finding the uncertainty on a predicted quantity by studying the variation of χ^2 as a function of that quantity
  - systematic reweighting of experiments (work in progress with John Collins)
  - random reweighting of experiments: a variant of the "well known" statistical bootstrap method

- The "Les Houches Accord" interface makes it easy to handle the large number of PDF solutions that are needed to characterize uncertainties. [hep-ph/0204316]
- Improvements in the treatment of heavy quark effects are in progress, and together with neutrino experiments they will allow improved flavor differentiation.
- Since PDFs summarize some fundamental nonperturbative physics of the proton, they should be considered a challenge to be computed! (Low moments of meson PDFs have indeed been calculated in lattice gauge theory.)
- Other non-perturbative methods, e.g. for $s(x) - \bar{s}(x)$, may be helpful.
- HERA and Fermilab run II data will provide the next major experimental steps forward, followed by LHC.
- Theoretical improvements such as resummation to use direct photon and W transverse momentum data will be useful.
- In view of possible isospin breaking, and the importance of nuclear shadowing & anti-shadowing effects, HERA measurements on deuterons would be highly welcome.

I thank the organizers of Ringberg '03 for an excellent workshop in an excellent setting. I thank my collaborators on this work, D. Stump, W.K. Tung, J. Huston, H. Lai, P. Nadolsky, F. Olness, S. Kuhlmann, J. Owens; S. Kretzer, and J. Collins for many valuable discussions.