Parton distribution functions

- Global analysis
- Practical applications

Pavel Nadolsky

Department of Physics
Southern Methodist University (Dallas, TX)

Lecture 2

Stages of the PDF analysis

1. Select experimental data
2. Assemble all relevant theoretical cross sections and verify their mutual consistency
3. Choose the functional form for PDF parametrizations
4. Perform a fit
5. Make the new PDFs and their uncertainties available to end users

1. Selection of experimental data

- **Neutral-current ep DIS data from HERA are most extensive and precise** among all data sets
  - In addition, their systematic errors were reduced recently by cross calibration of H1 and ZEUS detectors
- However, by their nature they constrain only a limited number of PDF parameters
- Thus, two complementary approaches to the selection of the data are possible

Two strategies for selection of experimental data

DIS-based analyses ⇒ focus on the most precise (HERA DIS) data

- NC DIS, CC DIS, NC DIS jet, c and b production *(H1, ZEUS, HERAPDF)*

Global analyses *(CT09, MSTW'2008, NNPDF1.1)*
⇒ focus on completeness, reliable flavor decomposition

- all HERA data + fixed-target DIS data
  - notably, CCFR and NuTeV νN DIS constraining $s(x, Q)$
- low-Q Drell-Yan (E605, E866), Run-1 W lepton asymmetry, Run-2 Z rapidity *(CT09, MSTW'08, upcoming NNPDF2.0)*
- Tevatron Run-2 jet production, W asymmetry *(CT09, MSTW'08)*

Toward CT09 PDF analysis

- An update of CTEQ6.6 study *(PRD 78, 013004 (2008))*
- New experimental data in the fit
  - CDF Run-2 and D0 Run-2 inclusive jet production
    - preliminarily explored in *J. Pumplin et al., arXiv:0904.0424; P.N., in preparation*
  - CDF Run-2 lepton asymmetry
  - CDF Z rapidity distribution
  - low-Q Drell-Yan p_T (E288, E605, R209) and Tevatron Run-1, Run-2 Z p_T distributions
- updated procedure for PDF error estimates

2. Theoretical cross sections

Process	Number of QCD loops	Mass scheme*	
Neutral current DIS	2	ZM	*Moch, Vermaseren, Vogt*
	2	GM	*Harris, Smith; Buza, Matiounine, Smith, van Neerven*
Charged current DIS	2	ZM	*Moch, Vermaseren, Vogt*
	1	GM	
$pN \xrightarrow{\gamma^*,W,Z} \ell\bar{\ell}^{(\prime)}X$	2	ZM	*Anastasiou, Dixon, Melnikov, Petriello*
$p\bar{p} \to jX$	1	ZM	
$ep \to jjX$	2	ZM	

*ZM/GM: zero-mass/general-mass approximation for c, b contributions

Although "NNLO" PDF fits include most of the NNLO hard cross sections, more work is needed to bring them to true NNLO accuracy

Meanwhile, the NLO PDFs can still be used in most cases

3. Requirements for PDF parametrizations

A. A valid set of $f_{a/p}(x,Q)$ must satisfy QCD sum rules

Valence sum rule

$$\int_0^1 [u(x,Q) - \bar{u}(x,Q)]\,dx = 2 \qquad \int_0^1 \left[d(x,Q) - \bar{d}(x,Q)\right]dx = 1$$

$$\int_0^1 [s(x,Q) - \bar{s}(x,Q)]\,dx = 0$$

A proton has net quantum numbers of 2 u quarks + 1 d quark

Momentum sum rule

$$[\text{proton}] \equiv \sum_{a=g,q,\bar{q}} \int_0^1 x\, f_{a/p}(x,Q)\,dx = 1$$

momenta of all partons must add up to the proton's momentum

Through this rule, normalization of $g(x,Q)$ is tied to the first moments of quark PDFs

3. Requirements for PDF parametrizations

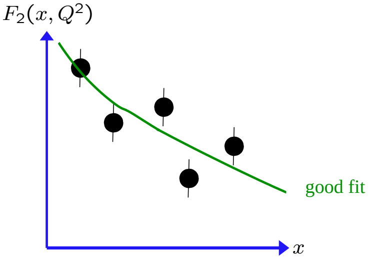
B. A valid PDF set must **not** produce unphysical predictions for observable quantities

Example

- Any concievable hadronic cross section σ must be non-negative: $\sigma \geq 0$
  - this is typically realized by requiring $f_{a/p}(x, Q) > 0$
- Any cross section asymmetry A must lie in the range $-1 \leq A \leq 1$
  - this constrains the range of allowed PDF pararametrizations

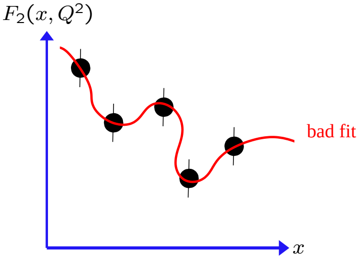
3. Requirements for PDF parametrizations

C. PDF parametrizations for $f_{a/p}(x,Q)$ must be "flexible just enough" to reach agreement with the data, without reproducing random fluctuations



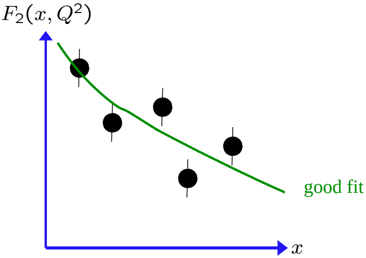
3. Requirements for PDF parametrizations

C. PDF parametrizations for $f_{a/p}(x, Q)$ must be "flexible just enough" to reach agreement with the data, without reproducing random fluctuations



3. Requirements for PDF parametrizations

C. PDF parametrizations for $f_{a/p}(x,Q)$ must be "flexible just enough" to reach agreement with the data, without reproducing random fluctuations



Traditional solution

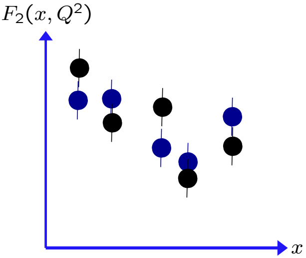
"Theoretically motivated" functions with a few parameters

$$f_{i/p}(x,Q_0) = a_0 x^{a_1}(1-x)^{a_2} \times F(x; a_3, a_4, ...)$$

- $x \to 0$: $f \propto x^{a_1}$ – Regge-like behavior
- $x \to 1$: $f \propto (1-x)^{a_2}$ – quark counting rules
- $F(a_3, a_4, ...)$ affects intermediate x; just a convenient functional form

3. Requirements for PDF parametrizations

C. PDF parametrizations for $f_{a/p}(x, Q)$ must be "flexible just enough" to reach agreement with the data, without reproducing random fluctuations



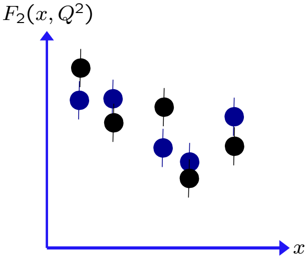
Radical solution

Neural Network PDF collaboration

- Generate N replicas of the experimental data, randomly scattered around the original data in accordance with the probability suggested by the experimental errors
- Divide the replicas into a fitting sample and control sample

3. Requirements for PDF parametrizations

C. PDF parametrizations for $f_{a/p}(x,Q)$ must be "flexible just enough" to reach agreement with the data, without reproducing random fluctuations



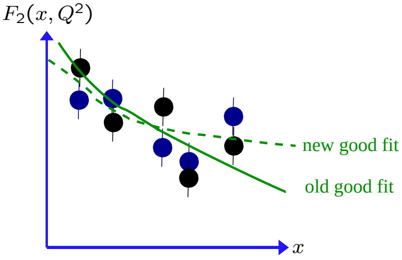
Radical solution

Neural Network PDF collaboration

- Parametrize $f_{a/p}(x,Q)$ by ultra-flexible functions — neural networks
- A statistical theorem states that any function can be approximated by a neural network with a sufficient number of nodes (in practice, of order 10)

3. Requirements for PDF parametrizations

C. PDF parametrizations for $f_{a/p}(x,Q)$ must be "flexible just enough" to reach agreement with the data, without reproducing random fluctuations



Radical solution

Neural Network PDF collaboration

- Fit the neural nets to the fitting sample, while demanding good agreement with the control sample
- Smoothness of $f_{a/p}(x,Q)$ is preserved, despite its nominal flexibility

4. Statistical aspects

J. Pumplin et al., JHEP 0207, 012 (2002), and references therein; J. Collins & J. Pumplin, hep-ph/0105207

Suppose there are N PDF parameters $\{a_i\}$, N_{exp} experiments, M_k data points and N_k correlated systematic errors in each experiment

Each systematic error is associated with a random parameter r_n, distributed in to be distributed according to a Gaussian distribution with unit dispersion

The best external estimate of syst. errors corresponds to $\{r_n = 0\}$; but we must allow for $r_n \neq 0$

The most likely combination of $\{a\}$ and $\{r\}$ is found by minimizing

$$\chi^2 = \sum_{k=1}^{N_{exp}} w_k \chi_k^2$$

$w_k > 0$ are weights applied to emphasize or de-emphasize contributions from individual experiments (default: $w_k = 1$)

4. Statistical aspects

J. Pumplin et al., JHEP 0207, 012 (2002), and references therein; J. Collins & J. Pumplin, hep-ph/0105207

χ^2 for one experiment is

$$\chi_k^2 = \sum_{i=1}^{M_k} \frac{1}{\sigma_i^2} \left(D_i - T_i(\{a\}) - \sum_{n=1}^{R_k} r_n \beta_{ni} \right)^2 + \sum_{n=1}^{R_k} r_n^2$$

D_i and T_i are **data** and **theory** values at each point

$\sigma_i = \sqrt{\sigma_{stat}^2 + \sigma_{syst,uncor}^2}$ is the total statistical + systematical **uncorrelated** error

$\sum_n \beta_{ni} r_k$ are **correlated** systematic shifts

β_{ni} is the **correlation** matrix; is provided with the data or theoretical cross sections before the fit

$\sum_n r_n^2$ is the penalty for deviations of r_n from their expected values, $r_n = 0$

4. Statistical aspects

J. Pumplin et al., JHEP 0207, 012 (2002), and references therein; J. Collins & J. Pumplin, hep-ph/0105207

Each χ_k can be **analytically** minimized with respect to **the Gaussian** r_n , with the result

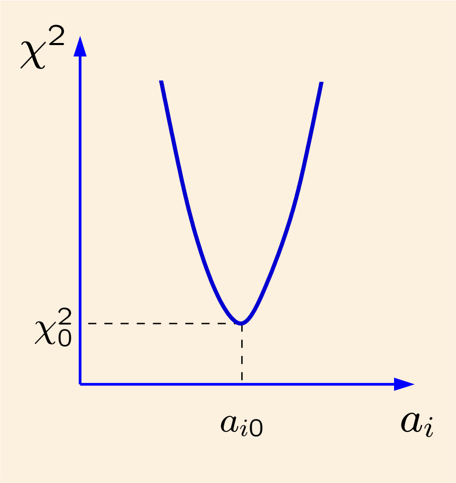
$$r_n\left(\{a\}\right) = \sum_{n'=1}^{R_k} (A^{-1})_{nn'} B_{n'}(\{a\})$$

$$A_{nn'} = \delta_{nn'} + \sum_{i=1}^{M_k} \frac{\beta_{ni}\beta_{n'i}}{\sigma_i^2}; \qquad B_n(\{a\}) = \sum_{i=1}^{M_k} \frac{\beta_{ni}(D_i - T_i)}{\sigma_i^2}$$

$$\chi_k^2 = \sum_{i=1}^{M_k} \frac{1}{\sigma_i^2} \left(D_{ki} - T_{ki}(\{a\})\right)^2 + \sum_{n,n'=1}^{R_k} B_n (A^{-1})_{nn'} B_{n'} \qquad (1)$$

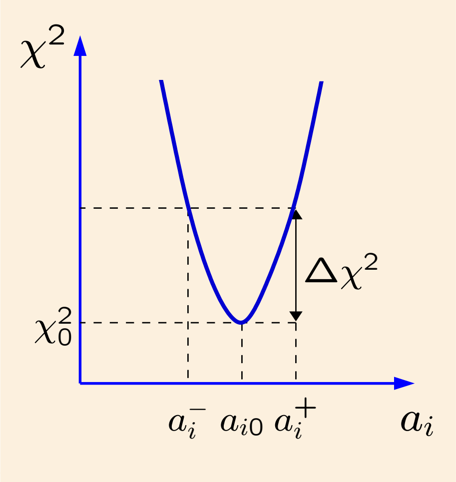
Numerical minimization of $\sum_k w_k \chi_k^2(a, r(a))$ (with χ_k from Eq. (1)) then establishes the region of acceptable $\{a\}$, which includes the largest possible variations of $\{a\}$ that are allowed by the systematic effects

Multi-dimensional error analysis



- Minimization of a likelihood function (χ^2) with respect to ~ 30 theoretical (mostly PDF) parameters $\{a_i\}$ and > 100 experimental systematical parameters

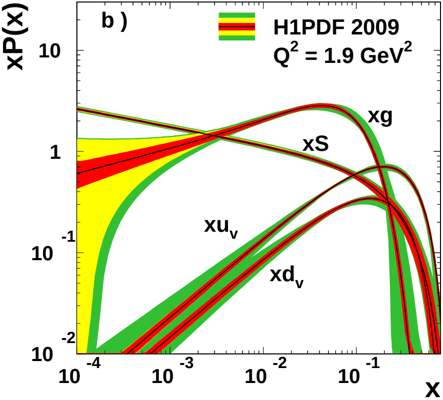
Multi-dimensional error analysis



- Establish a confidence region for $\{a_i\}$ for a given tolerated increase in χ^2
- In the ideal case of perfectly compatible Gaussian errors, 68% c.l. on a physical observable X corresponds to $\Delta\chi^2 = 1$ independently of the number N of PDF parameters

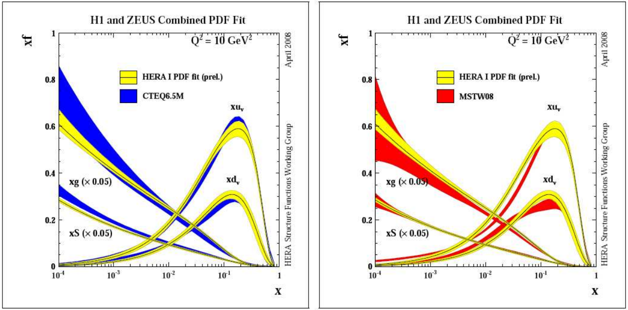
See, e.g., P. Bevington, K. Robinson, Data analysis and error reduction for the physical sciences

H1-2009 fit (arXiv:0904.3513)



- HERA-based fits are the closest to reproducing this ideal situation
- Example: the H1-2009 fit to the complete DIS data from HERA-1
- Color bands: experimental ($\Delta\chi^2 = 1$), theoretical, total uncertainty
- Heavy-flavor effects evaluated in GM-VFN scheme

HERAPDF0.1 set based on the combined H1+ZEUS data

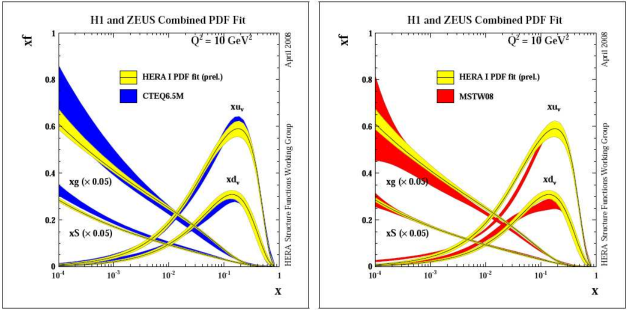


Updated HERAPDF0.2 fit was released this spring

The combined H1+ZEUS sample has a much smaller systematical uncertainty than the H1 and ZEUS samples individually

Nominally, very small uncertainty compared to CTEQ-MSTW-NNPDF!

HERAPDF0.1 set based on the combined H1+ZEUS data

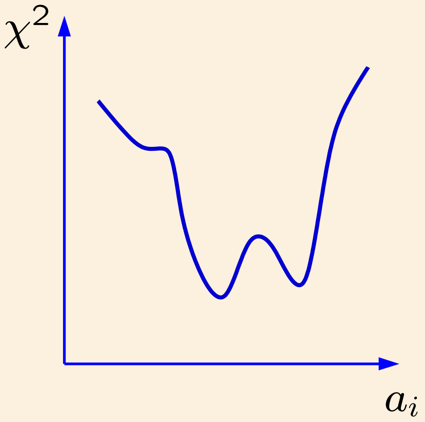


However:

- insufficient PDF flavor separation [neutral-current DIS probes only $4/9\,(u+\bar{u}+c+\bar{c})+1/9\,(d+\bar{d}+s+\bar{s})$]
- too rigid PDF parametrizations $\Rightarrow$ less flexibility to probe all allowed PDF behavior, notably at small x
- typical gluon forms, e.g., $g(x,Q_0)=Ax^B(1-x)^C\,(1+Dx)$, are ruled out by the Tevatron jet data *(Pumplin et al., arXiv:0904.2424)*

But if we combine the HERA data with the other experiments:

Multi-dimensional error analysis

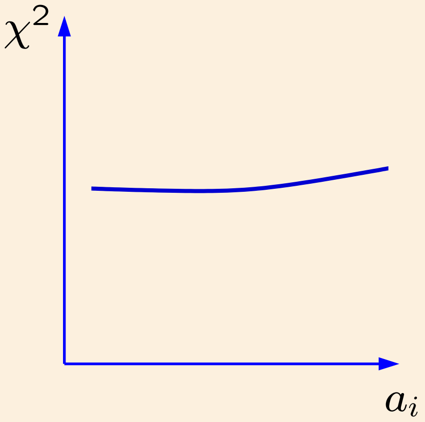


Pitfalls to avoid

- "Landscape"
  - disagreements between the experiments

In the worst situation, significant disagreements between M experimental data sets can produce up to $N \sim M!$ possible solutions for PDF's, with $N \sim 10^{500}$ reached for "only" about 200 data sets

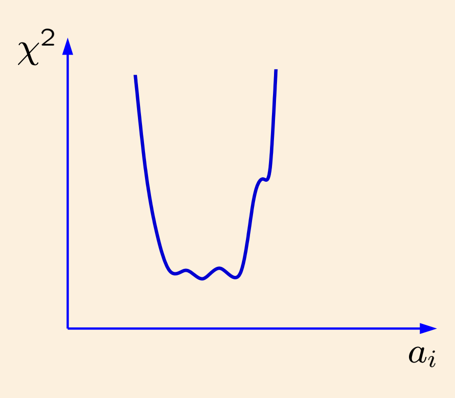
Multi-dimensional error analysis



Pitfalls to avoid

- Flat directions
  - unconstrained combinations of PDF parameters
  - dependence on free theoretical parameters, especially in the PDF parametrization
  - impossible to derive reliable PDF error sets

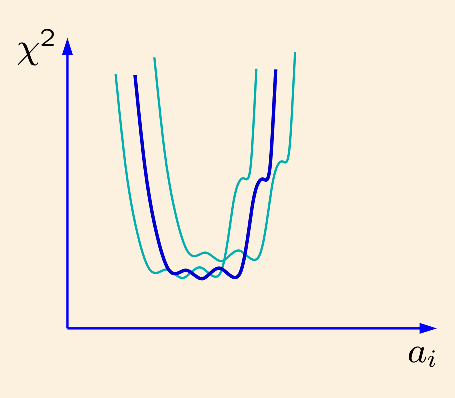
Multi-dimensional error analysis



The actual χ^2 function shows

- a well pronounced global minimum χ_0^2
- weak tensions between data sets in the vicinity of χ_0^2 (mini-landscape)
- some dependence on assumptions about flat directions

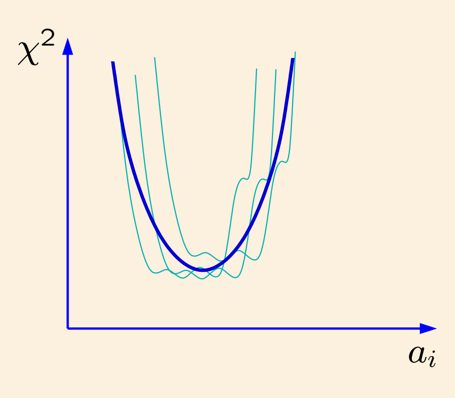
Multi-dimensional error analysis



The actual χ^2 function shows

- a well pronounced global minimum χ_0^2
- weak tensions between data sets in the vicinity of χ_0^2 (mini-landscape)
- some dependence on assumptions about flat directions

Multi-dimensional error analysis

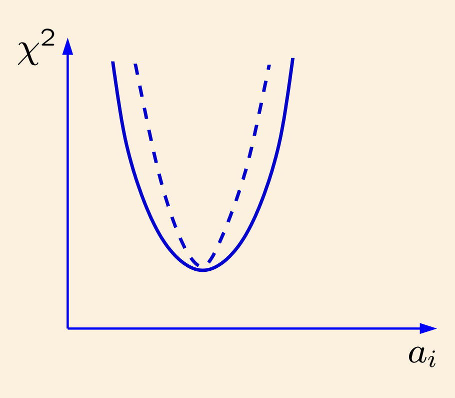


The actual χ^2 function shows

- a well pronounced global minimum χ_0^2
- weak tensions between data sets in the vicinity of χ_0^2 (mini-landscape)
- some dependence on assumptions about flat directions

The likelihood is approximately described by a quadratic χ^2 with a revised tolerance condition $\Delta\chi^2 \leq T^2$

Multi-dimensional error analysis

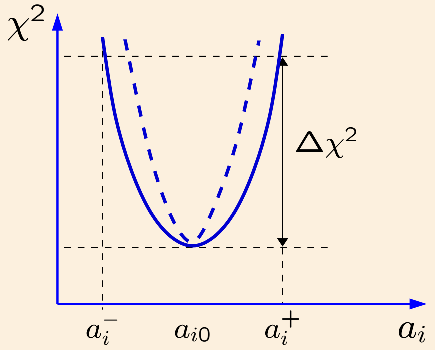


The actual χ^2 function shows

- a well pronounced global minimum χ_0^2
- weak tensions between data sets in the vicinity of χ_0^2 (mini-landscape)
- some dependence on assumptions about flat directions

The likelihood is approximately described by a quadratic χ^2 with a revised tolerance condition $\Delta\chi^2 \leq T^2$

Multi-dimensional error analysis



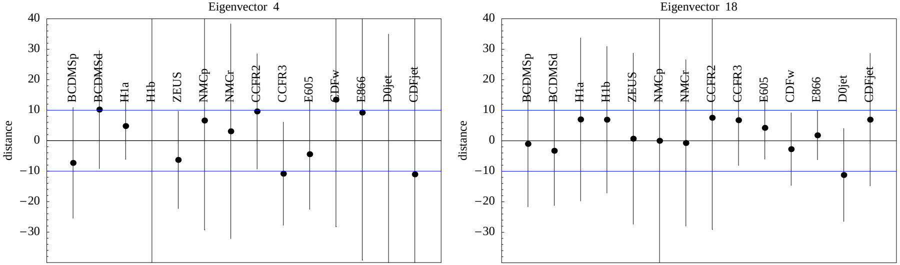
The actual χ^2 function shows

- a well pronounced global minimum χ_0^2
- weak tensions between data sets in the vicinity of χ_0^2 (mini-landscape)
- some dependence on assumptions about flat directions

The likelihood is approximately described by a quadratic χ^2 with a revised tolerance condition $\Delta\chi^2 \leq T^2$

CTEQ6 tolerance criterion (2001)

Acceptable values of PDF parameters must agree at ≈90% c.l. with all experiments included in the fit, *for a plausible range of assumptions about the PDF parametrization, scale dependence, experimental systematics, ...*

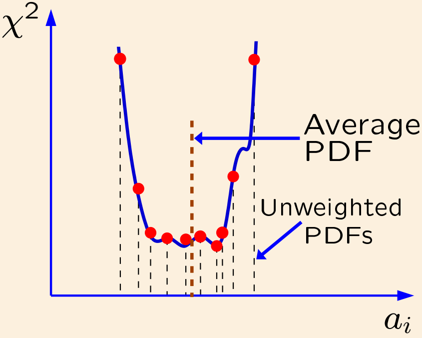


Can be crudely approximated (but does not have to) by assuming $T \approx 10$ for all PDF parameters

A somewhat stricter variant of this criterion is applied in the MSTW'08 analysis

Confidence intervals in global PDF analyses

Monte-Carlo sampling of the PDF parameter space



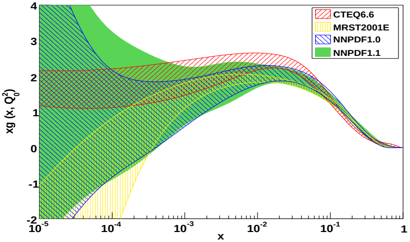
A very general approach that

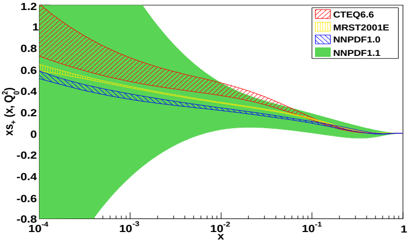
- realizes stochastic sampling of the probability distribution

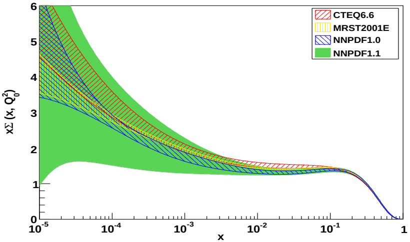
(Alekhin; Giele, Keller, Kosower; NNPDF)

- can parametrize PDF's by flexible neural networks (NNPDF)
- does not rely on smoothness of χ^2 or Gaussian approximations

NNPDF1.1 vs. other PDFs at $Q^2 = 2$ GeV^2 (arXiv:0811.2288)







At $x \lesssim 10^{-3}$, gluon g, strangeness $s_+ = (s + \bar{s})/2$, **and** singlet $\Sigma = \sum_i (q_i + \bar{q}_i)$ PDFs are **poorly** constrained;

determined by a "theoretically motivated" functional form in CTEQ/MSTW, flexible neural net in NNPDF; g, s_+ can be < 0!

PDF uncertainties: what they mean for you

Propagation of PDF errors into practical applications

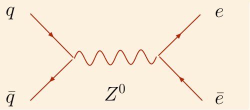
Z production at the LHC

Choose all that apply and select the x range

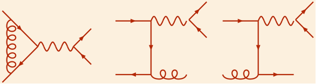
The PDF uncertainty in σ_Z is mostly due to...

1. u, d, $\bar{u}$, $\bar{d}$ PDF's
   at $x < 10^{-2}$ ($x > 10^{-2}$)
2. gluon PDF
   at $x < 10^{-2}$ ($x > 10^{-2}$)
3. s, c, b PDF's
   at $x < 10^{-2}$ ($x > 10^{-2}$)

Leading order



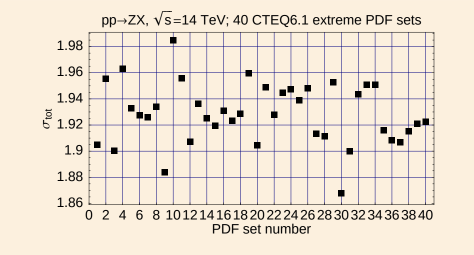
Next-to-leading order



An inefficient application of the error analysis

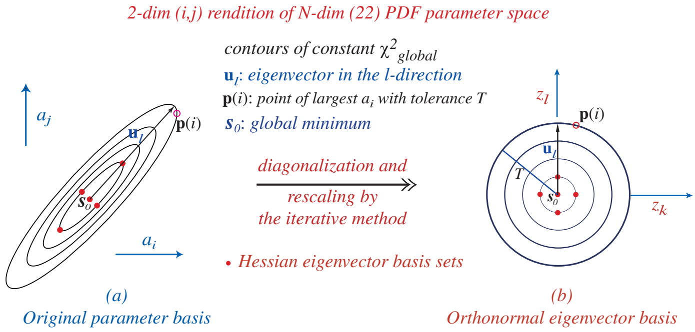
☺ Compute σ_Z for 40 (now 44) extreme PDF eigensets

☺ Find eigenparameter(s) producing largest variation(s), such as #9, 10, 30



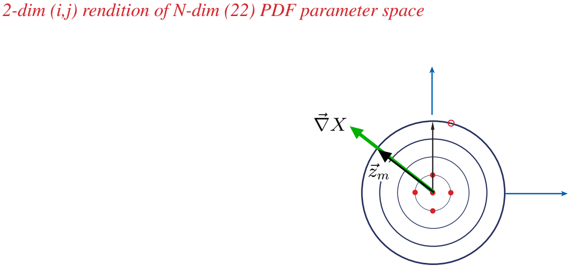
☹ It is not obvious how to relate abstract eigenparameters to physical PDF's $u(x)$, $d(x)$, etc.

Tolerance hypersphere in the PDF space



A hyperellipse $\Delta\chi^2 \leq T^2$ in space of N physical PDF parameters $\{a_i\}$ is mapped onto a filled hypersphere of radius T in space of N orthonormal PDF parameters $\{z_i\}$

Tolerance hypersphere in the PDF space



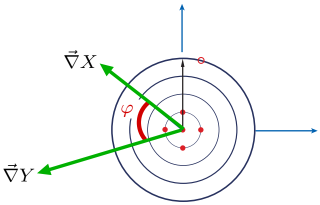
(b)
Orthonormal eigenvector basis

PDF error for a physical observable X is given by

$$\Delta X = \vec{\nabla} X \cdot \vec{z}_m = \left| \vec{\nabla} X \right| = \frac{1}{2} \sqrt{\sum_{i=1}^{N} \left(X_i^{(+)} - X_i^{(-)} \right)^2}$$

Tolerance hypersphere in the PDF space

2-dim (i,j) rendition of N-dim (22) PDF parameter space



(b)
Orthonormal eigenvector basis

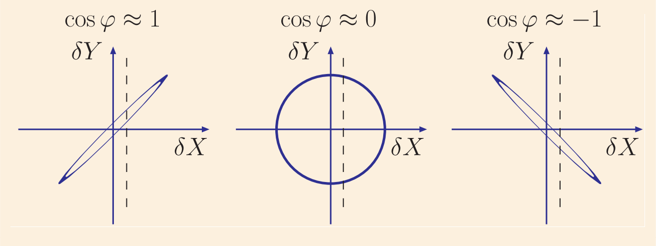
Correlation cosine for observables X and Y:

$$\cos\varphi = \frac{\vec{\nabla}X\cdot\vec{\nabla}Y}{\Delta X\,\Delta Y} = \frac{1}{4\Delta X\,\Delta Y}\sum_{i=1}^{N}\left(X_i^{(+)} - X_i^{(-)}\right)\left(Y_i^{(+)} - Y_i^{(-)}\right)$$

Correlation angle φ

Determines the parametric form of the $X - Y$ correlation ellipse

$$X = X_0 + \Delta X \cos\theta$$
$$Y = Y_0 + \Delta Y \cos(\theta + \varphi)$$



X_0, Y_0: best-fit values

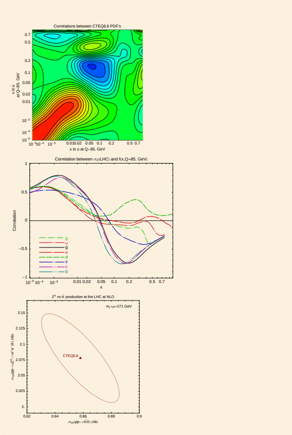
ΔX, ΔY: PDF errors

$\cos\varphi \approx \pm 1$: Measurement of X imposes tight constraints on Y
$\cos\varphi \approx 0$: Measurement of X imposes loose constraints on Y

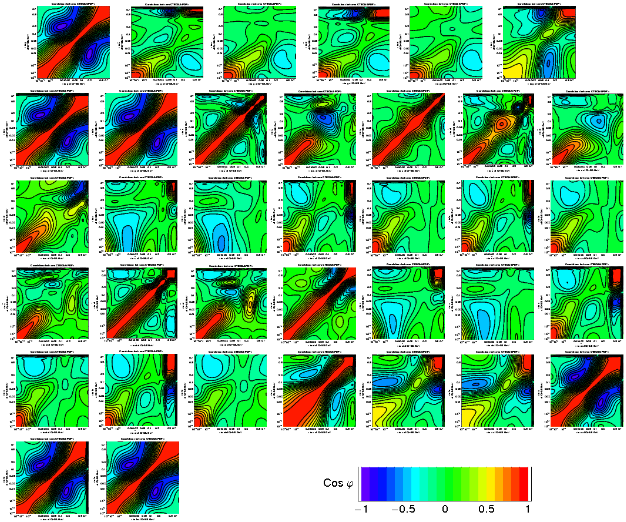
Types of correlations

X and Y can be

- two PDFs $f_1(x_1, Q_1)$ and $f_2(x_2, Q_2)$ (plotted as $\cos\varphi$ vs x_1 & x_2)
- a physical cross section σ and PDF $f(x, Q)$ (plotted as $\cos\varphi$ vs x)
- two cross sections σ_1 and σ_2



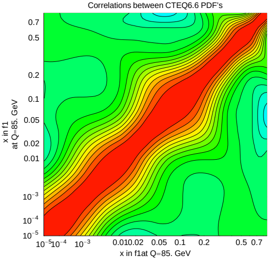
Correlations between $f_1(x_1, Q)$ and $f_2(x_2, Q)$ at $Q = 85$ GeV



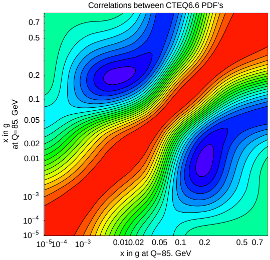
Figures from http://hep.pa.msu.edu/cteq/public/6.6/pdfcorrs/

Correlations between $f(x_1, Q)$ and $f(x_2, Q)$ at $Q = 85$ GeV

$u(x_1, Q)$ vs. $u(x_2, Q)$

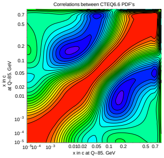


$g(x_1, Q)$ vs. $g(x_2, Q)$

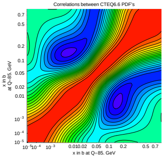


Correlation patterns look similar for g, c, b PDF's (no intrinsic charm here!)

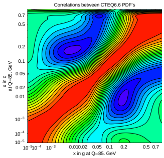
c vs. c



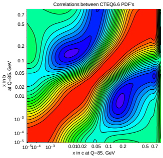
b vs. b



c vs. g

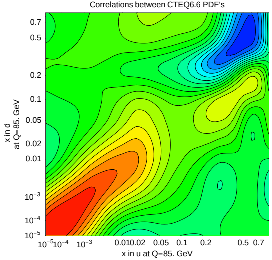


b vs. c

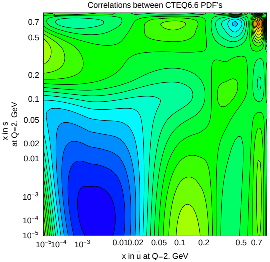


Correlations between $f_1(x_1, Q)$ and $f_2(x_2, Q)$ at $Q = 85$ GeV

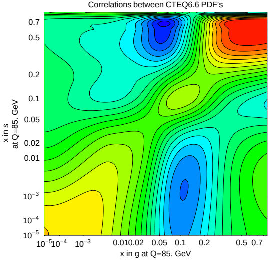
d vs u

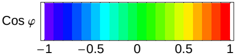


s vs $\bar{u}$ at Q=2 GeV



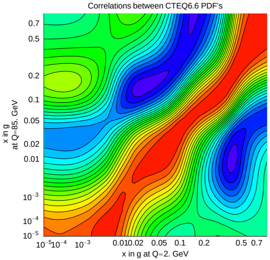
s vs. g





Sometimes there is a clear physics reason behind the correlation (e.g., sum rules or assumed Regge-like behavior); sometimes not

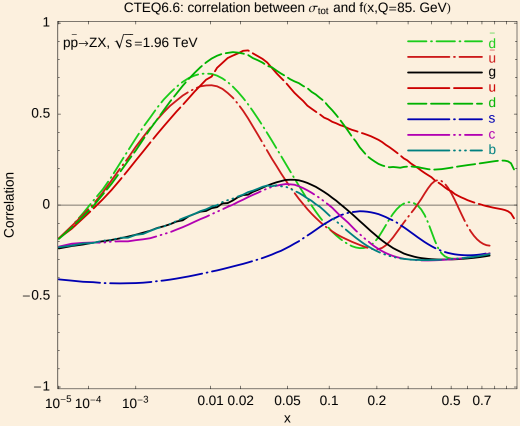
Correlations between $g(x_1, 2\ \text{GeV})$ and $g(x_2,\ 85\ \text{GeV})$



Gluons at $Q = 85$ GeV are correlated with gluons at $Q = 2$ GeV and larger x because of DGLAP evolution

Correlations $\cos\varphi$ between W, Z cross sections and PDF's

Tevatron Run-2

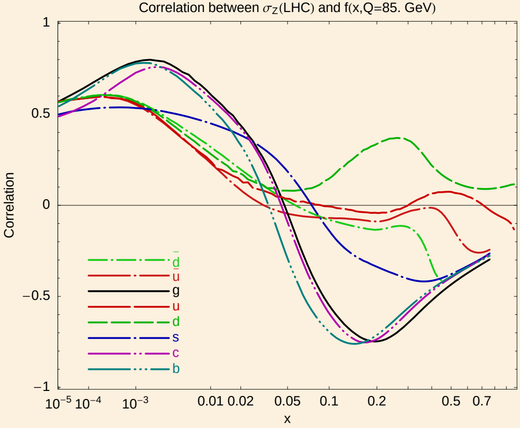


The largest correlations are with $u(0.05, M_Z)$ and $d(0.05, M_Z)$

Similar correlations for W production

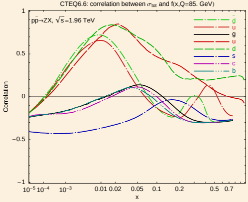
Correlations $\cos\varphi$ between W, Z cross sections and PDF's

LHC



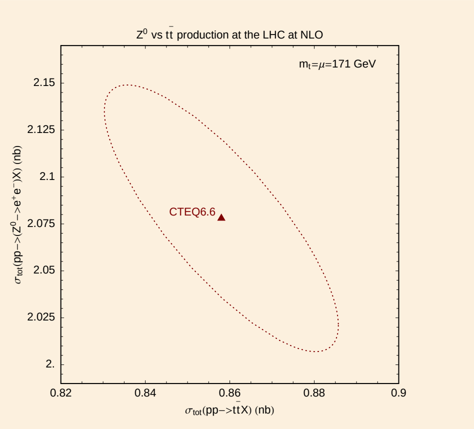
Strong correlation with $g(0.005, M_Z)$, c, b;
anticorrelation with $g(0.15, M_Z)$

Tevatron Run-2



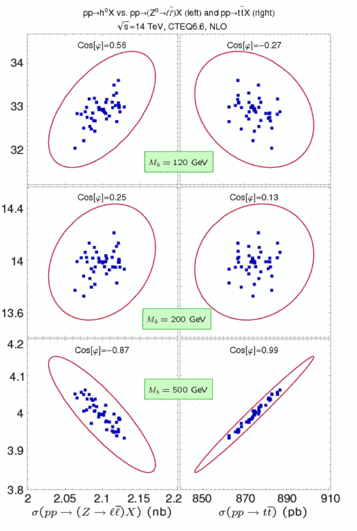
Similar correlations for W production

$t\bar{t}$ vs Z cross sections at the LHC



Measurements of $\sigma_{t\bar{t}}$ and σ_Z probe the same (gluon) PDF degrees of freedom at different x; they are anticorrelated because of the momentum sum rule (increasing $g(x,Q)$ at large x forces it to decrease at small x)

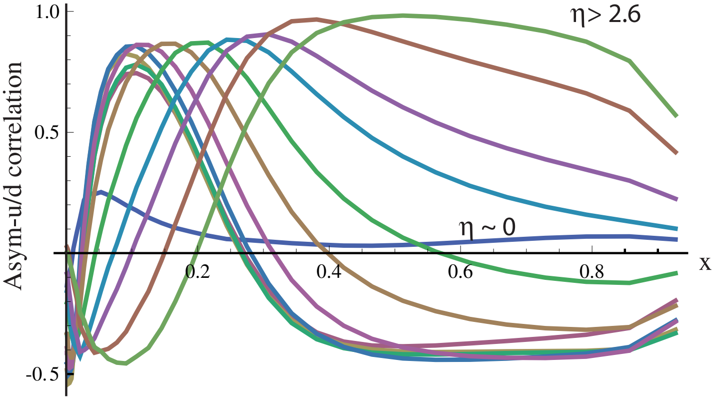
Correlations between $\sigma(gg \to H^0)$, σ_Z, $\sigma_{t\bar{t}}$



As M_H increases:

- $\cos\varphi(\sigma_H, \sigma_Z)$ decreases
- $\cos\varphi(\sigma_H, \sigma_{t\bar{t}})$ increases

CDF and D0 Run-2 W asymmetry $A_\ell(y)$

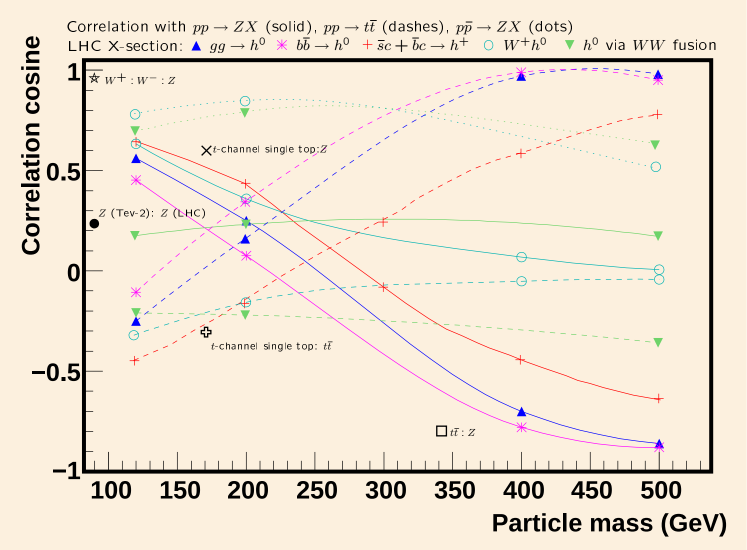


Correlation of $A_\ell(y)$ in different η_e bins ($p_{Te} > 35$ GeV) with $u(x)/d(x)$

(H. Schellman)

$\eta_e \sim 0$ is mostly sensitive to $d(x)/u(x)$ at $x \sim 0.1$; $\eta_e > 2.6$ to $x > 0.4$

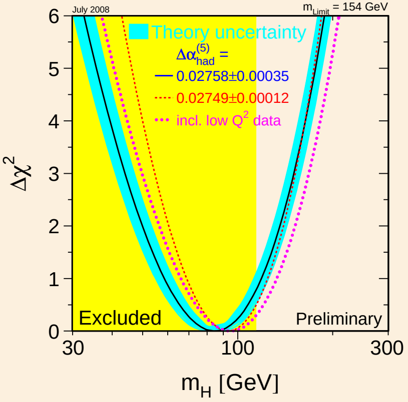
cos φ for various NLO Higgs production cross sections in SM and MSSM



Key Tevatron/LHC measurements require trustworthy PDFs

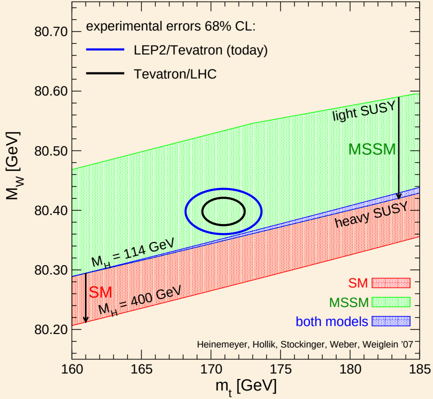
For example, leading syst. uncertainties in tests of electroweak symmetry breaking are due to insufficiently known PDFs

EW precision fits



A large part of δM_H arises from $\delta_{PDF} M_W$

EW fits + direct Higgs searches



SM band: $114 \leq M_H \leq 400$ **GeV**
SUSY band: random scan

Origin of differences between PDF sets

1. Corrections of wrong or outdated assumptions

lead to significant differences between new (≈post-2007) and old (≈pre-2007) PDF sets

- inclusion of (N)NLO QCD, heavy-quark hard scattering contributions
  - CTEQ6.6 and MSTW'2008 PDFs implement complete heavy-quark treatment; previous PDFs are obsolete without it
  - "NNLO" contributions are not automatically equivalent to better theory; to claim that, instabilities at small x or near heavy-quark thresholds must be also "tamed"
- relaxation of ad hoc constraints on PDF parametrizations
- improved numerical approximations

Origin of differences between PDF sets

2. PDF uncertainty

a range of allowed PDF shapes for plausible input assumptions, **partly** reflected by the PDF error band

is associated with

- the choice of fitted experiments
- experimental errors propagated into PDF's
- handling of inconsistencies between experiments
- choice of factorization scales, parametrizations for PDF's, higher-twist terms, nuclear effects,...

leads to non-negligible differences between the newest PDF sets

Conclusion

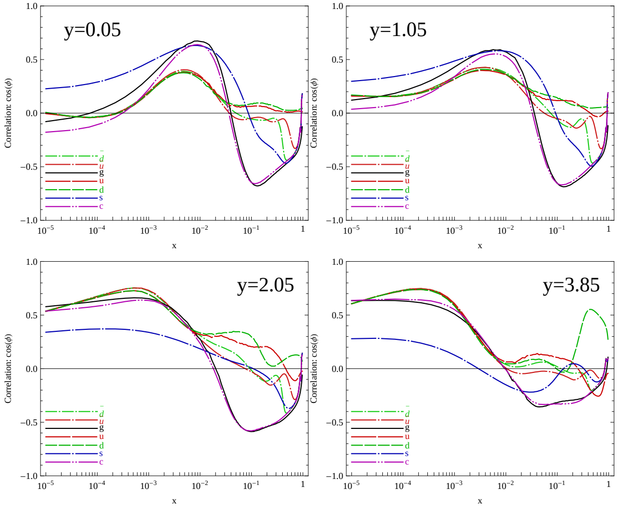
PDF analysis remains a fertile soil for novel contributions

- demand for **powerful calculations**
  - 2- and 3-loop QCD hard cross sections & matching coefficients for heavy-quark DIS, jet production
  - small-x, large-x, NLO EW effects (comparable to NNLO QCD)
- ample room for **physics judgement and ingenuity**
  - challenging issues in estimation, propagation of PDF uncertainties
  - flavor dependence of nonperturbative PDFs, isospin violation
- **close connection to experiment**
  - active studies of PDF effects in LHC, HERA, RHIC, Tevatron, and other measurements

Backup slides

Correlations between $d\sigma(pp \to Z^0 X)/dy$ and PDF's

$\cos\varphi$ between $d\sigma(pp \to Z^0 X)/dy$ at the LHC ($\sqrt{s} = 10$ TeV) and PDFs $f(x, Q = 85$ GeV$)$



Notice the change in sensitivity to parton flavors and the shift in the most relevant x range