Remarks on systematic uncertainties in PDF analyses

A proton at a collider

Pavel Nadolsky

Southern Methodist University

With A. Courtoy, T. Hobbs, X. Jing, J. Huston, K. Xie, C.-P. Yuan

and CTEQ-TEA (Tung Et Al.) Global QCD analysis group

CTEQ

Replicability risks for precision HEP

Nearly all complex STEM fields encounter replicability challenges.

Modern particle physics is not an exception.

1. It is complex! Is it rigorous enough?
   - Many approaches, especially AI-based ones, increase complexity and are not rigorously understood
2. It often uses wrong prescriptions for estimating epistemic uncertainties
   - Tens to hundreds of systematic uncertainties affect measurements, phenomenology, and lattice QCD

Ongoing studies of systematic uncertainties are essential and still insufficient

- from the experiment side

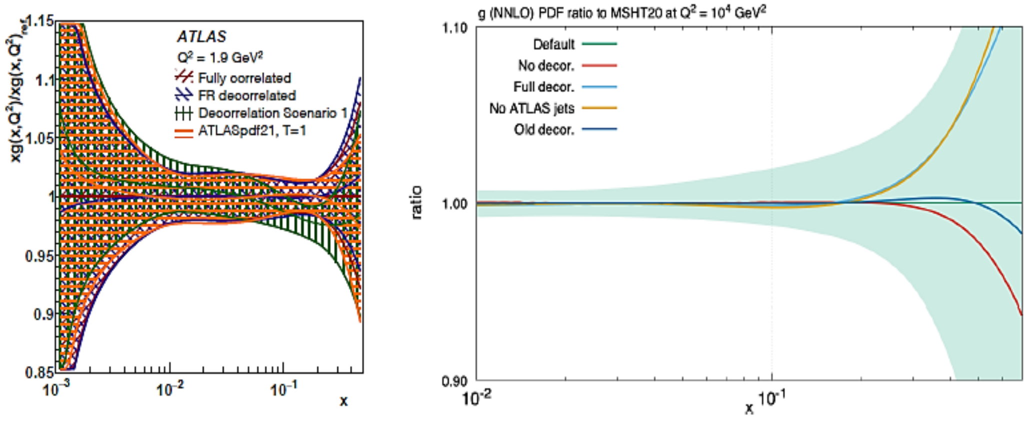


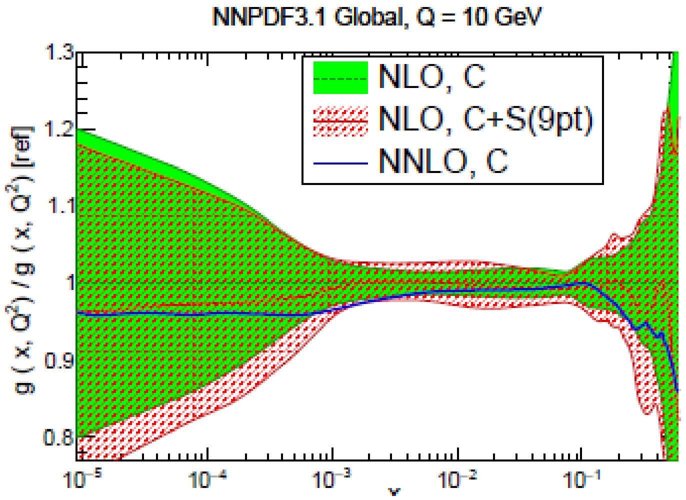
FIG. 9. Difference in the gluon PDF shown in ratio to the ATLASpdf21 (default) gluon(left). This default uses Decorrelation Scenario 2 and this is compared to the use of Full Correlation, Full decorrelation of the flavour response systematic and Decorrelation Scenario 1. The effect of no decorrelation, the default correlation of [9], the decorrelation in [362], and full decorrelation for the MSHT20 gluon (right).

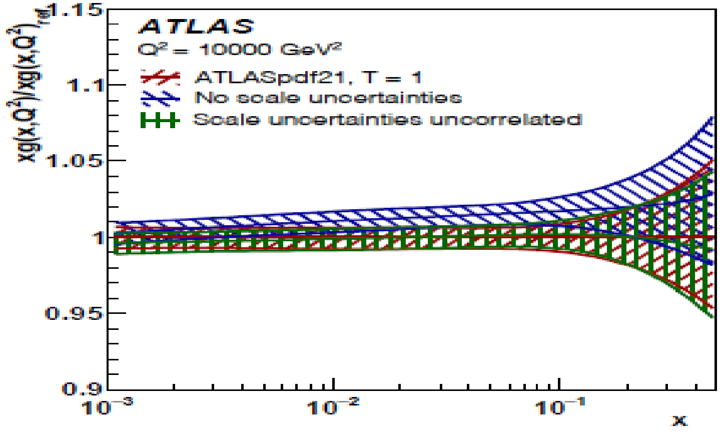
S. Amoroso et al., 2203.13923, Sec. 5.A

Strong dependence on the definition of corr. syst. errors raises a general concern:

Overreliance on Gaussian distributions and covariance matrices for poorly understood effects may produce very wrong uncertainty estimates
[N. Taleb, Black Swan & Antifragile]

- from the theory side



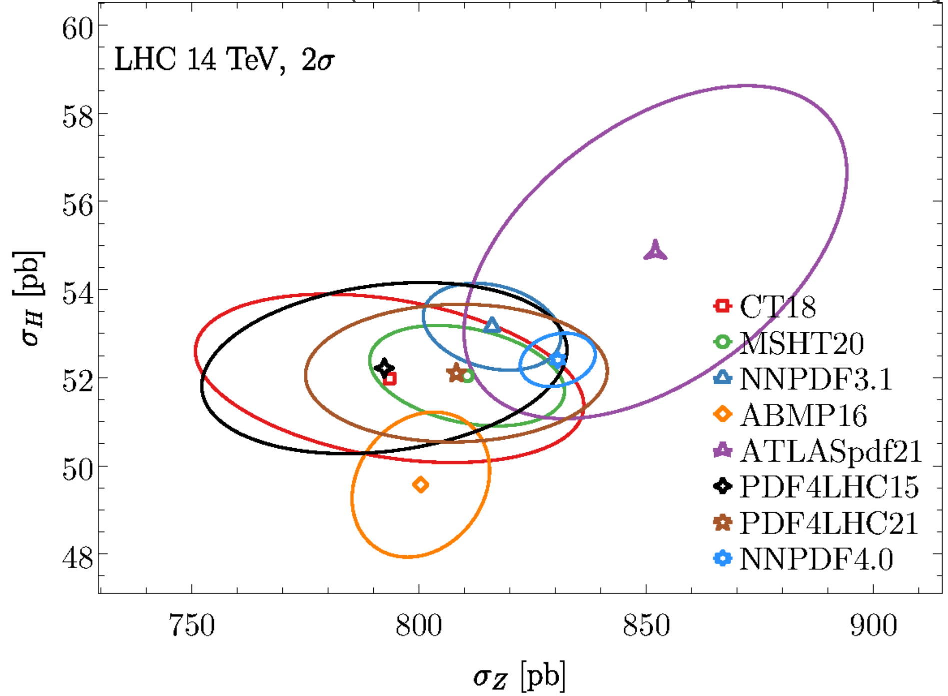


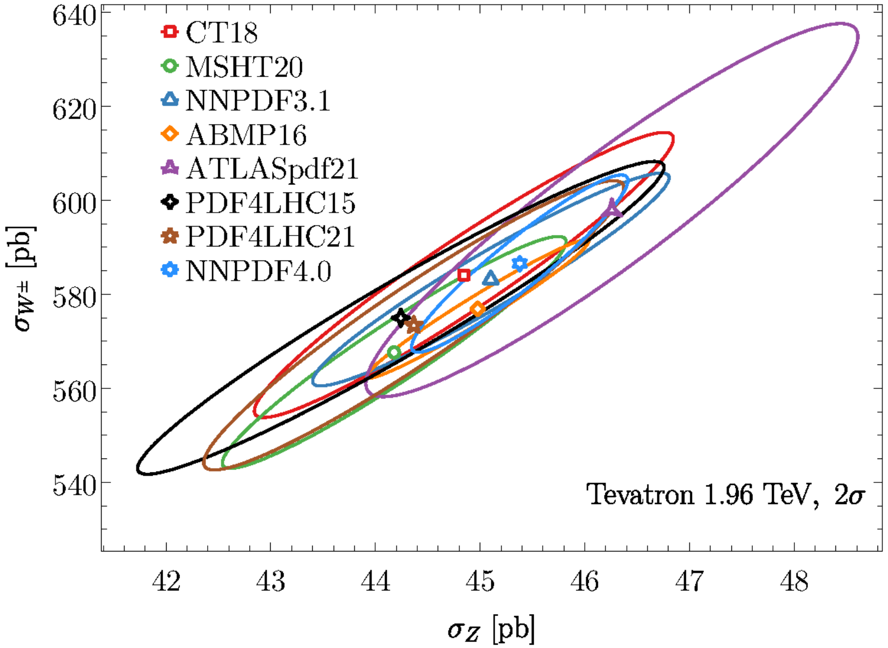
Examples: studies of theory uncertainties in the PDFs by NNPDF3.1 and ATLAS21

The tolerance puzzle

Why do groups fitting similar data sets obtain different PDF uncertainties?

Precision PDFs (Snowmass 21 WP) [2203.13923v2]





The answer has direct implications for high-stake experiments such as W boson mass measurement, tests of nonperturbative QCD models and lattice QCD, high-mass BSM searches, etc.

Tensions among experiments

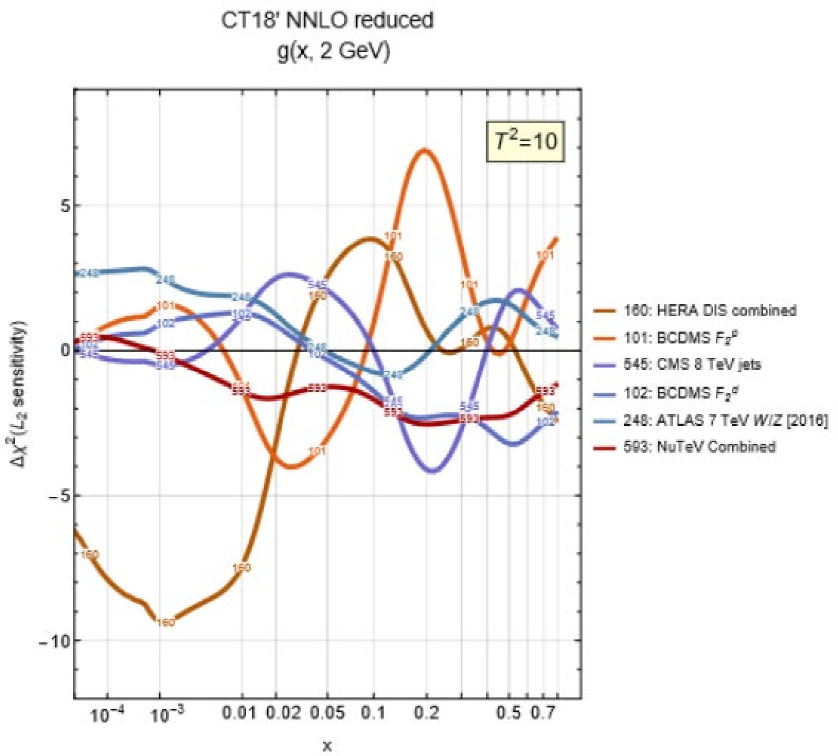
Explore using the L_2 sensitivity

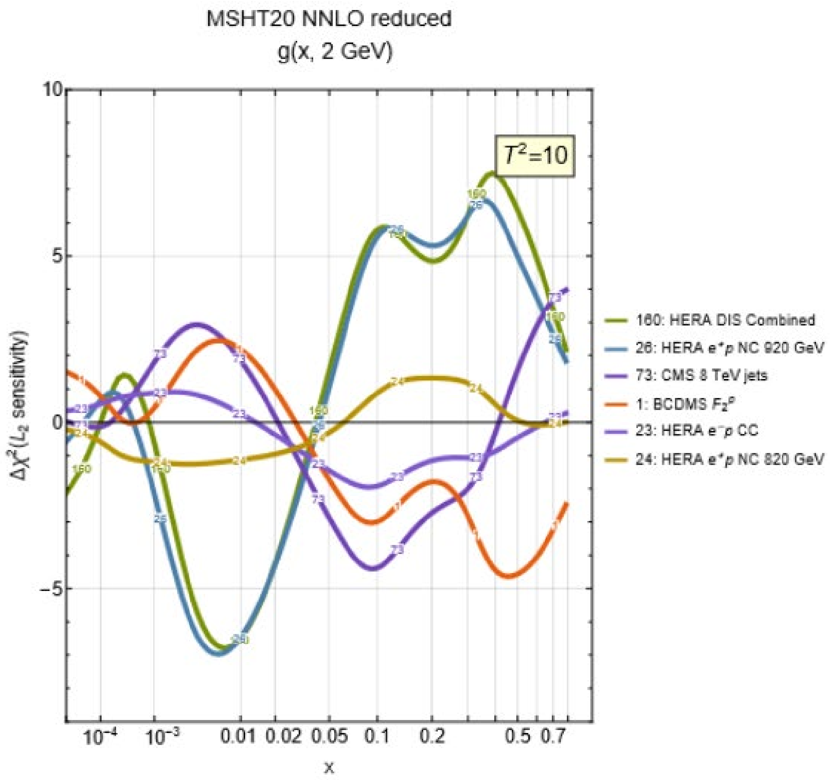
for Hessian PDFs

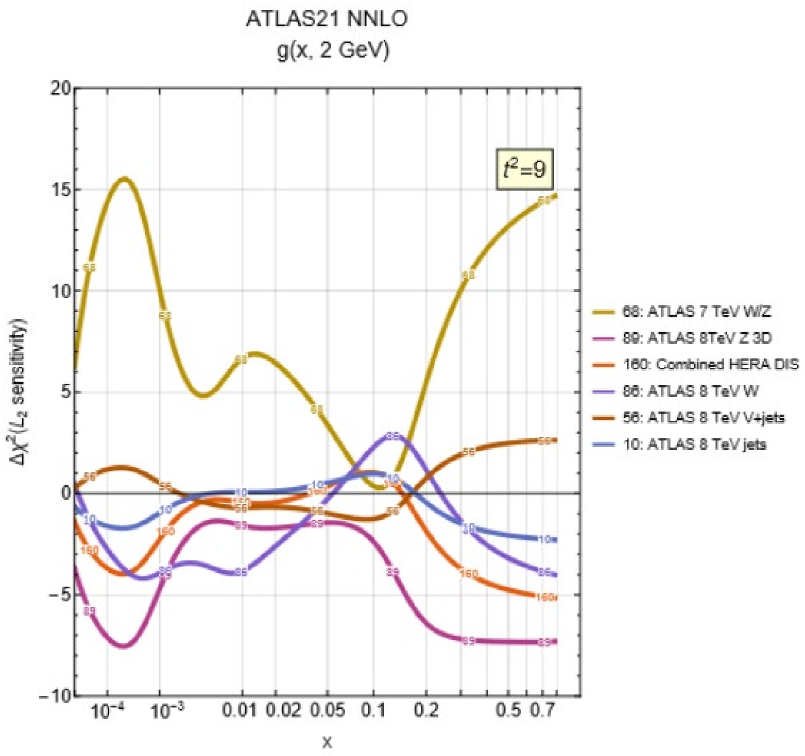
arXiv:2306.03918
by X. Jing, A. Cooper-Sarkar, A. Courtoy, T. Cridge, F. Giuli, L. Harland-Lang, T.J. Hobbs, J. Huston, P. N., R. S. Thorne, K. Xie, C.-P. Yuan

An ATLAS, CTEQ-TEA, and MSHT comparative study of NNLO and aN3LO PDF sensitivities

X.Jing et al.,arXiv:2306.03918







- Comparisons of strengths of constraints from individual data sets in 8 PDF analyses using the common L_2 sensitivity metric
- An interactive website (https://metapdf.hepforge.org/L2/) to plot such comparisons
  [2070 figures in total]

Dependence on implementations of systematic uncertainties

Explore using a hopscotch scan

for MC PDFs

arXiv:2205.10444 [PRD 107 (2023) 3, 034008]

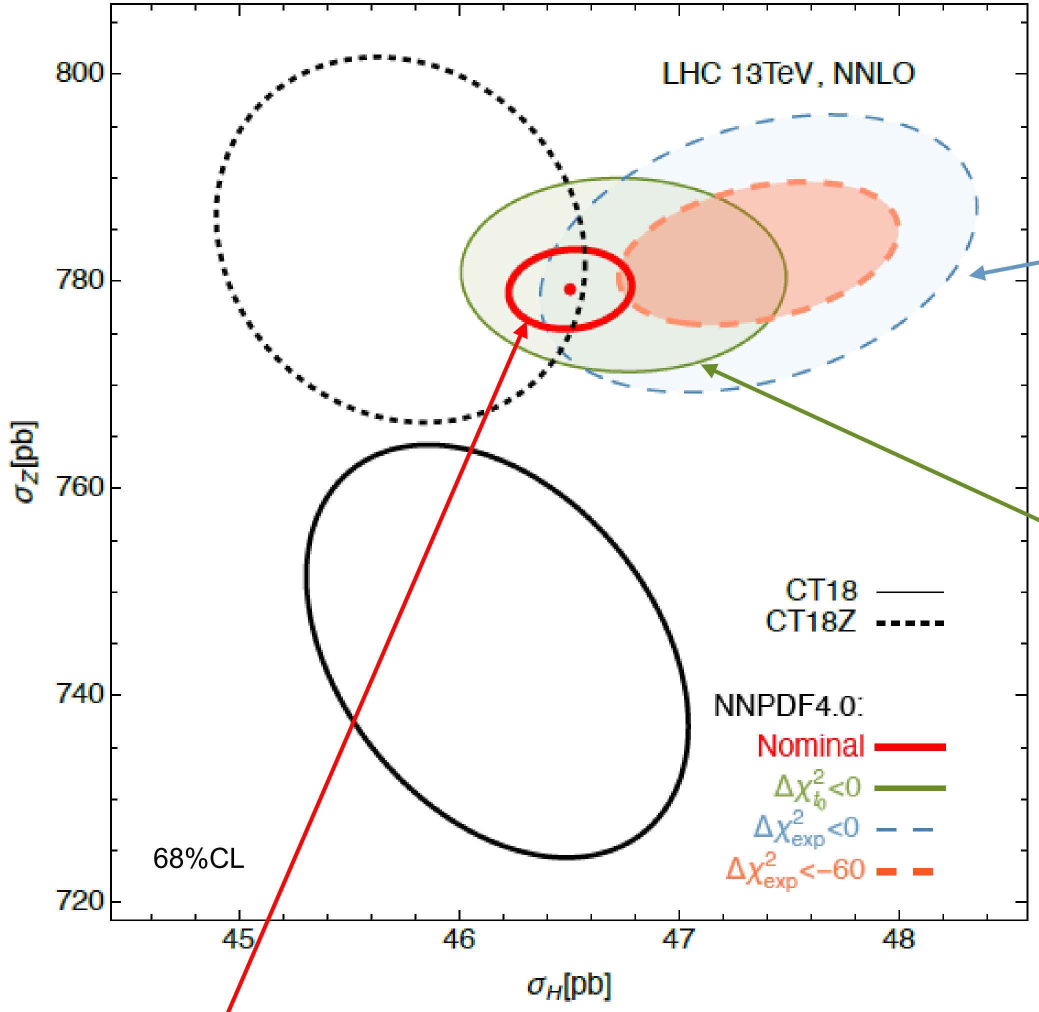
by A. Courtoy, J. Huston, P. N., K. Xie, M. Yan, C.-P. Yuan

Goodness-of-fit functions in PDF analyses

Analysis	χ^2 prescription to fit PDFs	χ^2 prescription to compare PDFs	Comments
HERAPDF	HERA	HERA	
CT	Extended T +addl. prior	Extended T, Experimental	
MSHT'20	T	T	
NNPDF4.0	t_0 +addl. prior with fluctuated cross-sampled data	Experimental or t_0 with unfluctuated full data	t_0 prescription has pre- and post-NNPDF3.0 versions
…			
Hopscotch'2022	N/A	Experimental or t_0 [2022] with unfluctuated data	

Different prescriptions reflect modeling of additive and multiplicative systematic errors in covariance matrices. Neither prescription is complete because of the bias-variance dilemma. The χ^2 definition affects the PDF uncertainty.

Hopscotch scan+sampling of PDF parametrizations



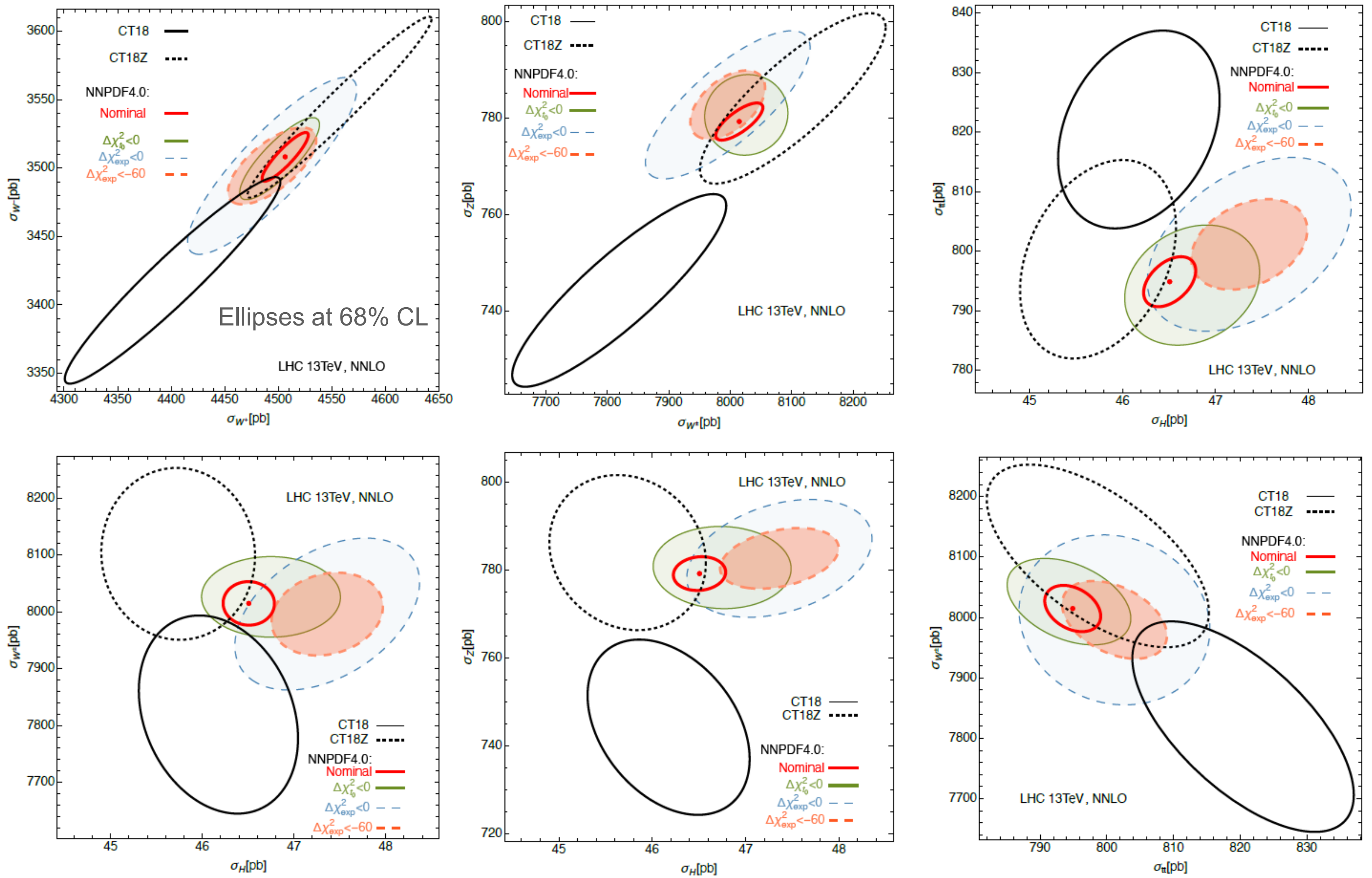
Regions containing (very) good solutions according to the experimental form of χ^2 (is used in χ^2 summary tables of the NN4.0 article, is used in the NN4.0 public code when not doing the fits)

Region containing good solutions according to the NNPDF3.0 t_0 form of χ^2 (used to train NN4.0 replicas)

Nominal NN4.0 Hessian or MC 68%cl

These regions are approximate, at least as large as shown arXiv:2205.10444

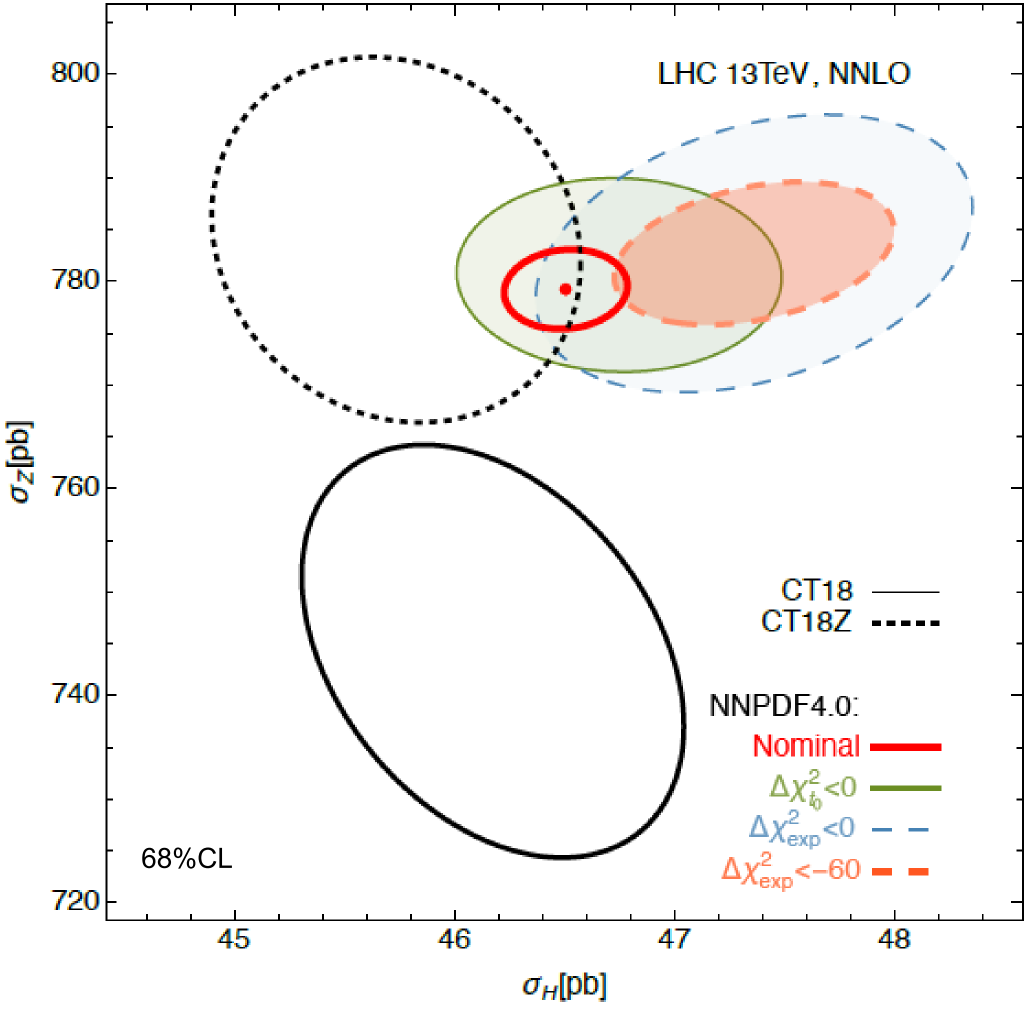
The hopscotch scans: NNPDF4.0 vs CT18 uncertainties



The ellipses are projections of 68% c.l. ellipsoids in N_{par}-dim. spaces

$N_{par} = 28$ and 50 for CT18 and NNPDF4.0 Hessian PDFs

Hopscotch scans realize the likelihood-ratio test



According to the LR test, the NN4.0 analysis discards PDFs in the green and blue regions based on the prior probabilities and differences in the likelihood definitions – both associated with prior terms

The allowed regions will change for the other acceptable χ^2 definitions, which exist in reflection of the bias-variance dilemma

A likelihood-ratio test of NN models T_1 and T_2

From Bayes theorem, it follows that

$$\frac{P(T_2|D)}{P(T_1|D)} = \frac{P(D|T_2)}{P(D|T_1)} \times \frac{P(T_2)}{P(T_1)}$$

$\equiv r_{\mathrm{posterior}}$ $\equiv r_{\mathrm{likelihood}}$ $\equiv r_{\mathrm{prior}}$

aleatory epistemic + aleatory probabilities

Suppose replicas T_1 and T_2 have the same χ^2 $[r_{\mathrm{likelihood}} = \exp\left(\frac{\chi_1^2-\chi_2^2}{2}\right) = 1]$, but T_2 is disfavored compared to T_1 $[r_{\mathrm{posterior}} \ll 1]$.

This only happens if $r_{\mathrm{prior}} \ll 1$: T_2 is discarded based on its **prior** probability.

Two forms of χ^2 in PDF fits

1. In terms of nuisance parameters $\lambda_{\alpha,exp}$

$$\chi^2 = \sum_{i=1}^{N_{pt}} \frac{\left[D_i + \sum_\alpha \beta_{i,\alpha}^{\rm exp} \lambda_{\alpha,\rm exp} - T_i\right]^2}{s_i^2} + \sum_\alpha \lambda_{\alpha,\rm exp}^2$$

2. In terms of the covariance matrix

$$\chi^2 = \sum_{i,j}^{N_{pt}} (T_i - D_i)(\mathrm{cov}^{-1})_{ij}(T_j - D_j)$$

$$(\mathrm{cov})_{ij} \equiv s_i^2 \delta_{ij} + \sum_{\alpha=1}^{N_\lambda} \beta_{i,\alpha}\beta_{j,\alpha}, \qquad \beta_{i,\alpha} = \sigma_{i,\alpha} X_i,$$

D_i, T_i, s_i are the central data, theory, uncorrelated error
$\beta_{i,\alpha}$ is the correlation matrix for N_λ nuisance parameters.

Experiments publish $\sigma_{i,\alpha}$. To reconstruct $\beta_{i,\alpha}$, we need to decide on the normalizations X_i.
Possible choices:

a. $X_i = D_i$: "**exp**erimental scheme"; can result in a bias
b. $X_i =$ fixed or varied T_i : "$\boldsymbol{t_0}$, T, extended T schemes"; can result in (different) biases

Systematic uncertainties and the bias-variance dilemma

$$\chi^2 = \sum_{i,j}^{N_{pt}} (T_i - D_i)(\mathrm{cov}^{-1})_{ij}(T_j - D_j) \qquad (\mathrm{cov})_{ij} = s_i^2 \delta_{ij} + \sum_{\alpha=1}^{N_\lambda} \beta_{i,\alpha}\beta_{j,\alpha}$$

$$\boxed{\beta_{i,\alpha} = \sigma_{i,\alpha} X_i}$$

D_i, T_i, s_i are the central data, theory, uncorrelated error

$\beta_{i,\alpha} \equiv \sigma_{i,\alpha}\hat{X}_i$ is the correlation matrix for N_λ nuisance parameters. Experiments publish $\sigma_{i,\alpha}$.

The "truth" normalizations $\hat{X}_i$ in the experiment are of order T_i or D_i. $\{\boldsymbol{\hat{X}_i}\}$ **are learned as a model** $\{\boldsymbol{X_i}\}$ **together with PDFs** $\boldsymbol{f}$ **and theory** $\{\boldsymbol{T_i(f)}\}$. For example, we can sample as $X_i = a_i D_i + b_i T_i$, with free $0 \le a_i, b_i \lesssim 1$.

Mean variation δ_X^2 of the model from truth on an ensemble of replicas, for data point i:

$$\delta_X^2 \equiv \left\langle \left(X_i - \hat{X}_i\right)^2 \right\rangle = \underbrace{\left\langle \left(\hat{X}_i - \langle X_i \rangle\right)^2 \right\rangle}_{\text{model bias}} + \underbrace{\langle (X_i - \langle X_i \rangle)^2 \rangle}_{\text{variance}} = \underbrace{\left\langle \left(\hat{X}_i - \langle X_i \rangle\right)^2 \right\rangle}_{\text{model bias}} - \underbrace{\langle (D_i - \langle X_i \rangle)^2 \rangle}_{\text{data bias}} + \underbrace{\langle (D_i - X_i)^2 \rangle}_{\chi^2(D_i, T_i)}$$

Experimental definition, $X_i = D_i$: $\left\langle \left(X_i - \hat{X}_i\right)^2 \right\rangle = \left(\hat{X}_i - D_i\right)^2 \equiv \delta_D^2$

t_0 definition, $X_i = t_{0i}$: $\left\langle \left(X_i - \hat{X}_i\right)^2 \right\rangle = \left(\hat{X}_i - t_{0i}\right)^2 \equiv \delta_{t_0}^2$

In general, not enough information to compare δ_D and δ_{t_0}

Systematic uncertainty from PDFs in W boson mass and α_s measurements

ATLAS-CONF-2023-004

PDF-Set	p_T^ℓ [MeV]	m_T [MeV]	combined [MeV]
CT10	$80355.6^{+15.8}_{-15.7}$	$80378.1^{+24.4}_{-24.8}$	$80355.8^{+15.7}_{-15.7}$
CT14	$80358.0^{+16.3}_{-16.3}$	$80388.8^{+25.2}_{-25.5}$	$80358.4^{+16.3}_{-16.3}$
CT18	$80360.1^{+16.3}_{-16.3}$	$80382.2^{+25.3}_{-25.3}$	$80360.4^{+16.3}_{-16.3}$
MMHT2014	$80360.3^{+15.9}_{-15.9}$	$80386.2^{+23.9}_{-24.4}$	$80361.0^{+15.9}_{-15.9}$
MSHT20	$80358.9^{+13.0}_{-16.3}$	$80379.4^{+24.6}_{-25.1}$	$80356.3^{+14.6}_{-14.6}$
NNPDF3.1	$80344.7^{+15.6}_{-15.5}$	$80354.3^{+23.6}_{-23.7}$	$80345.0^{+15.5}_{-15.5}$
NNPDF4.0	$80342.2^{+15.3}_{-15.3}$	$80354.3^{+22.3}_{-22.4}$	$80342.9^{+15.3}_{-15.3}$

Table 2: Overview of fitted values of the W boson mass for different PDF sets. The reported uncertainties are the total uncertainties.

ATLAS-CONF-2023-015

The statistical analysis for the determination of $\alpha_s(m_Z)$ is performed with the xFitter framework [60]. The value of $\alpha_s(m_Z)$ is determined by minimising a χ^2 function which includes both the experimental uncertainties and the theoretical uncertainties arising from PDF variations:

$$\chi^2(\beta_{\rm exp},\beta_{\rm th}) = \sum_{i=1}^{N_{\rm data}} \frac{\left(\sigma_i^{\rm exp} + \sum_j \Gamma_{ij}^{\rm exp}\beta_{j,\rm exp} - \sigma_i^{\rm th} - \sum_k \Gamma_{ik}^{\rm th}\beta_{k,\rm th}\right)^2}{\Delta_i^2} + \sum_j \beta_{j,\rm exp}^2 + \sum_k \beta_{k,\rm th}^2 \,. \quad (1)$$

profiling of CT and MSHT PDFs requires to include a tolerance factor $T^2 > 10$ as in the ePump code

[T.J. Hou et al., 1912.10053, Appendix F]

Also the next slide.

Augmented likelihood for PDFs with global tolerance

1. Start by defining the correspondence between $\Delta\chi^2$ and cumulative probability level: 68% c.l. $\Leftrightarrow$ $\Delta\chi^2 = T^2$.
2. Write the **augmented** likelihood density for this definition:

$$P(D_i|T_i) \propto e^{-\chi^2/(2T^2)}$$

3. When profiling 1 new experiment with the prior imposed on PDF nuisance parameters $\lambda_{\alpha,th}$:

$$\chi^2(\vec{\lambda}_{\rm exp},\vec{\lambda}_{\rm th}) = \sum_{i=1}^{N_{pt}} \frac{\left[D_i + \sum_\alpha \beta^{\rm exp}_{i,\alpha}\lambda_{\alpha,\rm exp} - T_i - \sum_\alpha \beta^{\rm th}_{i,\alpha}\lambda_{\alpha,\rm th}\right]^2}{s_i^2} + \sum_\alpha \lambda^2_{\alpha,\rm exp} + \sum_\alpha T^2\lambda^2_{\alpha,\rm th}. \qquad \beta^{\rm th}_{i,\alpha} = \frac{T_i(f_\alpha^+) - T_i(f_\alpha^-)}{2},$$

new experiment

priors on expt. systematics and PDF params

4. Alternatively, we can reparametrize $\chi^{2'} \equiv \chi^2/T^2$, so that 68% c.l. $\Leftrightarrow$ $\Delta\chi^{2\prime} = 1$. We have

$$P(D_i|T_i) \propto e^{-\chi^{2\prime}/2}$$

$$\chi^{2\prime}(\vec{\lambda}_{\rm exp},\vec{\lambda}_{\rm th}) = \sum_{i=1}^{N_{pt}} \frac{\left[D_i + \sum_\alpha \beta^{\rm exp}_{i,\alpha}\lambda_{\alpha,\rm exp} - T_i - \sum_\alpha \beta^{\rm th}_{i,\alpha}\lambda_{\alpha,\rm th}\right]^2}{s_i^2\ T^2} + \sum_\alpha \frac{\lambda^2_{\alpha,\rm exp}}{T^2} + \sum_\alpha \lambda^2_{\alpha,\rm th}.$$

consistent redefinition

5. Inconsistent redefinitions:

$$\chi^{2\prime}(\vec{\lambda}_{\rm exp},\vec{\lambda}_{\rm th}) = \sum_{i=1}^{N_{pt}} \frac{\left[D_i + \sum_\alpha \beta^{\rm exp}_{i,\alpha}\lambda_{\alpha,\rm exp} - T_i - \sum_\alpha \beta^{\rm th}_{i,\alpha}\lambda_{\alpha,\rm th}\right]^2}{s_i^2} + \sum_\alpha \lambda^2_{\alpha,\rm exp} + \sum_\alpha \lambda^2_{\alpha,\rm th}.$$

and $P(D_i|T_i) \propto e^{-\chi^{2\prime}/2}$

or $P(D_i|T_i) \propto e^{-\chi^{2\prime}/(2T^2)}$

[equivalent to $s_i \to s_i/T$ or $\lambda_{\alpha,th} \to \lambda_{\alpha,th}T$ without $\beta_{i,\alpha,th} \to \beta_{i,\alpha,th}/T$]

Why **augmented** likelihood?

The term is accepted in lattice QCD [G. P. Lepage et al., hep-lat/0110175] to indicate that the log-likelihood contains **prior terms**

$$\chi^2(\vec{\lambda}_{\rm exp}, \vec{\lambda}_{\rm th}) = \sum_{i=1}^{N_{pt}} \frac{\left[D_i + \sum_\alpha \beta_{i,\alpha}^{\rm exp}\lambda_{\alpha,\rm exp} - T_i - \sum_\alpha \beta_{i,\alpha}^{\rm th}\lambda_{\alpha,\rm th}\right]^2}{s_i^2} + \sum_\alpha \lambda_{\alpha,\rm exp}^2 + \sum_\alpha T^2 \lambda_{\alpha,\rm th}^2.$$

new experiment

priors on expt. systematics and PDF params

After minimization w.r.t. to $\lambda_{\alpha,exp}$, $\lambda_{\alpha,th}$, the prior terms are **hidden** inside the covariance matrix:

$$\chi^2 = \sum_{i,j}^{N_{pt}} (T_i - D_i)(\mathrm{cov}^{-1})_{ij}(T_j - D_j)$$

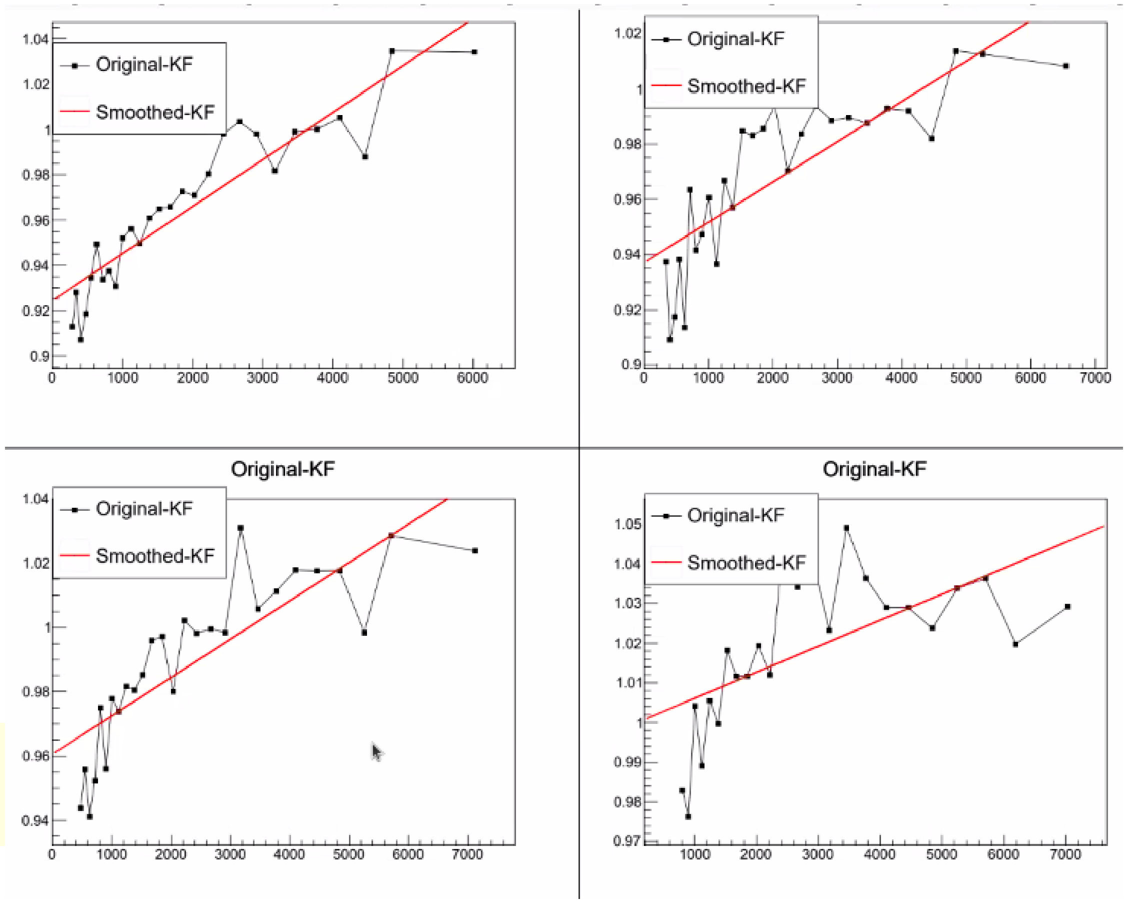
The usual χ^2 definition therefore contains a **prior** component, which may be handled differently by the various groups

Smoothing of K-factors

An analogous **bias-variance tradeoff** arises during smoothing of MC integration errors for K-factor tables

A smoother curve for theory reduces the χ^2 for the jet data, but the best-fit result retains some dependence on the fitted functional form

This dependence can be conservatively estimated by including an uncorrelated MC integration error



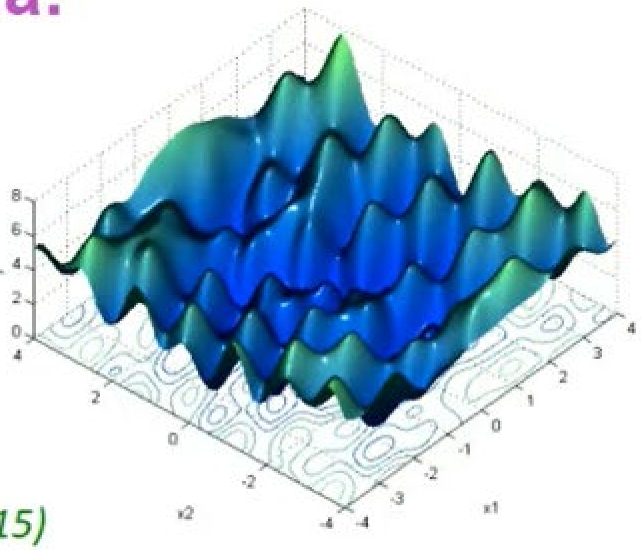
NNLO/NLO ratios for LHC 13 TeV jet production

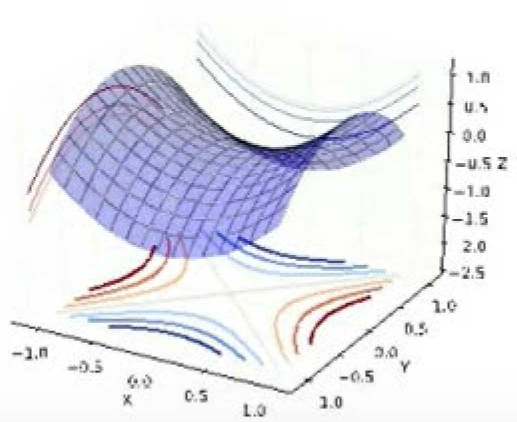
Not so terrible local minima: convexity is not needed

Myth busted:

- Local minima dominate in low-D, but saddle points dominate in high-D
- Most local minima are relatively close to the bottom (global minimum error)

(Dauphin et al NIPS'2014, Choromanska et al AISTATS'2015)





Global minimum: all $\frac{\partial^2\chi^2}{\partial a_i \partial a_j} > 0$ (improbable)

Saddle point: some $\frac{\partial^2\chi^2}{\partial a_i \partial a_j} > 0$ (probable)

An average global minimum: in properly chosen coordinates, $\frac{\partial^2\chi^2}{\partial z_i \partial z_j} > 0$ for dominant coordinate components

Y. Bengio, 2019 Turing lecture (YouTube)

Many dimensions introduce major difficulties with finding a global minimum…

An important question concerns the distribution of critical points (maxima, minima, and saddle points) of such functions. Results from random matrix theory applied to spherical spin glasses have shown that these functions have a combinatorially large number of saddle points. Loss surfaces for large neural nets have many local minima that are essentially equivalent from the point of view of the test error, and these minima tend to be highly degenerate, with many eigenvalues of the Hessian near zero.

We empirically verify several hypotheses regarding learning with large-size networks:

- For large-size networks, most local minima are equivalent and yield similar performance on a test set.
- The probability of finding a "bad" (high value) local minimum is non-zero for small-size networks and decreases quickly with network size.
- Struggling to find the global minimum on the training set (as opposed to one of the many good local ones) is not useful in practice and may lead to overfitting.

The Loss Surfaces of Multilayer Networks
A. Choromanska, M. Henaff, M. Mathieu, G. Ben Arous, Y. LeCun PMLR 38:192-204, 2015

Many dimensions introduce major difficulties with finding a global minimum…

…as well as with representative exploration of uncertainties

The Big Data Paradox in vaccine uptake

Article

Unrepresentative big surveys significantly overestimated US vaccine uptake

https://doi.org/10.1038/s41586-021-04198-4

Received: 18 June 2021

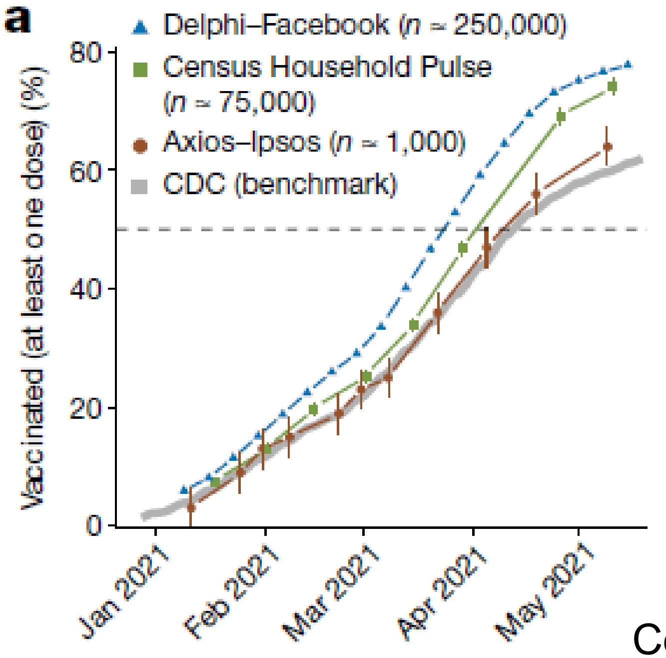
Accepted: 29 October 2021

Published online: 8 December 2021

Check for updates

Valerie C. Bradley[1,6], Shiro Kuriwaki[2,6], Michael Isakov[3], Dino Sejdinovic[1], Xiao-Li Meng[4] & Seth Flaxman[5]

Surveys are a crucial tool for understanding public opinion and behaviour, and their accuracy depends on maintaining statistical representativeness of their target populations by minimizing biases from all sources. Increasing data size shrinks confidence intervals but magnifies the effect of survey bias: an instance of the Big Data Paradox[1]. Here we demonstrate this paradox in estimates of first-dose COVID-19 vaccine uptake in US adults from 9 January to 19 May 2021 from two large surveys: Delphi–Facebook[2,3] (about 250,000 responses per week) and Census Household Pulse[4] (about 75,000 every two weeks). In May 2021, Delphi–Facebook overestimated uptake by 17 percentage points (14–20 percentage points with 5% benchmark imprecision) and Census Household Pulse by 14 (11–17 percentage points with 5% benchmark imprecision), compared to a retroactively updated benchmark the Centers for Disease Control and Prevention published on 26 May 2021. Moreover, their large sample sizes led to miniscule margins of error on the incorrect estimates. By contrast, an Axios–Ipsos online panel[5] with about 1,000 responses per week following survey research best practices[6] provided reliable estimates and uncertainty quantification. We decompose observed error using a recent analytic framework[1] to explain the inaccuracy in the three surveys. We then analyse the implications for vaccine hesitancy and willingness. We show how a survey of 250,000 respondents can produce an estimate of the population mean that is no more accurate than an estimate from a simple random sample of size 10. Our central message is that data quality matters more than data quantity, and that compensating the former with the latter is a mathematically provable losing proposition.



Nature v. 600 (2021) 695

Courtoy et al., PRD 107 (2023) 034008