Exploring sensitivity of hadronic experiments to the nucleon structure

Bo Ting Wang,[1,*] Sean Doyle,[2,3,†] Jun Gao,[2,3] Timothy Hobbs,[1,‡]
Tie-Jiun Hou,[4,§] Pavel Nadolsky,[3,¶] and Fredrick I. Olness[3,‡]

[1]*Department of Physics, Southern Methodist University, Dallas, TX 75275-0181, U.S.A.*
[2]*School of Physics and Astronomy, INPAC, Shanghai Key Laboratory for Particle Physics and Cosmology, Shanghai Jiao-Tong University, Shanghai 200240, China*[**]
[3] *Department of Physics, Southern Methodist University, Dallas, TX 75275-0181, U.S.A.*
[4]*School of Physics Science and Technology, Xinjiang University, Urumqi, Xinjiang 830046 China*

We demonstrate a collection of tools and metrics for quantitatively studying the sensitivity of hadronic measurements to the underlying Parton Distribution Functions (PDFs).

Replace PACS with PhySH:https://www.aps.org/publications/apsnews/201602/classification.cfm

PACS numbers: 12.15.Ji, 12.38 Cy, 13.85.Qk
Keywords: parton distribution functions; large hadron collider; Higgs boson

Contents

*Electronic address: botingw@smu.edu
†Electronic address: seand@smu.edu
‡Electronic address: olness@smu.edu
§Electronic address: tiejiun.hou@foxmail.com
¶Electronic address: nadolsky@smu.edu
**Electronic address: jung49@sjtu.edu.cn

I. INTRODUCTION

We'll polish this part near the end

A. Statement of the Problem

Parton distribution functions (PDFs) are crucial for understanding the behavior of hadron collisions and then exploring the Standard Model (SM). PDFs describe the structure of hadrons, which affect the configurations of the final particles in the collisions. Therefore, the magnitudes of physical observables in hadron collisions strongly depend on PDFs. Currently, The Large Hadron Collider (LHC) produces a lot of experimental data. Owning to the fact that uncertainties in measurements constantly decrease, reducing the PDF uncertainties of physical observables and using the higher order PDFs will make it easier to find the inconsistency between SM and the data sets collected by the LHC and then discover new physics. Incorporating more (LHC) new data sets in the global fits of PDFs is a naive way to generate better PDF sets with small uncertainties.

However, incorporating more experimental data points will substantially increase the time for fitting PDF sets, especially when we fit higher order PDF sets. From here we know that how to select data sets in global fits will become extremely important in the near future. It is essential to know which data sets will effective constrain the higher order PDFs for the global fits in the limited time of computation. In addition, because physical predictions are sensitive to respective flavors and regions of $\{\xi, \mu\}$ in PDFs, we need to narrow down uncertainties of the specific regions of $\{\xi, \mu\}$ (in the PDFs). Where partonic ξ are momentum fractions and μ are QCD factorization scales. For example, if PDF values for the leading $\{\xi, \mu\}$ ranges and flavors that characterize kinematical quantities for Higgs production processes (e.g. at $\mu = 125\,GeV$) are tightly constrained, the theoretical predictions for these processes are reliable (precise).

1. methods

   Using correlation between PDF uncertainties in two observables have been proposed to study constraints on PDFs and constraints on observables imposed by PDFs [1][2][3].

2. the evaluation of the methods

   The approach can help us to find the $\{\xi, \mu\}$ ranges of PDFs affecting physical observables such as total cross section [3]. It is yet be established that how to know the ranges specifying PDFs constrained by experimental data sets.

B. Objectives

Thus, establishing a better understanding of the relationships between the strength of constraints on PDF and experimental data sets will be a significant and beneficial contribution to particle physics.

C. Methodology

I have developed and tested a systematic method to study the constraints on PDFs imposed by the experimental data sets. I will use established statistical observables to quantify the strength of these constraints.

After that, I will introduce a statistical technique to visualize the regions of partonic momentum fractions ξ and QCD factorization scales μ where the experiments impose strong constraints on the PDFs. Recent experimental data will be considered in the analysis in order to provide better constraints to various ranges of PDFs.

1. Scope, Limitations and Assumptions

2. Significance

3. Structure of This Paper

D. Overview of paper

The article proceeds as follows. First, in the II A, we give a brief overview of PDF fitting. Second, the method used to "see" PDF from experimental data is introduced in II B. Third, an idea to estimate constraint of PDFs that we have seen from this method is provided in II C. Fourth, advantages of using correlation, residual uncertainty, and sensitivity to study PDF constraint is discussed in III A. Two methods are discussed in III A and **??**. The corresponding Mathematica code is introduced in **??**.

II. OVERVIEW OF GLOBAL FITTING AND CONSTRAINTS ON PDFS

...

A. Introduction of PDF fitting

How much do they need to know? Whether it is Hessian or MC, we always have a central set, and then a distribution of error sets.

Specifically, for the Hessian method:

1. We parameterize $f_i(x, Q_0^2) = a_0 x^{a_1}(1-x)^{a_2} F(a_3, a_4, ...)$ at Q0
2. determine the best-fitted a_0, a_1, a_2..... of the parametrization functions by minimizing χ^2
3. determine uncertainties of the parametrization functions by requiring $\chi^2 < \chi^2_{min} + \chi^2_{torelance}$
4. convert from $\{a_0, a_1, a_2, ...\}$ basis to an eigenvector basis $\{r_i\}$

~~To fit PDF sets, we first need to determine the input theoretical model and data sets. The model includes the selection of quark mass, coupling constants and the order considered in the correction of perturbation (i.e. LO, NLO and etc). Then we determine the parametrization function form $f_q(\xi, Q_0) = a_0 \xi^{a_1}(1-\xi)^{a_2} F(a_3, a_4, ...)$ at the lowest factorization scale Q_0, for which we need to take some physical rules into account. For instance, $a_0\xi^{a_1}(1-\xi)^{a_2}$ term requires that the probabilities of the partons with momentum fraction $\xi = 1$ or $\xi = 0$ are 0. Besides, the momentum conservation $(\int_0^1 \xi \sum_i f_i(\xi, Q) dx = 1)$ requires that the total momentum of all subparticles in each hadron should equal to the kinematical momentum of that hadron. $(\int_0^1 (u(x) - \bar{u}(x)) dx = 2$ and $\int_0^1 (d(x) - \bar{d}(x)) dx = 1)$ require that protons consist of uud at the low factorization scale. we use χ^2 minimization to explore the best fit parameters a_0, a_1, a_2 and etc. χ^2 analysis applies the ratios of the deviations between theoretical predictions and experimental values to experimental error bars to quantify the goodness of fits. Here are steps of the PDF fitting:~~

1. ~~select the experimental data sets and theoretical model in global fits~~
2. ~~write down parametrization functions of all flavors~~
3. ~~determine the best-fitted a_0, a_1, a_2..... of the parametrization functions by minimizing χ^2~~
4. ~~determine uncertainties of the parametrization functions by requiring $\chi^2 < \chi^2_{min} + \chi^2_{torelance}$~~

B. PDF fitting with χ^2 Method

This is just defining the residuals. Is that all we need???

χ^2 test is a way to evaluate the goodness of fits. We assume good fits are theoretical predictions within experimental error bar. Thus, by residuals (r_i) of data points i, We can estimate the goodness of the fit of this point. For data sets, the goodness of fits is the sum of all points i and r_i, which means that we use the squared distance to evaluate the agreement between the model and the data sets. Hence, we can obtain the best coefficients of the

model by minimizing χ^2. If $\chi^2 \gg N_{data}$ for a data set and a theory, this fit is bad. If $\chi^2 \leq N_{data}$, this fit is better than expected. If $\chi^2 \ll N_{data}$, this fit is over-fitted. To sum up, r_i and χ^2 could be defined as follows:

$$r_i = \frac{T_i - D_i}{\sigma_i}$$

$$\chi^2 = \sum_{i \in \text{Expt data}} r_i^2$$

Where i r_i, T_i, D_i, and σ_i are the residual, theoretical prediction, experimental central value, and experimental uncertainty in data point i. When we take systematic uncertainties into account, the experimental central values should be modified to $D_{shift,i} = D_i + shift_i$ since the averages of the measurements are shifted from the real values. Here we provide criteria of good PDF fittings

1. $\chi^2 \simeq N_{data}$, the smaller a χ^2, the better a fitting
2. if residuals of points i are small ($|r| < 1$, r is called residual $\frac{T_i - D_{shift,i}}{\sigma_i}$), the fit of these points is good

C. How to estimate PDF constraints

Here we show samples of the types of plot output the program can make.

In general, we estimate the uncertainties of the parameters describing PDFs by constraining χ^2 value. In other words, we identify the region representing to the parameter uncertainties by the region with χ^2 smaller than $\chi^2_{min} + \chi^2_{torelance}$. We learn that constraints on PDFs are from constraining the upper bound of the goodness of fits. When we fit theoretical models to match experimental data so that r_i for data points i and χ^2 are not too large, PDFs are constrained. Thus, we use χ^2 and r_i to see the constraints on PDFs because they represent the criteria of the goodness of fits and $f_a(\xi, \mu)$ values are constrained to meet this criteria. In other words, the criteria could determine the range of $f_a(\xi, \mu)$. For instance, Fig. 1 is the comparison of two different fluctuations of theoretical values. Even though mean values of red crosses and blue crosses are the same, we can find the fluctuation of red crosses is easier to be detected because it's affection on residual values is larger than the fluctuation of blue crosses. Fig. 2 is the comparison of two different fluctuations of residuals depending on $f_a(\xi, \mu)$. Although both of red circles and blue circles are strongly correlated, red circles are more sensitive to $f_a(\xi, \mu)$ because the $f_a(\xi, \mu)$ fluctuation of red circles strongly affects values of residuals. Therefore, To understand the relationship between data sets and the constraints on PDFs imposed by these data sets, we should study whether χ^2 and r_i are sensitive to the variation of $f_a(\xi, \mu)$ values.

FIG. 1: Theoretical predictions and an experimental data point measurement with the error bar. Red crosses and blue crosses are two sets of theoretical prediction uncertainties.

III. SYSTEMATIC METHOD FOR STUDYING CONSTRAINTS ON PDFS IMPOSED BY EXPERIMENTAL DATA SETS

framework

I have developed and tested a systematic method to study the constraints on PDFs imposed by experimen-

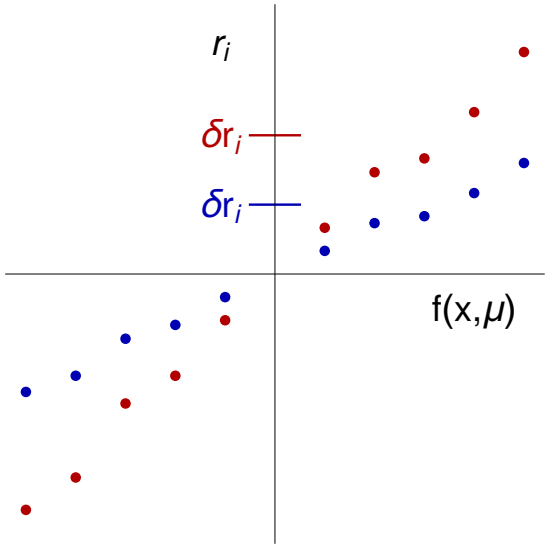

FIG. 2: The sensitivity of a data point to a PDF. Red circles and blue circles are residuals of two data points versus f(x,Q) values in PDF error sets

tal data sets. I use established statistical observables to quantify the strength of these constraints. After that, I introduce a statistical technique to visualize the regions of partonic momentum fractions ξ and QCD factorization scales μ where the experiments impose strong constraints on the PDFs. Recent experimental data is considered in the analysis in order to provide better constraints to various ranges of PDFs.

To test the effectiveness of the proposed method, I study constraints on CT14NNLO parton distributions [4] from various data sets. I include various types of experimental data sets in the analysis, including DIS processes, $Z \to l^+l^-, d\sigma/dy(l)$, $W \to l\nu$, and jet productions $(p_1p_2 \to jjX)$.

visualization method

For data sets of interest, we can demonstrate and identify values of correlation/sensitivity data by different colors on the $\xi - \mu$ plane ($2D - \xi - \mu$ figure), such as Figs. 2, which help us to rapidly estimate the distribution of the strength of constraints on the $\xi - \mu$ plane. We can also know the number of data points constraining PDFs by the histograms of the statistical quantities.

A. Statistical quantities for constraints

introduction correlation and sensitivity

Among various quantities that characterize the sensitivity of the experimental data to the PDFs, the correlations $Corr(f_a(\xi,\mu), r_i)$ of PDFs $f_a(\xi,\mu)$ and residuals r_i can determine whether there exist predictive relationships between PDFs and goodness of fit to data points. Here a is the flavor index, and r_i is the residual of data point i. We can also define a factor $\delta r_i \times Corr(f_a(\xi,\mu), r_i)$ to quantify the sensitivity of the experimental datum to a variation of the PDF. Both correlation and sensitivity are useful for constraining PDFs.

correlation

correlation's advantages

Correlation illustrates the strength of the predictive relation between any two observables X and Y. We can use values of one observable to predict values of another observable very well when their correlation is close to ± 1. correlations of Hessian uncertainties [1] have been used to see the simultaneous constraint on observables X and Y, and to get constraints on PDFs [1][2][3]. First, via measuring one physical observable, we are able to predict the value of another observable precisely. In addition, strong correlations are highly likely to show the signs of some physical relations, such as causation, between the two observables.

Hessian correlation: definition and physical meaning

There are several ways to evaluate uncertainties on PDFs such as the Hessian method [1], the Monte Carlo method [5][6], and the Lagrange Multiplier [7]. Our PDF set input is CT14NNLO, which uses the Hessian method to estimate uncertainties information. This idea is based on the quadratic assumption. According to the quadratic assumption, we will get an elliptical shape of PDF parameter space around the best fit parameters $\vec{a_0}$ for a given tolerance parameter $\chi^2_{tolerance}$ satisfying $\chi^2(\vec{a}) < \chi^2(\vec{a_0}) + \chi^2_{torelance}$. If errors of an observable X along the $\pm$ directions of i-th dimension of the ellipse are

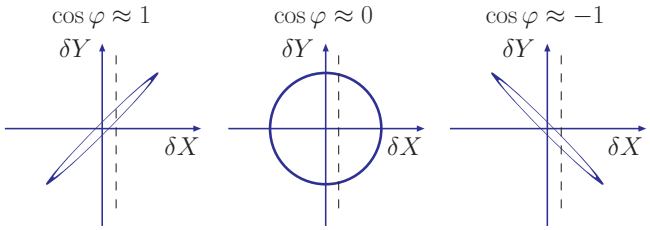

FIG. 3: Lissajous figure of observables X and Y for different $cos\phi$

X_i^+ and X_i^-, the uncertainty of X based on the variation of parameter at i-th dimension could be approximated by $(X_i^+ - X_i^-)/2$. According to the principle of error propagations, the X uncertainty via PDF parameter space is $\Delta X = \frac{1}{2}\sqrt{\sum_i (X_i^+ - X_i^-)^2}$.

Move these definitions to an appendix???

Our idea of studying PDF constraints from data sets uses the correlation between PDF Hessian sets and residual Hessian sets, where the Hessian correlation of two observables is defined as $cos\phi = \sum_i (X_i^+ - X_i^-)(Y_i^+ - Y_i^-)/4\Delta X \Delta Y$. The correlation of any two observables X and Y could be used to see the simultaneous constraint of X and Y [1]. The ellipse of simultaneous constraint could be described by Lissajous figure

$$X = X_0 + \Delta X\, sin(\theta + \phi)$$

$$Y = Y_0 + \Delta Y\, sin(\theta + \phi)$$

where $0 < \theta < 2\pi$ traces the shape of the ellipse, and whether the shape is needle-like or circle-like is controlled by ϕ. If $|cos\phi| \simeq 1$, the shape is needle-like, which strongly constrains Y for a given δX (see Fig. 3). Thus, the correlations $Corr(f_a(\xi,\mu), r_i)$ of PDFs $f_a(\xi,\mu)$ and the residuals r_i can determine the strength of constraints on PDFs imposed by r_i in experimental data points.

construct more representative statistical quantity for constraints:

strength of constraints on PDFs imposed by r_i $(SOC(PDF))$

Although we can know the predictivities between PDFs and measurements through $Corr(f_a(\xi,\mu), r_i)$, $Corr(f_a(\xi,\mu), r_i)$ could not specify the strength of constraints on PDFs imposed by r_i $(SOC(PDF))$. For instance, the measurements with large uncertainties cannot effectively constrain $f_a(\xi,\mu)$ no matter how large $Corr(f_a(\xi,\mu), r_i)$ is, since r_i is not sensitive to the variation of $f_a(\xi,\mu)$. Therefore, we want to find a more representative statistical quantity for $SOC(PDF)$. To study $SOC(PDF)$ between PDFs and data sets, we study the variation of χ^2 and r_i associated with the fluctuation in $f_a(\xi,\mu)$. Fig. 2 shows r_i in data points depending on the variation in $f_a(\xi,\mu)$ error sets. We find that despite the fact that the correlation between r_i in two data points (red circles and blue circles) and $f_a(\xi,\mu)$ are the same, the fluctuation for $f_a(\xi,\mu)$ imposes different levels of impact to r_i. The r_i, represented by red circles, are more sensitive to $f_a(\xi,\mu)$, which indicates that when we constrain χ^2 for getting the new fitted $f_a(\xi,\mu)$ error sets, the data point represented by the red circles will more dramatically narrow down the range of the new $f_a(\xi,\mu)$ error sets so that r_i for error sets become smaller. Here we find the δr_i, which evaluates the fraction of theoretical and experimental uncertainties, indicating whether the theoretical uncertainties are apt to be constrained after the fitting. For the above reasons, we advise using $\delta r_i \times Corr(f_a(\xi,\mu), r_i)$ to quantify the sensitivity $(Sen(f_a(\xi,\mu), r_i))$ for r_i to $f_a(\xi,\mu)$, and using the sensitivity to estimate $SOC(PDF)$ for data point i.

Same point method

objective

In principle, we can use the correlation & sensitivity mentioned above to quantify $SOC(PDF)$ for any points on the $\xi - \mu$ plane and data point i. Therefore, we can identify which regions in the $\xi - \mu$ plane have strong $SOC(PDF)$. Our objective is to characterize the strongly constrained ranges (Strong $SOC(PDF)$

Regions) imposed by the given data sets.

difficulties in the analysis and solutions

Acquiring the corresponding Strong $SOC(PDF)$ Regions for each point i still could not tell us which ranges are constrained by each data set because the amount of information in all data points is too large to analyze it. Therefore, we present a simple method as follows. For each data point i, we select the points (or the ranges) in the $\xi - \mu$ plane whose $f_a(\xi_i, \mu_i)$ are constrained most by the point i (Max $SOC(PDF)$ Regions). Here we assume that in scattering processes, sizes of physical observables are mainly contributed by the ranges near $\{\xi_i, \mu_i\}$. Thus, it is highly possible that the measurement at point i will impose the strongest constraints on the $f_a(\xi_i, \mu_i)$ in these ranges. As a result, the combination of these ranges describes the most constrained ranges for all data points in each data set.

capability

It is possible to evaluate those PDF ranges. For each experimental data point i, we can establish an approximate relation between the kinematical quantities for that data point, and unobserved quantities a, ξ, and μ specifying the PDFs, where a, ξ, and μ are flavor, momentum fraction, and resolution scale of partons. For example, in DIS, ξ and μ are approximately equal to Bjorken x and momentum transfer Q according to the Born-level kinematic relation. However, this relation is violated in high-order radiative contributions. Nevertheless, this relation will approximately hold in most scattering events. Therefore, we derive the relation between the kinematical quantities and unobserved quantities we mentioned for data sets in our analysis, including DIS, $dXsec/dy(l)$ of $Z \to l_+l_-$, $dXsec/dy(l)$ of $W \to l\nu$, and (di)jet productions. In practice, our fitted PDF sets are not perfect, so even some ranges of the real PDF dominate a physical observable, the PDF sets in these ranges are not always strongly correlated to that physical observable.

Following are formulas (code part: selectExptxQv2 in **??**) connecting experimental data points and their leading ξ, μ points of PDFs:

Process	Experimental	Theortical	notes	ref???
DIS	$\{x, Q^2\}$	$\{\xi, \mu\}$		
Z Prog	$\{\tau, y\}$			
Jet Pro	$\{y, p_T\}$			

TABLE I: table

DIS: $x, Q_{data} = \xi, \mu_{PDF}$

$Z \to l_+l_-, dXsec/dy(l)$: $(Q/\sqrt{S}) \times \exp(\pm y), Q_{data} = \xi, \mu_{PDF}$

$W \to l\nu$: same as $Z \to l_+l_-, dXsec/dy(l)$

JP ($q_1q_2 \to j_1j_2$, estimate ξ_1, ξ_2 of jet as peak of $y(j_1), y(j_2)$): $(2P_T/\sqrt{S}) \times \exp(\pm y), P_{T\,data} = \xi, \mu_{PDF}$

Here we give physics of these formulas. DIS processes are just mentioned in the example. In lepton pair production and (di)jet production, the rapidities of the final-state pairs are small for most events. If y is integrated out, we set $y = 0$ $\tau = Q/\sqrt{S}$ $\xi_1 = \xi_2 = \tau$. If y of the lepton pair or jet pair is known, we set $\xi_{1,2} = \tau \cdot \exp(\pm y)$. For jet production, $\tau = 2p_T^{\rm jet}/\sqrt{S}$ at the leading order. In most events, if rapidity y_l of the lepton is known yet y of the boson is unknown, we use the fact that $y_l \sim y \pm 1$ for most events. You can still estimate ξ_1 and ξ_2 as $\xi_{1,2} = \tau \cdot \exp(\pm y)$, where $y \sim y_l$ (up to an error of less than 1 unit).

advantages

Because Max $SOC(PDF)$ Regions are obtained from a physical relation, we know the constraints in these ranges are the real physical constraints rather than the results of other factors, such as parametrization function dependency.

practical procedure (step by step)

Finally, we provide steps of Same point method. Steps are as follows:

1. calculate Max $SOC(PDF)$ Regions $\{\xi_i, \mu_i\}$ from corresponding experimental data points i by using a suitable transformation formula describing the approximate relation between $\{\xi_i, \mu_i\}$ and kinemat-

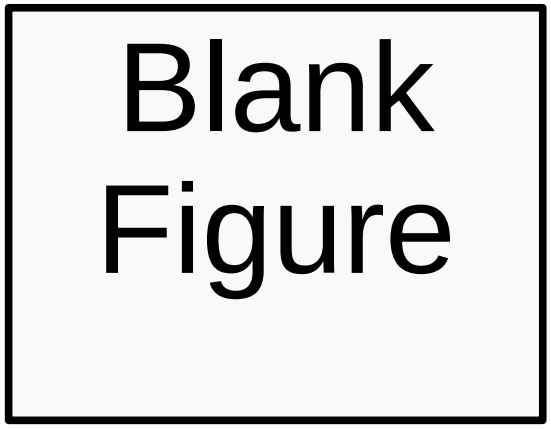

FIG. 4: The correlation and sensitivity of the Hessian uncertainty of data set 281(D0 Run-2) and CT14NNLO depending on the momentum fraction and the factorization scale.

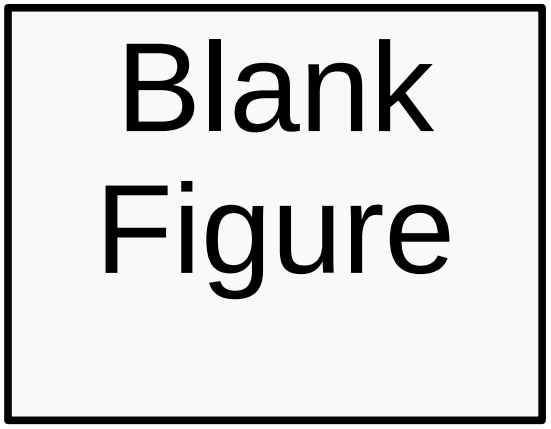

FIG. 5: The correlation and sensitivity of the Hessian uncertainty of data set 225(cdfLasy) and CT14NNLO depending on the momentum fraction and the factorization scale.

ical quantities for points i (code part: selectExptxQv2 in **??**)

2. for each data point, calculate all flavours of PDF values for the same $\{\xi_i, \mu_i\}$ (executable part: fxQsamept.nb in **??**)

3. calculate correlation $Corr(f_a(\xi_i, \mu_i), r_i)$, sensitivity $\delta r_i * Corr(f_a(\xi, \mu), r_i)$, and other statistical quantities ($r_{i,central\,value}$, δr_i, and experimental error ratio in this code) for every point i and every flavor a (executable: fxQsamept_corr.nb in **??**)

4. draw the histograms and $2D - \xi - \mu$ figures for the statistical quantities derived in step 3 (executable: run_v3.nb in **??**, example figure: 4)

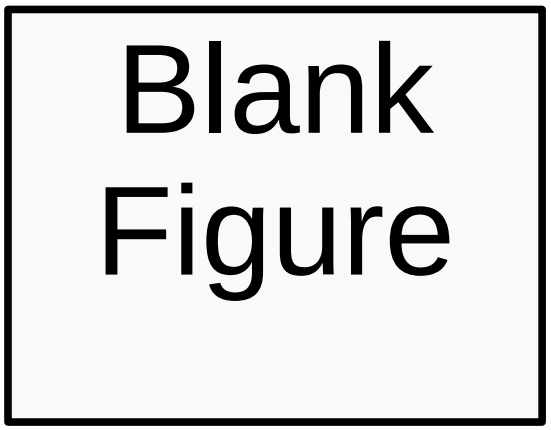

FIG. 6: The correlation and sensitivity of the Hessian uncertainty of data set 247(ATL7Zpt) and CT14NNLO depending on the momentum fraction and the factorization scale.

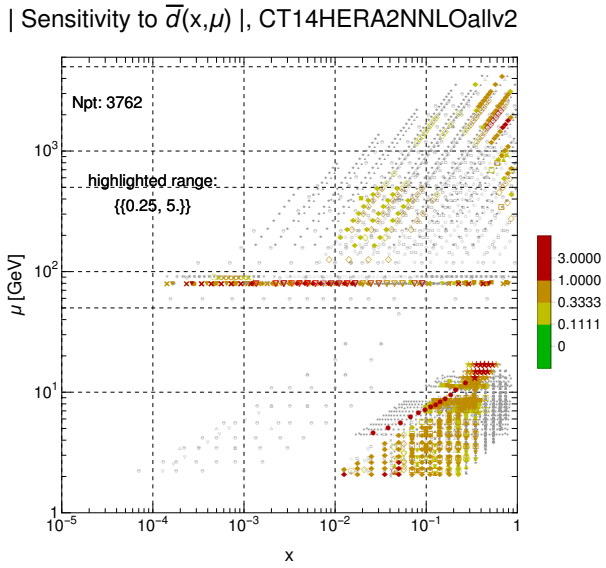

FIG. 7: Sample Figure.

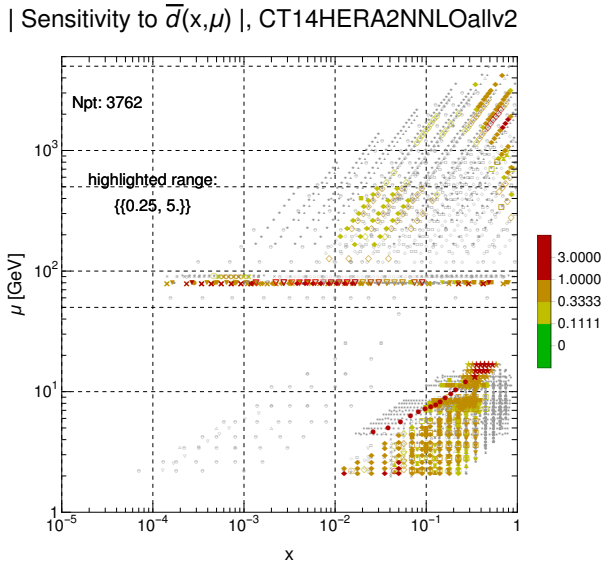

FIG. 8: Sample Figure.

IV. THE METRICS:

A. The metrics

We briefly introduce the various metrics which we can use to evaluate the relative sensitivity of separate experimental measurements to the individual PDF flavors.

1. Correlations

In Figure 9 we display the computed correlation between the top quark production process and the gluon PDFs. The correlations are defined in Eq.**** and the values range over the interval $[-1, 1]$.

In Figure 9-a) we show the histogram of the absolute values of the correlations for each of the 35 data points in the top quark production set. Larger magnitudes imply a stronger correlation, and we have arbitrarily drawn a line at 0.7 to focus on the larger values. In Figure 9-a) there are 11 out of 35 data points with an absolute correlation above the 0.7 mark.

In Figure 9-b) we display the absolute value of the correlation in the $\{x, \mu\}$ plane. This allows us to see the specific kinematic region where the data contributes. The points are color-coded according to the legend in the figure, and the points above the 0.7 threshold are enlarge to emphasize their contributions.

For $\mu \sim 350\,\text{GeV}$ there is a grouping of data points corresponding to pair production of top–anti-top quarks. There is also a group of points *** [describe the points on the slope] ***

2. Sensitivity

Although the correlation is a useful measure, this has a potential drawback that if the uncertainties of the measurement are large then the net constraining power of the data could be minimal. To address this deficiency, we can construct a new quantity which we will label as the sensitivity and define as the residual times the correlation. Thus, data sets with both a strong correlation and a strong pull will have a large impact on the fit, and hence a large sensitivity.

In Figure 10 we display the computed sensitivity between the top quark production process and the gluon PDFs. In contrast to the correlations which are constrained to the interval $[-1, 1]$, the sensitivity can in principle take on any value: $[-\infty, \infty]$; in practice, we expect the sensitivity to be centered about zero with a roughly Gaussian distribution.

In Figure 10-a) we show the histogram of the absolute values of the sensitivity for each of the 35 data points in the top quark production set. Larger magnitudes imply a stronger sensitivity, and we have arbitrarily drawn a line at 1.0 to focus on the larger values. In Figure 10-a) there are only 1 out of 35 data points with an absolute correlation above the 1.0 mark.

In Figure 10-b) we display the absolute value of the sensitivity in the $\{x, \mu\}$ plane. This allows us to see the specific kinematic region where the data contributes. The points are color-coded according to the legend in the figure, and the points above the 0.25 threshold are enlarge to emphasize their contributions.

Again we see the distribution of the data points are grouped along a line at $\mu \sim 350\,\text{GeV}$ corresponding to pair production of top–anti-top quarks. There is also a group of points *** [describe the points on the slope] ***

V. SELECTED CASE STUDIES:

Using the correlations and the sensitivity described in the previous chapter, we can examine individual data sets to answer detailed questions about the influence on the PDF flavors.

Here we will present two example analyses demonstrating the utility of these metrics. First we will study the influence of both the top quark and jet production data on the gluon PDF. Next we will look at a future LHeC

facility and investigate how new data might further constrain the PDFs.

A. Constraining the Gluon

Before high precision top production data sets were available, the primary constraints on the gluon PDF came from the inclusive jet cross section measurements. With the recent LHC runs, we additionally have high-precision measurements of top quark production available. Thus, we can now study the extent to which each data set constrains the gluon PDF. We will find the correlations and sensitivities introduced in the previous section are helpful in addressing this queston. We begin by comparing the correlation and sensitivity relating the gluon PDF with the top quark production (discussed in the previous section) to the same quantities relating the gluon PDF to the jet cross section. Figure 11 displays the correlations, and Figure 12 displays the sensitivity.

Examining the correlations between the jet cross sections and the gluon PDFs, in Fig. 11-a) we show a histogram of the correlations for all 538 jet data points. As before, we indicate a arbitrary threshold at a correlation of 0.7; there are XXX data points in this region. Recall, for the top production cross section we had a total of 35 points with 11 above the 0.7 cut.

In Fig. 11-b) we display the absolute value of the correlation in the $\{x, \mu\}$ plane. As before, the points are color-coded according to the legend in the figure, and the points above the 0.7 threshold are enlarge to emphasize their contributions. Clearly, comparing Fig. 11-b) to Fig. 9-b) we see the jet data has many more points with significant correlation covering a much larger region of the $\{x, \mu\}$ plane.

Turning to the sensitivity measurements, we see in Fig. 12-a) a histogram of the sensitivities for all 538 jet data points, and we have marked an arbitrary threshold at 1.0; there are XXX data points in this region. By comparison, for the top production cross section we had a total of 35 points with 1 above the 1.0 cut.

In Fig. 12-b) we display the absolute value of the sensitivty in the $\{x, \mu\}$ plane; as in Fig. 10-b) the points above the 0.25 threshold are enlarged to emphasize their contributions. As with the correlations, comparing the top production to the jet cross section data we see the jet data has many more points with significant correlation covering a much larger region of the $\{x, \mu\}$ plane.

In this example, the combination of the correlation and sensitivity plots provides us a distinct picture of which data impacts the gluon PDF, and what the relevant $\{x, \mu\}$ kinematic regions are involved. This provides us with a set of incisive tools to answer detailed questions about the PDFs.

x

x

x

B. LHeC Predictions

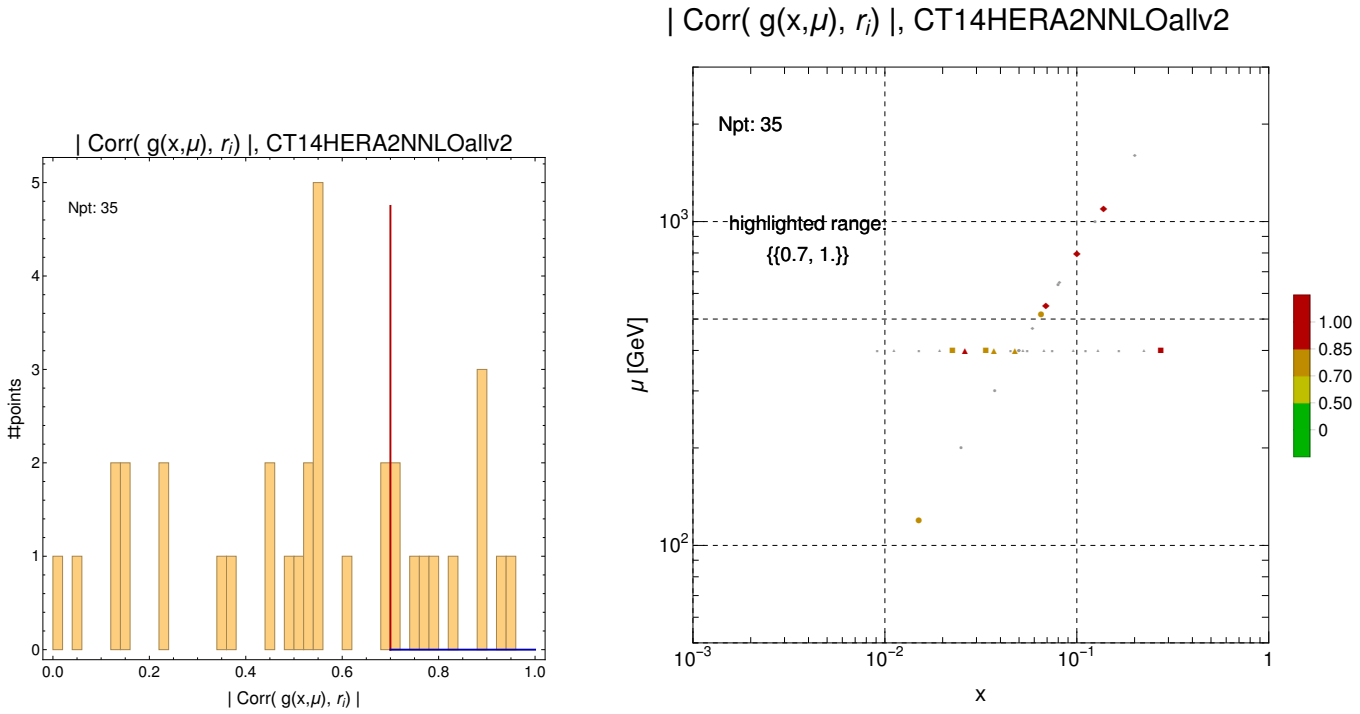

FIG. 9: The xQ-plot and histogram of sensitivities between top quark processes and gluon.

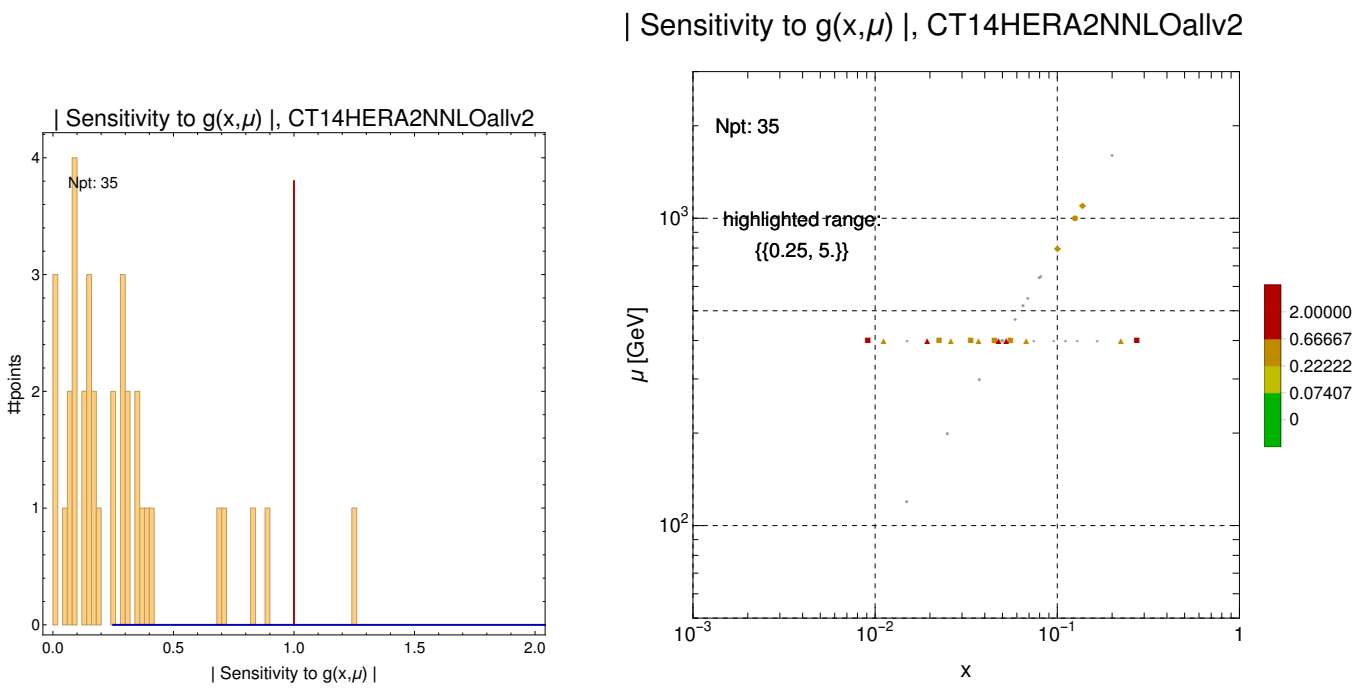

FIG. 10: The xQ-plot and histogram of sensitivities between top quark processes and gluon.

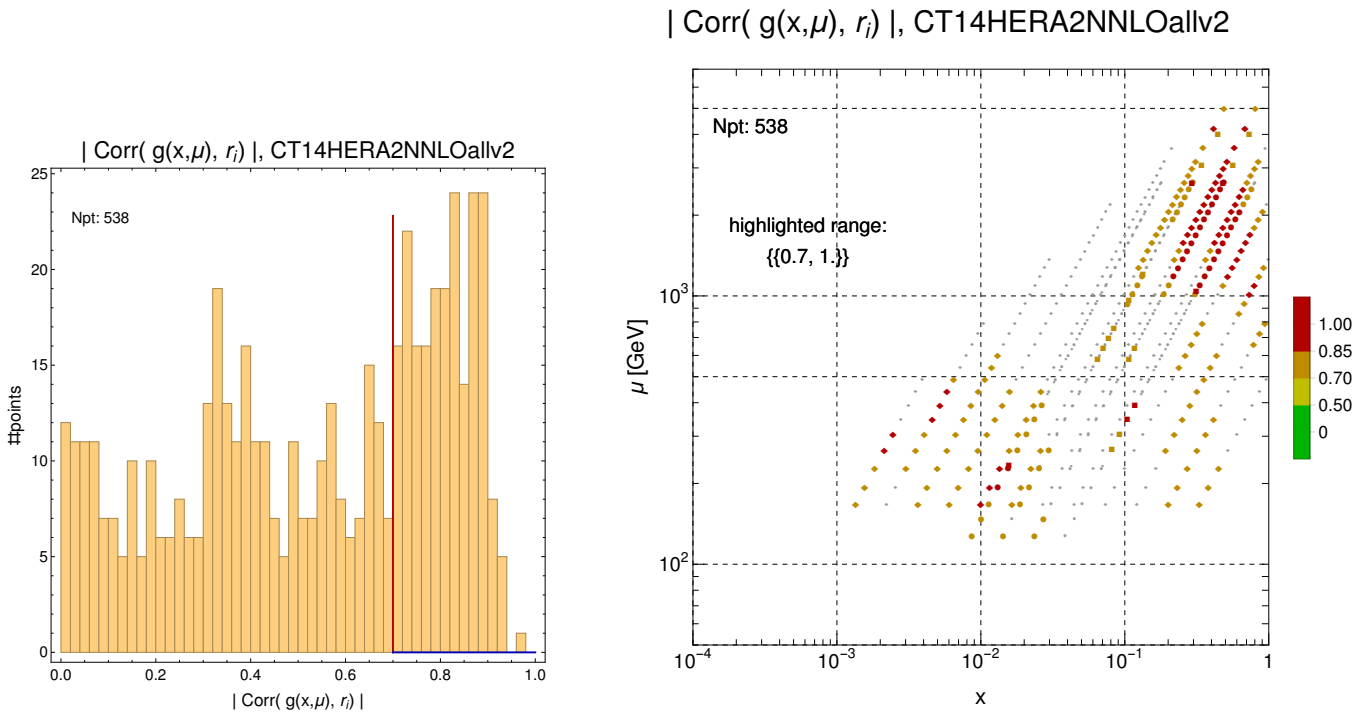

FIG. 11: The xQ-plot and histogram of correlations between jet processes and gluon.

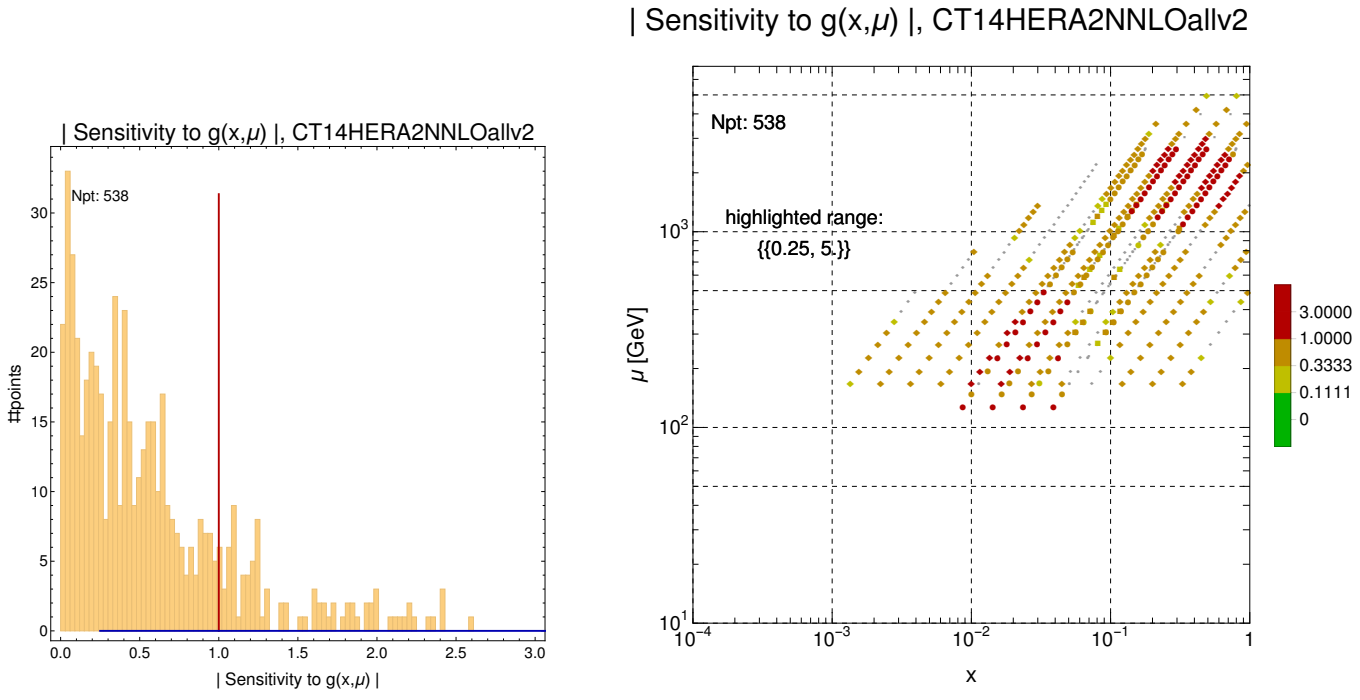

FIG. 12: The xQ-plot and histogram of sensitivities between jet processes and gluon.

VI. CONCLUSIONS

This work was supported in part by the U.S. Department of Energy under Grant No. DE-SC0010129 and by the National Natural Science Foundation of China under the Grant No. 11465018. The work of J.G. is sponsored by Shanghai Pujiang Program.

[1] J. Pumplin, D. Stump, R. Brock, D. Casey, J. Huston, J. Kalk, H. L. Lai, and W. K. Tung, Phys. Rev. **D65**, 014013 (2001), hep-ph/0101032.

[2] P. M. Nadolsky and Z. Sullivan, eConf **C010630**, P510 (2001), hep-ph/0110378.

[3] P. M. Nadolsky, H.-L. Lai, Q.-H. Cao, J. Huston, J. Pumplin, D. Stump, W.-K. Tung, and C.-P. Yuan, Phys. Rev. **D78**, 013004 (2008), 0802.0007.

[4] S. Dulat, T.-J. Hou, J. Gao, M. Guzzi, J. Huston, P. Nadolsky, J. Pumplin, C. Schmidt, D. Stump, and C.-P. Yuan, Phys. Rev. **D93**, 033006 (2016), 1506.07443.

[5] W. T. Giele and S. Keller, Phys. Rev. **D58**, 094023 (1998), hep-ph/9803393.

[6] W. T. Giele, S. A. Keller, and D. A. Kosower (2001), hep-ph/0104052.

[7] D. Stump, J. Pumplin, R. Brock, D. Casey, J. Huston, J. Kalk, H. L. Lai, and W. K. Tung, Phys. Rev. **D65**, 014012 (2001), hep-ph/0101051.

[8] W.-K. Tung, H.-L. Lai, A. Belyaev, J. Pumplin, D. Stump, and C.-P. Yuan, JHEP **02**, 053 (2007), hep-ph/0611254.

This is stolen from Pavel's paper. We'll take what we need and refer to his paper for the rest.

APPENDIX A: CORRELATIONS DUE TO PDFS

In many applications, it is instructive to study the correlations between the PDFs and the experimental observables. We review the relevant theoretical framework as presented in Ref.***.

Let X be a variable that depends on the PDFs. We consider X as a function of the parameters $\{a_i\}$ that define the PDFs at the initial scale μ_0. Thus we have $X(\vec{a})$, where $\vec{a}$ forms a vector in an N-dimensional PDF parameter space, with N being the number of free parameters in the global analysis that determines these PDFs. In the Hessian formalism for the uncertainty analysis developed in [1] and used in all of our recent work, this parton parameter space is spanned by a set of orthonormal eigenvectors obtained by a self-consistent iterative procedure [8**?**].

If $\vec{a}_0$ represents the best fit obtained with a given set of theoretical and experimental inputs, the variation of $X(\vec{a})$ for parton parameters $\vec{a}$ in the neighborhood of $\vec{a}_0$ is given, within the Hessian approximation, by a linear formula

$$\Delta X(\vec{a}) = X(\vec{a}) - X(\vec{a}_0) = \vec{\nabla} X|_{\vec{a}_0} \cdot \Delta\vec{a}, \tag{A1}$$

where $\vec{\nabla} X$ is the gradient of $X(\vec{a})$, and $\Delta\vec{a} = \vec{a} - \vec{a}_0$. As explained in detail in Refs. [1, 8**?**], the uncertainty range of the PDFs in our global analysis is characterized by a tolerance factor T, equal to the radius of a hypersphere spanned by maximal allowed displacements $\Delta\vec{a}$ in the orthonormal PDF parameter representation. T is determined by the criterion that all PDFs within this tolerance hypersphere should be consistent with the input experimental data sets within roughly 90% c.l. The detailed discussions and the specific iterative procedure used to construct the eigenvectors can be found in Refs. [1, 8**?**].

In practice, the results of our uncertainty analysis are characterized by $2N$ sets of published eigenvector PDF sets along with the central fit. We have 2 PDF sets for each of the N eigenvectors, along the $(\pm)$ directions respectively, at the distance $|\Delta\vec{a}| = T$. The i-th component of the gradient vector $\vec{\nabla} X$ may be approximated by

$$\frac{\partial X}{\partial a_i} \equiv \partial_i X = \frac{1}{2}(X_i^{(+)} - X_i^{(-)}), \tag{A2}$$

where $X_i^{(+)}$ and $X_i^{(-)}$ are the values of X computed from the two sets of PDFs along the $(\pm)$ direction of the i-th eigenvector. The uncertainty of the quantity X due to its dependence on the PDFs is then defined as

$$\Delta X = \left|\vec{\nabla} X\right| = \frac{1}{2}\sqrt{\sum_{i=1}^{N}\left(X_i^{(+)} - X_i^{(-)}\right)^2}, \tag{A3}$$

where for simplicity we assume that the positive and negative errors on X are the same.*

We may extend the uncertainty analysis to define a *correlation* between the uncertainties of two variables, say $X(\vec{a})$ and $Y(\vec{a})$. We consider the projection of the tolerance hypersphere onto a circle of radius 1 in the plane of the gradients $\vec{\nabla} X$ and $\vec{\nabla} Y$ in the parton parameter space [1, 2]. The circle maps onto an ellipse in the XY plane. This "tolerance ellipse" is described by Lissajous-style parametric equations,

$$X = X_0 + \Delta X \cos\theta, \tag{A4}$$

$$Y = Y_0 + \Delta Y \cos(\theta + \varphi), \tag{A5}$$

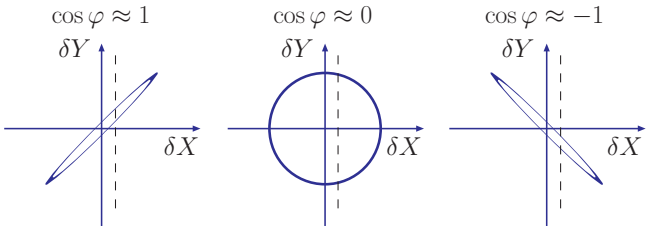

FIG. 13: Dependence on the correlation ellipse formed in the $\delta X - \delta Y$ plane on the value of $\cos\varphi$.

* A more detailed equation for ΔX accounts for differences between the positive and negative errors [2**?**]. It is used for $t\bar{t}$ cross sections in Table **??** and Fig. **??**.

where the parameter θ varies between 0 and 2π, $X_0 \equiv X(\vec{a}_0)$, and $Y_0 \equiv Y(\vec{a}_0)$. ΔX and ΔY are the maximal variations $\delta X \equiv X - X_0$ and $\delta Y \equiv Y - Y_0$ evaluated according to Eq. (A3), and φ is the angle between $\vec{\nabla} X$ and $\vec{\nabla} Y$ in the $\{a_i\}$ space, with

$$\cos\varphi = \frac{\vec{\nabla} X \cdot \vec{\nabla} Y}{\Delta X \Delta Y} = \frac{1}{4\Delta X\, \Delta Y} \sum_{i=1}^{N} \left(X_i^{(+)} - X_i^{(-)}\right)\left(Y_i^{(+)} - Y_i^{(-)}\right). \tag{A6}$$

The quantity $\cos\varphi$ characterizes whether the PDF degrees of freedom of X and Y are correlated ($\cos\varphi \approx 1$), anti-correlated ($\cos\varphi \approx -1$), or uncorrelated ($\cos\varphi \approx 0$). If units for X and Y are rescaled so that $\Delta X = \Delta Y$ (e.g., $\Delta X = \Delta Y = 1$), the semimajor axis of the tolerance ellipse is directed at an angle $\pi/4$ (or $3\pi/4$) with respect to the ΔX axis for $\cos\varphi > 0$ (or $\cos\varphi < 0$). In these units, the ellipse reduces to a line for $\cos\varphi = \pm 1$ and becomes a circle for $\cos\varphi = 0$, as illustrated by Fig. 13. These properties can be found by diagonalizing the equation for the correlation ellipse,

$$\left(\frac{\delta X}{\Delta X}\right)^2 + \left(\frac{\delta Y}{\Delta Y}\right)^2 - 2\left(\frac{\delta X}{\Delta X}\right)\left(\frac{\delta Y}{\Delta Y}\right)\cos\varphi = \sin^2\varphi. \tag{A7}$$

A magnitude of $|\cos\varphi|$ close to unity suggests that a precise measurement of X (constraining δX to be along the dashed line in Fig. 13) is likely to constrain tangibly the uncertainty δY in Y, as the value of Y shall lie within the needle-shaped error ellipse. Conversely, $\cos\varphi \approx 0$ implies that the measurement of X is not likely to constrain δY strongly.[†]

The parameters of the correlation ellipse are sufficient to deduce, under conventional approximations, a Gaussian probability distribution $P(X, Y|\text{CTEQ6.6})$ for finding certain values of X and Y based on the pre-LHC data sets included in the CTEQ6.6 analysis. If the LHC measures X and Y nearly independently of the PDF model, a new confidence region for X and Y satisfying both the CTEQ6.6 and LHC constraints can be determined by combining the prior probability $P(X, Y|\text{CTEQ6.6})$ with the new probability distribution $P(X, Y|\text{LHC})$ provided by the LHC measurement. For this purpose, it suffices to construct a probability distribution

$$P(X, Y|\text{CTEQ6.6+LHC}) = \\ = P(X, Y|\text{CTEQ6.6})P(X, Y|\text{LHC}) \tag{A8}$$

which establishes the combined CTEQ6.6+LHC confidence region without repeating the global fit.

The values of ΔX, ΔY, and $\cos\varphi$ are also sufficient to estimate the PDF uncertainty of any function $f(X, Y)$ of X and Y by relating the gradient of $f(X, Y)$ to $\partial_X f \equiv \partial f/\partial X$ and $\partial_Y f \equiv \partial f/\partial Y$ via the chain rule:

$$\Delta f = \left|\vec{\nabla} f\right| = \sqrt{(\Delta X\ \partial_X f\ )^2 + 2\Delta X\ \Delta Y\ \cos\varphi\ \partial_X f\ \partial_Y f + (\Delta Y\ \partial_Y$$

(A9)

Of particular interest is the case of a rational function $f(X, Y) = X^m/Y^n$, pertinent to computations of various cross section ratios, cross section asymmetries, and statistical significance for finding signal events over background processes [2]. For rational functions Eq. (A9) takes the form

$$\frac{\Delta f}{f_0} = \sqrt{\left(m\frac{\Delta X}{X_0}\right)^2 - 2mn\frac{\Delta X}{X_0}\ \frac{\Delta Y}{Y_0}\ \cos\varphi\ + \left(n\frac{\Delta Y}{Y_0}\right)^2}. \tag{A10}$$

For example, consider a simple ratio, $f = X/Y$. Then $\Delta f/f_0$ is suppressed ($\Delta f/f_0 \approx |\Delta X/X_0 - \Delta Y/Y_0|$) if X and Y are strongly correlated, and it is enhanced ($\Delta f/f_0 \approx \Delta X/X_0 + \Delta Y/Y_0$) if X and Y are strongly anticorrelated.

As would be true for any estimate provided by the Hessian method, the correlation angle is inherently approximate. Eq. (A6) is derived under a number of simplifying assumptions, notably in the quadratic approximation for the χ^2 function within the tolerance hypersphere, and by using a symmetric finite-difference formula (A2) for $\{\partial_i X\}$ that may fail if X is not monotonic. With these limitations in mind, we find the correlation angle to be

[†] The allowed range of $\delta Y/\Delta Y$ for a given $\delta \equiv \delta X/\Delta X$ is $r_Y^{(-)} \leq \delta Y/\Delta Y \leq r_Y^{(+)}$, where $r_Y^{(\pm)} \equiv \delta\cos\varphi \pm \sqrt{1-\delta^2}\sin\varphi$.

a convenient measure of interdependence between quantities of diverse nature, such as physical cross sections and parton distributions themselves. For collider applications, the correlations between measured cross sections for crucial SM and beyond SM processes will be of primary interest, as we shall illustrate in Sec. **??**. As a first example however, we shall present some representative results on correlations between the PDFs in the next section.