

# VALIDATION OF HADRON MASS CORRECTION SCHEMES IN DEEP INELASTIC SCATTERING AT LOW ENERGY TRANSFER

Andy Krause

Advisor: Alberto Accardi

Hampton University

## **Abstract:**

This research focuses on validating the range at which factorization theorems still accurately describe the momentum distribution of quarks. Specifically, we are interested in values of large momentum distribution, referred to as high- $x_B$ , where quarks start to experience a phenomenon known as quark confinement. In the high- $x_B$  regimen, modern techniques used to calculate quantum interactions such as perturbation theory (pQCD), are unable to accurately describe the distribution of the quarks momentum. Thus, correction terms, known as “Hadron Mass Corrections”, are needed and are a focus of study at Jefferson Lab. The goal of our research is to assess the viability of the proposed extended version of the QCD factorization theorem, whose validity can be explicitly tested in the model and applied with renewed confidence to experimental data. To accomplish this, an analytically calculable distribution function is used to simulate Deep Inelastic Scattering experiments, such as those done at Jefferson Lab. These simulated experiments are then used to analytically test the viability of the proposed factorization theorems.

- for example, an electron off a proton, whereby the lepton interacts with the quarks that compose the target in a bid to study its internal structure. That’s how the prediction by Murray Gell-Mann’s of the very existence of these elementary particles, indirectly derived from the the structure and regularity of the hadron spectrum in the 1960s, were confirmed a decade later by DIS experiments. Under normal conditions, indeed, quarks only exist confined inside heavier composite particles called hadrons, of which the proton is an example. There, they vigorously interact with other quarks via gluon exchange by means of the “strong nuclear force”. Combined, the quark and gluon interactions give rise to the observed properties of their host hadron, yet these particle cannot be detected by themselves. Conversely, high-energy scattering experiments between electrons, protons, and even nuclei, have provided the scientific community with a flexible tool for finer and finer measurements of hadron structure, and the dynamics of their constituent “partons”: the quarks and gluons.

## **Introduction**

### **Background**

Deep Inelastic Scattering (DIS) is the scattering of a high energy lepton off a target hadron

### **Motivation**

The focus of this research is on understanding confinement at ordinary density and temperature by means of DIS experiments. What enables the investigation we have in mind is the QCD factorization theorem, that reformulates the quarks cross sections as a convolution

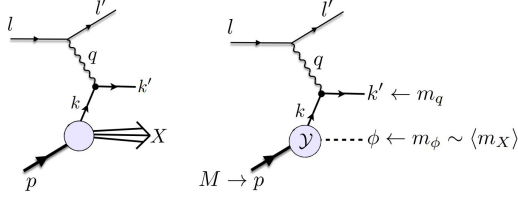


Figure 1: *Left*: Electron scattering off a proton resulting in the emission of a quark and a "jet" of hadrons. *Right*: Electron scattering off proton is diquark view.

of theoretically calculable parton-level interactions and so-called Parton Distribution Functions (PDF) that characterizes the quark (q) and gluon (g) momentum distribution within the hadron. These can then be extracted in global QCD fits by comparing the calculated and measured cross sections.

### Research Goal

**The goal of our research is to assess the viability of the proposed extended version of the QCD factorization theorem, whose validity can be explicitly tested in the model and applied with renewed confidence to experimental data.** Extrapolations of these measurements to  $x_B \rightarrow 1$  will then inform one on the way QCD dynamically confines quarks and gluons inside the proton.

### Deep Inelastic Scattering

The process under investigation is a high energy electron scattering off quarks inside the proton via an emitted virtual photon of 4-momentum  $q$ , as shown in Fig. 1, *Left*. The virtuality of the photon,  $Q^2 = -q \cdot q$ , can be thought of as the squared 4-momentum transferred during the scattering process. While at large enough  $Q^2$  the photon essentially scatters on a quasi-free quark in a calculable manner, the confining nature of the strong nuclear interactions precludes a the-

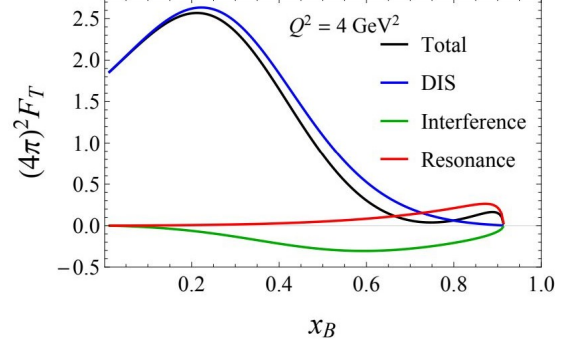


Figure 2: Gauge invariant decomposition of the transverse structure function  $F_T$  at  $Q^2 = 4 \text{ GeV}^2$  for model parameters fitted to phenomenologically determined quark PDFs. On the horizontal axis, the Bjorken variable  $x_B = Q^2/(2p^\mu q_\mu)$  can be interpreted as the fraction of momentum carried by the quark.

oretical calculation of the quark PDF, that can only be experimentally measured as discussed above.

In order to test the factorization theorem regime of applicability, and extend this to the largest possible  $x_B$  and smallest possible  $Q^2$ , we utilize a model theory which describes the essence of QCD in DIS, see Fig. 1, *Right*. In the model one can calculate not only can the DIS scattering off the quark, to which one applies pQCD factorization, but also (as it happens in real processes) the resonant excitation of the target and the interference of these two processes, see Fig. 2. as shown in a recent paper by Guerrero and Accardi [GA20]. The model's factorized cross section can then be analytically compared to the full one.

More specifically, the values of the large momentum distribution of quarks confined in a hadron are particularly sensitive to the mechanism by which the strong interactions confines them, but lies at the edge of the applicability of perturbative factorization techniques (pQCD). In this regime, which is currently been experimentally probed at Jefferson Lab, a number of corrections to the standard theoretical cal-

culations are need to interpret the experimental data and extract the quark distributions [GA20]. These corrections, known as "Hadron Mass Corrections", become non-negligible at lower values of  $Q^2$ .

Within the adopted model, as opposed to real process, one can analytically calculate the full scattering cross section, and separate the purely DIS component whereby the photon scatters off the quark [GA20]. One can also derive the quark distribution,  $q(x)$ , analytically in terms of the masses of the particles involved and of the quark's momentum fraction  $x$  [GA20].

The factorization theorem can then be tested by generating pseudo-data for the model's scattering cross section, approximating this with the factorized formula, and using the latter to extract a quark PDF by fitting the data. The factorized formula will be deemed successful if the fitted PDF statistically compares well enough with the analytical one. We will thus be able to find the validity of factorized pQCD in energy regions where hadron mass corrections are non-negligible and thereby piece together a new factorized formulation of QCD to the prescribed region.

### Event Simulation

The initial objective of this project is to simulate the inclusive electron-proton scattering events described by the model in the previous section. Each event consists of a measurement made of the 4-momentum transferred between the electron and the proton, measured by the pair of variables  $(x_B, Q^2)$ . These events will then be used to test the factorization formulas. A more detailed description of the statistical techniques used to analyze the event generation can be found in works such as [DS10] and/or [Tay97] as well as [KNS20].

In order to generate the events a standard Monte-Carlo method was used. That is, given a probability distribution function (p.d.f.),  $f(x)$ ,

of a stochastic variable  $x$ , then the probability of an event,  $x$ , falling between  $x_a$  and  $x_b$ ,  $\Pr[x_a \leq x \leq x_b]$ , is calculated via:

$$\Pr[x_a \leq x \leq x_b] = \int_{x_a}^{x_b} f(x) dx . \quad (1)$$

One may then create an array of events,  $\{x_i\}$ , distributed according to  $f(x)$  by repeatedly generating a random pair of coordinates,  $(x, y)$ , uniformly over a domain of  $x \in [x_{min}, x_{max}]$  and  $y \in [y_{min}, y_{max}]$ . If the values of  $y$  fall under the function, i.e.,

$$y < f(x) \quad (2)$$

then the corresponding  $x$  value is added to the array of events. Otherwise the pair is discarded. This process continues until the desired number,  $N$ , of events are collected within the events array

$$\text{Events} = \{x_1, x_2, \dots, x_N\} . \quad (3)$$

In this case the events,  $\{(x_i, Q_i^2)\}$ , are the measurements of a scattered electron. Thus the Monte-Carlo method was applied by generating a set of coordinates  $(x_B, Q^2, y)$  across a 3 dimensional domain via

$$f(x_B, Q^2) = \frac{1}{A} F_T(x_B, Q^2) \quad (4)$$

where

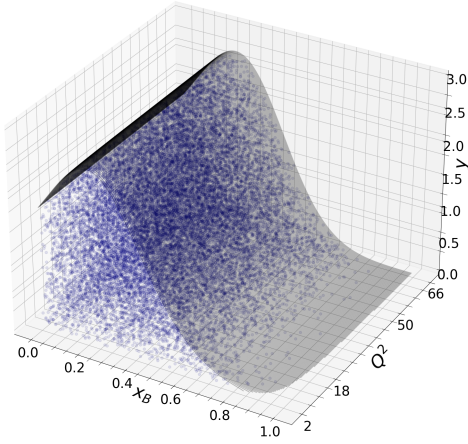
$$A_{DIS} = \int_0^1 dx_B dQ^2 F_T(x_B, Q^2) . \quad (5)$$

from [GA20]. Fig. 3 shows a sample of accepted events distributed via the model DIS function, where the shaded area represents the DIS structure function.

### Binning

Binning is the process of dividing the the domain into discrete intervals and grouping the events which fall into those intervals. For simplicity the intervals are divided into,  $K$ , rectangles of size  $\delta x_B$  by  $\delta Q^2$  such as that in Fig. 3. Choosing bin sizes is not arbitrary, however, and special care must be taken.

Distribution of 25k Events generated via  $F_T^{DIS}$



Phase Space of 25k Events generated via  $F_T^{DIS}$

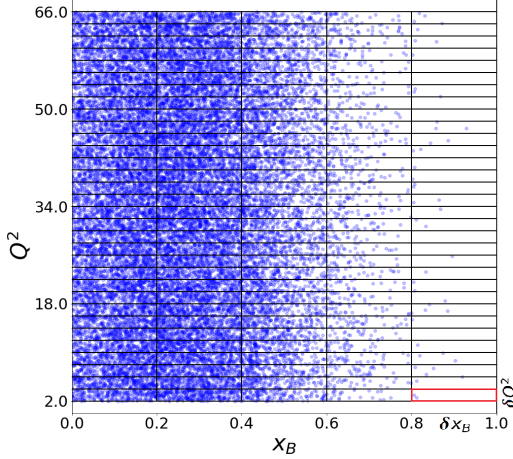


Figure 3: Distribution of 25000 events generated via  $F_T^{DIS}(x_B, Q^2)$ . The left figure shows the generated  $(x_B, Q^2, y)$  coordinate with the underlying function in black. The figure on the right shows the corresponding phase space, such as if one were to down upon the figure on the left.

### Bin Sizes

Although different bin choices cannot improve the overall statistics of the generated events, it allows one to interpret them differently. Different bin sizes will inherently change the amount of events that fall into it, ultimately changing the statistics for the individual bins. Generally, one wants to take as small bins as possible to maximize the information information of the shape

of the underlying function. However, too small of bin sizes can severely affect the uncertainty of our counts. Thus careful consideration had to be taken to ensure an ideal bin size was being used.

### High- $x_B$ region

Our area of interest is that of high energy partons on the edge of confinement. This corresponds to measurements made in the high- $x_B$  region. However, this region also corresponds to a highly unlikely state of the parton and needs careful consideration.

If there are not enough points in a bin, the measurement will not be statistically significant. At the very minimum, 5 events must be in the bin to be statistically significant. Specifically, we needed a way to find a way to increase the statistical significance of the high  $x_B$  region, in order to precisely validate the factorization theorems.

### Enhanced Data

Although larger bin sizes could capture more points in the high- $x_B$ , we would lose resolution. Thus, we enhanced the data in the high- $x_B$  region. A new array of  $m$  events were generated over a domain of  $\kappa = 0.6 \leq x_B \leq 1$  and added to an array of events in which there are already  $n_\kappa$  events within that same region.

The enhanced data must be re-scaled, or weighted, appropriately, however, or else it will not represent the actual distribution function. That is, given a total of  $m + n_\kappa$  events, it must be re scaled to represent it's original number of events  $n_\kappa$ .

### Statistical Analysis of the Event Generation

With an ideal scheme for binning, one may generate a histogram of the counts in each bin and begin to parameterize a fit for it. However, before fitting the data, it is useful to make sure that

the data statistically corresponds to the underlying p.d.f. used to generate it.

Before trying to fit the data and test the sub-asymptotic factorization formulas, it is first important to make sure that the generated events match the underlying p.d.f,  $f(x)$ , used to generate them. That is, we will compare the observed counts,  $Ob_k$ , to the expected counts,  $Ex_k$ . This is for two reasons. First, to test that the events are being generated correctly. Otherwise, the eventual fits will not match the p.d.f.,  $f(x)$ . Second, to ensure that the statistical analysis program which was developed for this project is also correctly analyzing the data.

Assuming the events are properly generated, we know they will be statistically representative of the underlying p.d.f., which is analytically known in our case. Thus, if everything is running properly, the analysis of the events will also demonstrate what one should expect to see when the events are compared to a fitted function that statistically represents its underlying distribution.

### **Observed Counts for the Generated Events**

Once the events are divided into bins, a discrete distribution function can be created via a histogram as seen below. The counts of the generated events will be referred to as the observed counts,  $Ob_k$ , since they represent measurements which would be observed in a DIS experiment. The histogram generates the basic shape which will then be used to generate a continuous function.

### **Expected Counts for Continuous p.d.f.s**

In order to compare it to the underlying p.d.f., one first must find the expected amount of counts,  $Ex_k$ , for the underlying p.d.f.,  $f(x)$ , in each bin. That is, for each bin,  $k$ , whose boundaries are  $x_a$  and  $x_b$ , the expected number of

counts are:

$$Ex_k = N \int_{x_a}^{x_b} f(x) dx . \quad (6)$$

This could naturally be expanded to any number of dimensions. With the observed and expected values for each bin, a series of tests can be performed to see how well the data matches what one would expect the data for a known p.d.f.

### **Ratio (Relative Deviation)**

With a set of observed data points and expected values, analytical tests, such as the ratio between two values centered around 0.

$$\text{Ratio} = \frac{Ob_k - Ex_k}{Ex_k} \pm \frac{\sqrt{Ob_k}}{Ex_k} \quad (7)$$

This value is simply the relative deviation of the observed counts from the expected count in a given bin, and one would expect them to fluctuate around 0. While this test is able to show where the two fits diverge up to a given percentage, it does not take into account the level of precision of the generated data. Therefore this only gives a rough estimate of the goodness of fit.

### **Residual (Statistical Uncertainty)**

In order to find the level of uncertainty in the fit, one may consider the residual between the two counts. The residual simply takes the difference between the observed and expected values and scales it to the expected standard deviation or uncertainty.

$$Res_k = \frac{(Ob_k - Ex_k)}{\sqrt{Ex_k}} \pm \frac{\sqrt{Ob_k}}{\sqrt{Ex_k}} \quad (8)$$

### **Residual Distribution**

In fact, for a distribution of residuals values,

$$R = \{Res_1, Res_2, \dots, Res_K\} , \quad (9)$$

as  $K \rightarrow \infty$  the distribution must approach a normal distribution, as this, by definition, is what makes a random process normally distributed according to some underlying p.d.f.. Thus, the residual distribution is a good indicator of both the magnitude and direction of the fluctuations. As mentioned, the average residual should correspond with 0, if the data is fluctuating around the correct values, and should have a standard deviation equal to 1 indicating the expected error for a normal distribution.

### Chi-Squared, $\chi^2$

Another test, known as the  $\chi^2$  test, is used, where:

$$\chi_K^2 = \sum_k \text{Res}_k^2 = \sum_k \frac{(\text{Ob}_k - \text{Ex}_k)^2}{\text{Ex}_k} \quad (10)$$

The chi squared test squares the residuals and then takes the sum. Since the residuals are squared, the sign of the residual disappears and only the magnitude of the residual becomes apparent. This allows one to find the total "error" caused by the fluctuations. If the events are representative of the underlying p.d.f., on average the bin's should have a standard deviation of 1, introducing the concept of  $\chi^2$  per degree of freedom.

$$\chi_K^2 / \text{DoF} = \sum_k \frac{\text{Res}_k^2}{\text{DoF}} = \sum_k \frac{(\text{Ob}_k - \text{Ex}_k)^2}{K \text{Ex}_k} \quad (11)$$

where the degrees of freedom, DoF are, for now, equal to the number of bins,  $K$ . Thus, to make the statement that a set of data is representative of some underlying distribution, the bins on average should be 1 standard deviation away from the expected values. This corresponds to  $\chi^2 \approx K$  and  $\chi^2 / \text{DoF} \approx 1$ .

Ultimately, both tests are needed and used to tell if the data is representative of a given distribution function,  $f(x_B, Q^2)$ .

### Chi-Squared Distribution

The  $\chi^2$  value for a single experiment gives a numerical value for the quality of fit, the distribution of these values provides insight into how reliable the experiments are. One important feature of the  $\chi^2$  distribution is that as  $k \rightarrow \infty$  the shape of the distribution approaches that of a Gaussian distribution of width  $\sqrt{2k}$  centered around  $k$ . Thus, for a distribution of  $\chi^2$  values,

$$X = \{X_1, X_2, \dots, X_J\} \quad (12)$$

which are distributed approximately to that of  $f_k(x)$  then

$$\frac{(X - k)}{\sqrt{2K}} \quad (13)$$

should approach a normal distribution as  $k \rightarrow \infty$ .

### PDF Fitting Framework

With the events binned and scaled properly a parameterized function is used to fit the Data. With a function fit to the data, the same type of analysis as done on the discrete data can be done on the function to see how well it fits the underlying distribution function. Generating a good fit not only depends on how well the data is binned, but it also depends on the parameterization of the fits.

### Parameterizing $\Phi(x)$ and $\Omega(x_B, Q^2)$

With the discrete distribution function, a set of parameters can be found which minimizes the  $\chi^2$  value for that fit compared to the data. Specifically, the code uses the Levenberg-Marquardt algorithm as a method of least squares optimization. If a fit is able to be perfectly parameterized, then the fit should have a  $\chi^2 / \text{DoF} \approx 1$ .

With our goal of validating the range of applicability of the factorized structure functions we need two functions, which I call  $\Phi(x)$  and

$\Omega(x_B, Q^2)$ . The first will be the leading order term,  $\Phi(x)$ , which will be compared to  $q(x)$ . The other,  $\Omega(x_B, Q^2)$ , will absorb higher order terms based on which observable we are taking into account and will be compared to  $F_T^{DIS}$ .

### **Bernstein Polynomials**

Rather than create our own polynomial we chose to use Bernstein Polynomials:

$$B_\eta(x) = \sum_{\nu=0}^{\eta} \beta_\nu b_{\nu,\eta}(x) \quad (14)$$

Where the Bernstein basis:

$$b_{\nu,\eta}(x) = \binom{\eta}{\nu} x^\nu (1-x)^{\eta-\nu} \quad (15)$$

form a basis for the vector space of polynomials no more than degree  $\eta$ . The coefficients  $\beta_\eta$  are known as the Bernstein coefficients and are the free parameters which will be used to fit our function.

### **Parameterizing the Quark PDF, $\Phi(x)$**

In conjunction with the Bernstein Polynomials, we multiplied the entire polynomial by the restriction,  $(1-x)^B$ , in order to set the restriction  $q(x) \xrightarrow{x \rightarrow 1} 0$ . This we simply called a modified Bernstein Polynomial,  $mB(x)$ . More specifically, a 8th degree  $mB_8$  was used to fit the data such that:

$$\Phi(x) = mB_8(x) = (1-x)^B \sum_{\nu=0}^8 \beta_\nu b_{\nu,8}(x) \quad (16)$$

is used to fit the experimental data with respect to  $B, \beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8$ .

### **Parameterizing the observables, $\Omega(x_B, Q^2)$**

Just as there are multiple scaling values of  $\bar{x}$  used to generate data [GA20], the different scaling values of  $\bar{x}$  can also be used to fit the data.

Fitting the data with different values of  $\bar{x}$  can be thought as fitting different observables to a set of data. That is, for each set of data, each observable can be fit via:

$$\Omega^{\bar{x}}(x_b, Q^2) = \Phi(x) \big|_{x=\bar{x}} \quad (17)$$

This model can then be used to see how well the Bernstein Polynomials are able to fit the generated data. The scaling variable  $\bar{x}$  can also, again then be used on the generated fits. That is, given a fit which used  $\bar{x}^{Gen}$  to generate the fit,  $\bar{x}$  may then scale the fit to give.

### **Higher Twist Corrections $\Omega^{HT}(x_b, Q^2)$**

The second way we will attempt to separate the qPDF from the higher order twist terms is by directly making a  $Q^2$  dependant term. This correction term will be fit such that the dependence on  $Q^2$  will be extracted from the fit and put in higher correction terms  $H(x, Q^2)$ . The observable in this case:

$$\Omega^{HT}(x_b, Q^2) = \Phi(x) + \frac{H(x)}{Q^2} \quad (18)$$

This will leave a model of the  $Q^2$  independent  $\Phi(x)$  to analyze directly with  $q(x)$ .

### **The Average Fit**

And finally an average fit including error can be generated by simply taking the mean and standard deviation of the individual fits. That is:

$$\mu\Phi_j(x) \pm \sigma\Phi_j(x) = \frac{1}{J} \sum \Phi_j \pm \sqrt{\frac{1}{J} \sum (\Phi_j - \mu\Phi_j)^2} \quad (19)$$

## **Analysis and Conclusion**

### **Continuous Residual Distribution Function**

Again, with our goal of finding the range a validity between the DIS structure function and

our factorized models it will be more useful to define our residuals in terms of a continuous function rather than discrete counts. The scaled residuals between two p.d.f.s can be thought of as a residual distribution function, which I define as:

$$\text{RDF}(x_k) = \frac{\phi(x_k) - q(x_k)}{\sqrt{q(x_k)}} \sqrt{\frac{K}{N}} \quad (20)$$

where:

$$\text{Res}_k \delta x \approx \int_{x_k}^{x_{k+1}} \text{RDF}(x) dx \quad (21)$$

This allows the residual to be expressed in terms of the discrete distribution functions.

$$\text{Res}(x_k) = \frac{\phi(x_k) - q(x_k)}{\sqrt{q(x_k)}} \sqrt{\frac{N}{K}} \quad (22)$$

### Analysis of $\Phi(x)$ and $\Omega(x)$

After fitting the DIS data, the fits were then compared to the DIS data as shown in Figures 4 and 5. The top plots show the difference between the functions. The middle plots shows the residual distributions between  $\Phi(x)$  and  $q(x)$ , while the bottom plot shows the residual distributions between the observables  $\Omega(x_B, Q^2)$  with  $F_T^{DIS}$ .

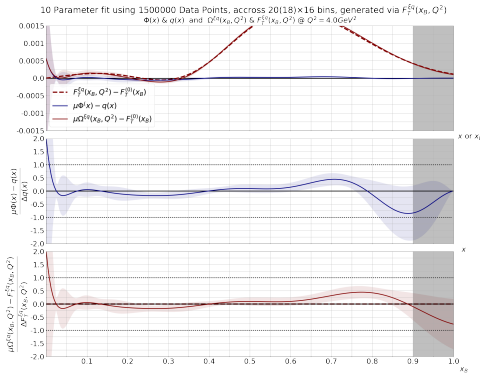


Figure 4: Fit of an 8th degree polynomial to that of the DIS function.

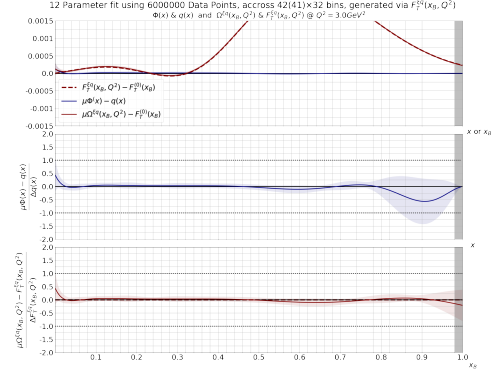


Figure 5: Fit of an 8th degree polynomial with an added higher power term to that of the DIS function.

### Conclusions

In conclusion, it is clear that the factorized formulas are able statistically viable up to  $0.8x_B$  when different choices of  $\bar{X}$  are chosen. However with the additional higher power term, the factorized formulas are able to extract the individual shape of the DIS function better. As shown in Fig. 5, the added higher power term was able to boost the statistical viability up to about  $0.9x_B$ . In the end we will continue to try different ways of parameterize our fit functions, in order to try improve the range of our factorized formula viability.

---

## References

- [DS10] Morris H DeGroot and Mark J Schervish. *Probability and Statistics*. 4th ed. Upper Saddle River, NJ: Pearson, Dec. 2010.
- [GA20] Juan V. Guerrero and Alberto Accardi. “Collinear Factorization at sub-asymptotic kinematics and validation in a diquark spectator model”. In: (Oct. 2020). arXiv: 2010.07339 [hep-ph].
- [KNS20] Karol Kovařík, Pavel M. Nadolsky, and Davison E. Soper. “Hadronic structure in high-energy collisions”. In: *Rev. Mod. Phys.* 92.4 (2020), p. 045003. DOI: 10 . 1103 / RevModPhys . 92 . 045003. arXiv: 1905.06957 [hep-ph].
- [Tay97] John R Taylor. *Introduction to Error Analysis, second edition*. en. 2nd ed. Sausalito, CA: University Science Books, July 1997.