REGRESSION NETWORKS: PRECISION AND UNCERTAINTY ESTIMATION

Stefano Forte

Università di Milano & INFN

with tutorials by

Tommaso Giani

NIKHEF

UNIVERSITÀ DEGLI STUDI DI MILANO

DIPARTIMENTO DI FISICA

Istituto Nazionale di Fisica Nucleare

Machine Learning in Particle Theory

MITP, July 3, 2023

LECTURE PLAN

- I: THE PHYSICS PROBLEM
  - PARTON DISTRIBUTIONS AND THEIR DETERMINATION
  - THE PROBLEM OF PDF UNCERTAINTIES
- II: REGRESSION & UNCERTAINTY
  - MONTE CARLO UNCERTAINTIES
  - NEURAL NETWORK REGRESSION
- III: PROPER LEARNING
  - OVERLEARNING AND CROSS-VALIDATION
  - HYPEROPTIMIZATION AND GENERALIZATION
- IV: VALIDATION & TESTING
  - TESTING: DATA REGION AND EXTRAPOLATION REGION
  - THE MEANING OF CORRELATIONS
- V: UNDERSTANDING RESULTS
  - DISTRIBUTION OF RESULTS AND FAITHFULNESS
  - OUTLIERS AND GENERALIZATION

SOME GENERAL REFERENCES

REGRESSION AND STATISTICS

- C. M. Bishop, "Pattern recognition and machine learning", Springer, 2006-2009
- G. James, D. Witten, T. Hastie, R. Tibishrani, "An introduction to statistical learning", Springer, 2021

GENERAL MACHINE LEARNING

- I. Goodfellow, Y. Bengio, A. Courville, "Deep learning", MIT press, 2016
- A. Géron, "Hands-on machine learning with Scikit-Learn & TensorFlow", O'Reilly, 2017

PARTON DISTRIBUTIONS

- S. Forte, "Parton Distributions at the dawn of the LHC", Acta Phys. Pol. **B41** (2010) 2859, `arXiv:1011.5247`
- S. Forte and S. Carrazza, "Parton distribution functions", in "Artificial intelligence for High Energy Physics", P. Calafiura, D. Rousseau and K. Terao, eds., World Scientific 2022, pag 715, `arXiv:2008.12305`
- L. Del Debbio, T. Giani, T. Wilson, "Bayesian approach to inverse problems: an application to NNPDF closure testing", Eur. Phys. J. **C82** (2022) 33, `arXiv 2111.057887`

THE NNPDF CODE

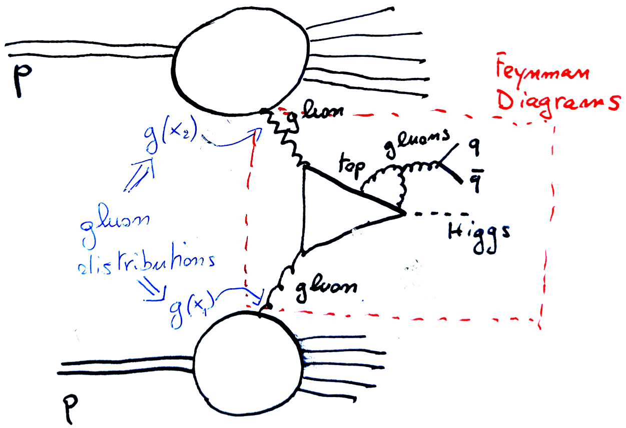
`https://nnpdf.mi.infn.it/nnpdf-open-source-code/`

I: THE PHYSICS PROBLEM

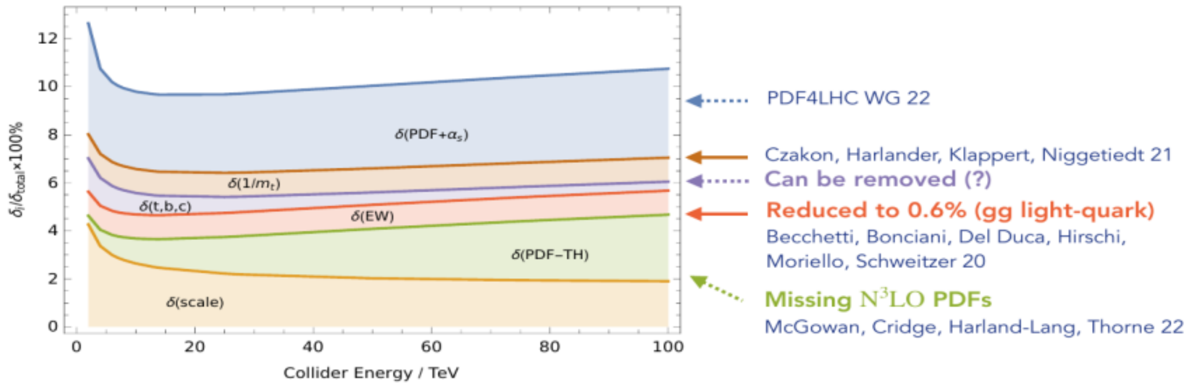
- FACTORIZATION AND PDFs
  - FACTORIZATION: LEPTONIC AND HADRONIC PROCESSES
  - PDFs AND THEIR PROPERTIES
- PDF DETERMINATION
  - FROM DATA TO PDFs
  - PHYSICAL PROCESSES
- UNCERTAINTIES
  - EXPERIMENTAL UNCERTAINTIES AND THEIR CORRELATIONS
  - MISSING HIGHER ORDERS AND THEORY UNCERTAINTIES
- THE PROBLEM OF MODEL UNCERTAINTIES
  - THE PROBLEM OF PDF UNCERTAINTIES
  - PLOTNOMIAL REGRESSION AND ITS PITFALLS

PDFs AND THEIR UNCERTAINTIES

QCD FACTORIZATION



UNCERTAINTIES: HIGGS IN GLUON FUSION



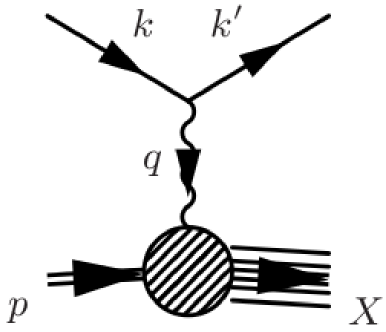
(R. Roentsch, Les Houches, June 2023)

- PDF ESPRESS THE "PROBABILITY" OF QUARKS OR GLUONS (PARTONS) TO ENTER A COLLISION
- THEIR KNOWLEDGE IS REQUIRED FOR THE COMPUTATION OF ANY PROCESS AT THE LHC
- THEIR KNOWLEDGE IS A DOMINANT SOURCE OF UNCERTAINTY

FACTORIZATION

FACTORIZATION IN DEEP-INELASTIC SCATTERING

STRUCTURE FUNCTIONS



Lepton fractional energy loss: $y = \frac{p\cdot q}{p\cdot k}$;
gauge boson virtuality: $q^2 = -Q^2$
Bjorken x: $x = \frac{Q^2}{2p\cdot q}$
lepton-nucleon CM energy: $s = \frac{Q^2}{xy}$;
virtual boson-nucleon CM energy $W^2 = Q^2\frac{1-x}{x}$;

$$\frac{d^2\sigma^{\lambda_p\lambda_\ell}(x,y,Q^2)}{dxdy} = \frac{G_F^2}{2\pi(1+Q^2/m_W^2)^2}\frac{Q^2}{xy}\left\{\left[-\lambda_\ell\, y\left(1-\frac{y}{2}\right)xF_3(x,Q^2) + (1-y)F_2(x,Q^2)\right.\right.$$

$$\left.\left.+y^2xF_1(x,Q^2)\right] - 2\lambda_p\left[-\lambda_\ell\, y(2-y)xg_1(x,Q^2) - (1-y)g_4(x,Q^2) - y^2xg_5(x,Q^2)\right]\right\}$$

$\lambda_l \to$ lepton helicity
$\lambda_p \to$ proton helicity

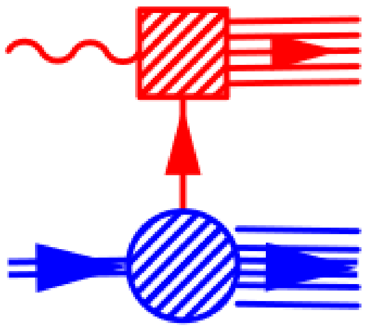
	PARITY CONS.	PARITY VIOL.
UNPOL.	F_1, F_2	F_3
POL.	g_1	g_4, g_5

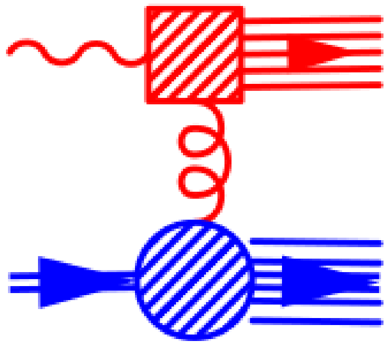
DEEP-INELASTIC SCATTERING

STRUCTURE FUNCTIONS AND PDFs

STRUCTURE FUNCTION=HARD COEFF. (PARTONIC STRUCTURE FUNCTION) $\otimes$PARTON DISTN.

HARD COEFF.: XSECT FOR INCOMING PROTON MOMENTUM $p \Rightarrow$ INCOMING PARTON MOMENTUM $\hat{p} = xp$





$$F_2(x,Q^2) = x\sum_i \int_1^1 \frac{dy}{y} C_i\left(\alpha_s(Q^2), \frac{x}{y}\right)\left[q_i(y,Q^2)+\bar{q}_i(y,Q^2)\right] + C_g\left(\alpha_s(Q^2), \frac{x}{y}\right) g(y,Q^2)$$

q_i quark, $\bar{q}_i$ antiquark, g gluon

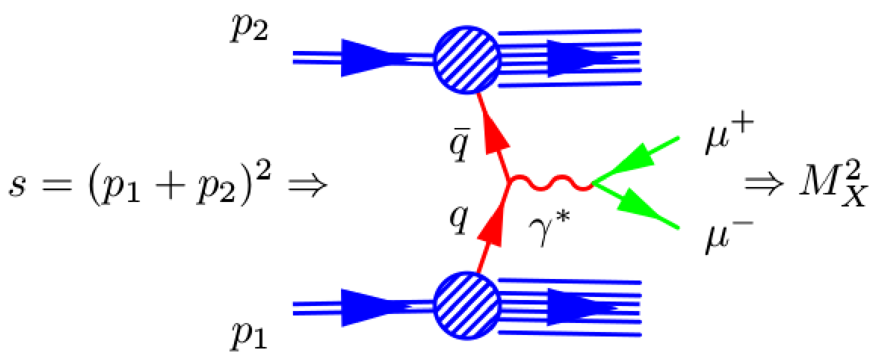
FACTORIZATION
HADRONIC PROCESSES
THE PARTON LUMINOSITY

$$\sigma_X(s, M_X^2) = \sum_{a,b} \int_{x_{\min}}^1 dx_1\, dx_2\, f_{a/h_1}(x_1) f_{b/h_2}(x_2) \hat\sigma_{q_a q_b \to X}\left(x_1 x_2 s, M_X^2\right)$$

$$\sigma_X(s, M^2) = \sigma_0 \sum_{a,b} \int_\tau^1 \frac{dx}{x} \mathcal{L}_{ab}\left(\frac{\tau}{x}\right) C\left(x, \alpha_s(M_H^2)\right)$$

- PARTON LUMINOSITY $\mathcal{L}_{ab}(\tau) = \int_\tau^1 \frac{dx}{x} f_{a/h_1}(x) f_{b/h_2}(\tau/x)$
- COEFFICIENT FUNCTION $\hat\sigma_{q_a q_b \to X}\left(x_1 x_2 s, M_X^2\right) = \sigma_0 C\left(\frac{M_X^2}{x_1 x_2 s}, \alpha_s(M_H^2)\right)$

EXAMPLE: THE DRELL-YAN PROCESS
AT LEADING ORDER



- Hadronic c.m. energy: $s = (p_1 + p_2)^2$
- Momentum fractions $x_{1,2} = \sqrt{\frac{\hat s}{s}} \exp \pm y$;
  Lead. Ord. $\hat s = M^2$
- Partonic c.m. energy: $\hat s = x_1 x_2 s$
- Invariant mass of final state X (dilepton, Higgs,…): $M_W^2 \Rightarrow$ scale of process
- Scaling variable $\tau = \frac{M_X^2}{s}$

$$\Rightarrow M^2 \frac{d\sigma}{dM^2} = \sigma_0 \mathcal{L}(\tau); \quad \sigma_0 = \frac{4}{9}\pi\alpha \frac{1}{s};$$

FACTORIZATION SUMMARY & MELLIN TRANSFORM

- x-SPACE FACTORIZED EXPRESSIONS:
  - LEPTON-HADRON
    $F(x,Q^2)=\sum_i \int_x^1 \frac{dy}{y} C_i\left(\alpha(Q^2),\frac{x}{y}\right) f_i(y,Q^2)=\left[C_i(\alpha(Q^2))\otimes f_i(Q^2)\right](x)$
  - HADRON-HADRON
    $\sigma(x,Q^2)=\sum_{ij}\int_x^1 \frac{dy}{y}\hat\sigma_{ij}\left(\alpha(Q^2),\frac{x}{y}\right)\mathcal{L}_{ij}=(y,Q^2)\left[\hat\sigma_{ij}(\alpha(Q^2))\otimes\mathcal{L}_{ij}(Q^2)\right](x)$
    $\hat\sigma_{ij}\left(\alpha(Q^2),x\right)=\sigma_0 C_{ij}\left(\alpha(Q^2),x\right);$
    $\mathcal{L}_{ij}(Q^2)(x)=\int_x^1\frac{dy}{y} f_i\left(\frac{x}{y}\right) f_j(y,Q^2)=\left[f_i\otimes f_j\right](x)$

- MELLIN TRANSFORM: $F(N)=\int_x^1 x^{N-1} f(x)\Leftrightarrow (x)=\int_{-i\infty}^{+i\infty} x^{-N} f(N)$
  $h(x)=\int_x^1\frac{dy}{y} f\left(\frac{x}{y}\right) g(y)=\left[f\otimes f\right](x)\Leftrightarrow H(N)=F(N)G(N)$

- N-SPACE FACTORIZED EXPRESSIONS:
  - LEPTON-HADRON $F(N,Q^2)=\sum_i C_i\left(\alpha(Q^2),N\right) f_i(N,Q^2)$
  - HADRON-HADRON $\sigma(x,N)=\sum_{ij}\hat\sigma_{ij}\left(\alpha(Q^2),N\right)\mathcal{L}_{ij}(N,Q^2)$
    $\mathcal{L}_{ij}(Q^2,N)=f_i(N)f_j(N)$

LARGE/SMALL x $\Leftrightarrow$ LARGE/SMALL N

THE SCALE DEPENDENCE OF PDFS
EVOLUTION EQUATIONS

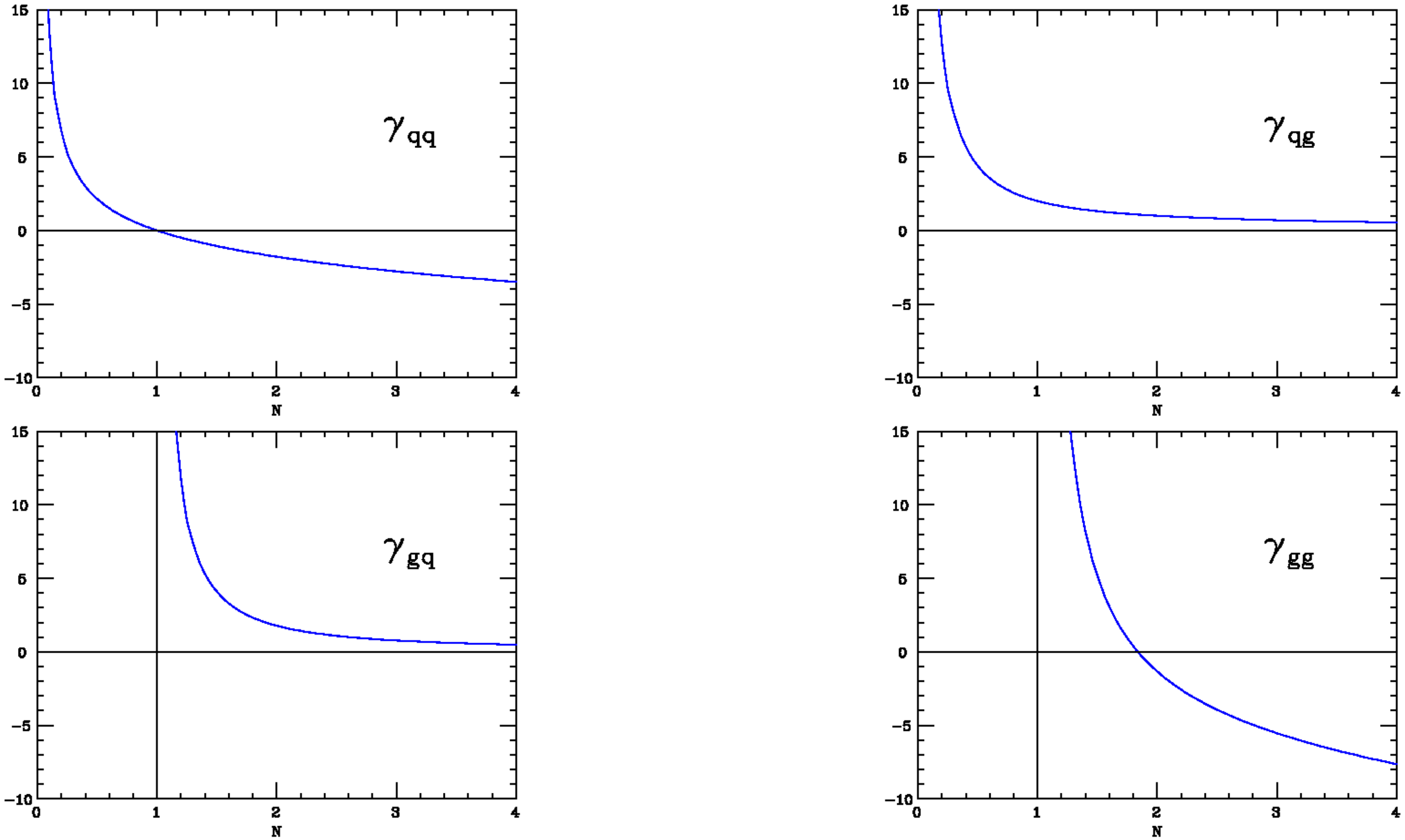
$$\frac{d}{dt} q_{NS}(N, Q^2) = \frac{\alpha_s(t)}{2\pi} \gamma_{qq}^{NS}(N, \alpha_s(t)) q_{NS}(N, Q^2),$$

$$\frac{d}{dt} \begin{pmatrix} \Sigma(N, Q^2) \\ g(N, Q^2) \end{pmatrix} = \frac{\alpha_s(t)}{2\pi} \begin{pmatrix} \gamma_{qq}^S(N, \alpha_s(t)) & 2n_f \gamma_{qg}^S(N, \alpha_s(t)) \\ \gamma_{gq}^S(N, \alpha_s(t)) & \gamma_{gg}^S(N, \alpha_s(t)) \end{pmatrix} \begin{pmatrix} \Sigma(N, Q^2) \\ g(N, Q^2) \end{pmatrix}.$$

- LOG SCALE $t = \ln \frac{Q^2}{\Lambda^2}$:
- ANOMALOUS DIMENSIONS VS. SPLITTING FUNCTIONS
  $\gamma(N, \alpha_s(t)) \equiv \int_0^1 dx\, x^{N-1} P(x, \alpha_s(t))$
- SINGLET $\Sigma(x, Q^2) = \sum_{i=1}^{n_f} \left(q_i(x, Q^2) + \bar{q}_i(x, Q^2) \right)$ VS.
  NONSINGLET $q^{NS}(x, Q^2) = q_i(x, Q^2) - q_j(x, Q^2)$
  COMBINATIONS OF QUARK PDFS
- PERTURBATIVE EXPANSION OF ANOMALOUS DIMENSION
  $\gamma_i(N, \alpha_s(t)) = \gamma_i^{(0)}(N) + \alpha_s(t) \gamma_i^{(1)}(N) + \ldots \Rightarrow$
  LOG RESUMMATION: LO $\Leftrightarrow$ LLQ^2; NLO $\Leftrightarrow$ LLQ^2, $\ldots$

SCALE DEPENDENCE:

ANOMALOUS DIMENSIONS (LEADING ORDER)



QUALITATIVE FEATURES

- AS Q^2 INCREASES, PDFS DECREASE AT LARGE x & INCREASE AT SMALL x DUE TO RADIATION
- GLUON SECTOR SINGULAR AT $N = 1 \Rightarrow$ GLUON GROWS MORE AT SMALL x
- $\gamma_{qq}(1) = 0 \Rightarrow$ NUMBER OF QUARKS CONSERVED

PARTON KINEMATICS VS. HADRON KINEMATICS

$$\sigma(\tau) = \int_\tau^1 \frac{dy}{y} \sum_{ij} \mathcal{L}_{ij}(y) \hat{\sigma}_{ij}\left(\frac{\tau}{z}\right); \quad \mathcal{L}_{ij}(y) \equiv \int_y^1 \frac{dx_1}{x_1} q_i(y) q_j\left(\frac{y}{x_1}\right)$$

- WHICH PARTON MOMENTUM FRACTIONS CONTRIBUTE TO A GIVEN HADRONIC PROCESS ?

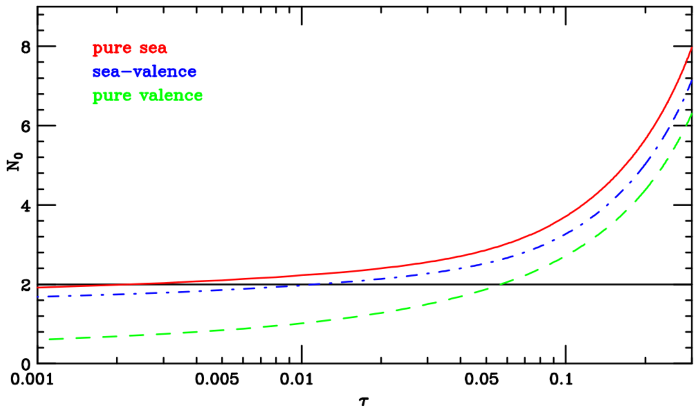
INVERSION OF MELLIN TRANSFORMS

$$f_n = \int_x^1 x^{n-1} f(x) \Leftrightarrow F(x) = \int_{-i\infty}^{+i\infty} x^{-n} f_n$$

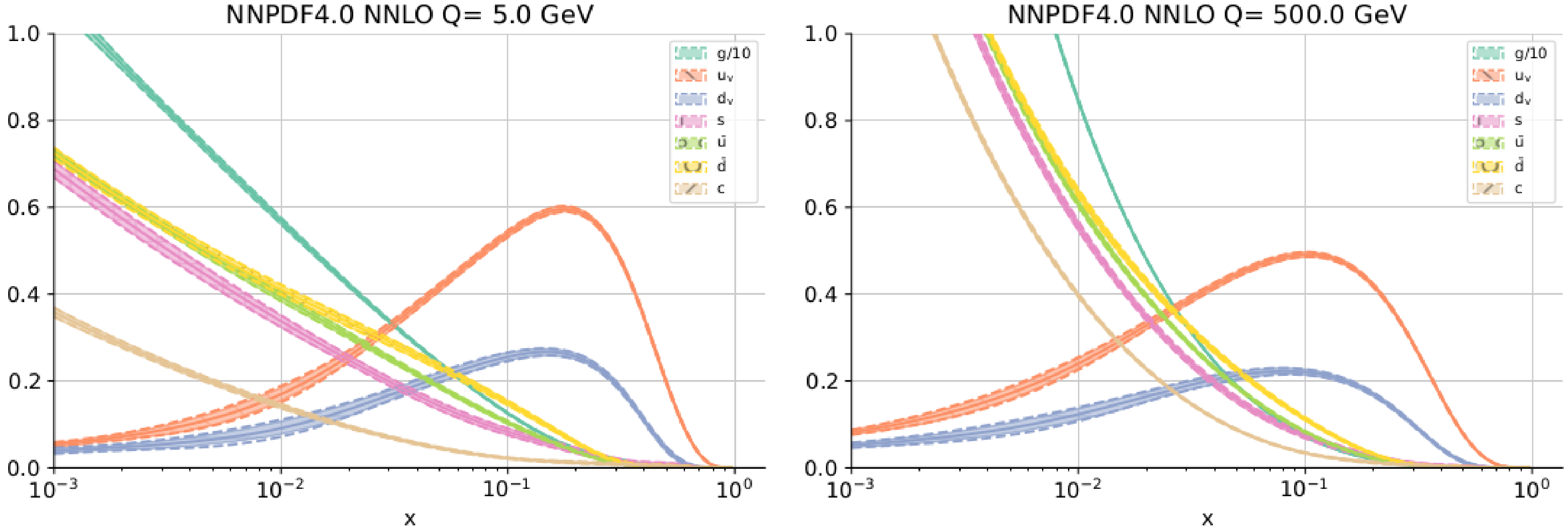
integrate to the right of convergence abscissa

- MELLIN INVERSION DOMINATED BY SADDLE POINT
- POSITION OF SADDLE DEPENDS ON HADRONIC KINEMATICS, CONTROLLED BY PARTON LUMINOSITY DEPENDENCE ON x OF $\mathcal{L}$ POWERLIKE, OF $\hat{\sigma}$ LOGARITHMIC
- PDF PEAKED AT SMALL x ("SEA" $\bar{q}$ VS. "VALENCE" $q - \bar{q}$) $\Rightarrow$ LUMI PEAKS AT SMALL N

SADDLE VS $\tau = Q^2/s$



PDFs: QUALITATIVE FEATURES



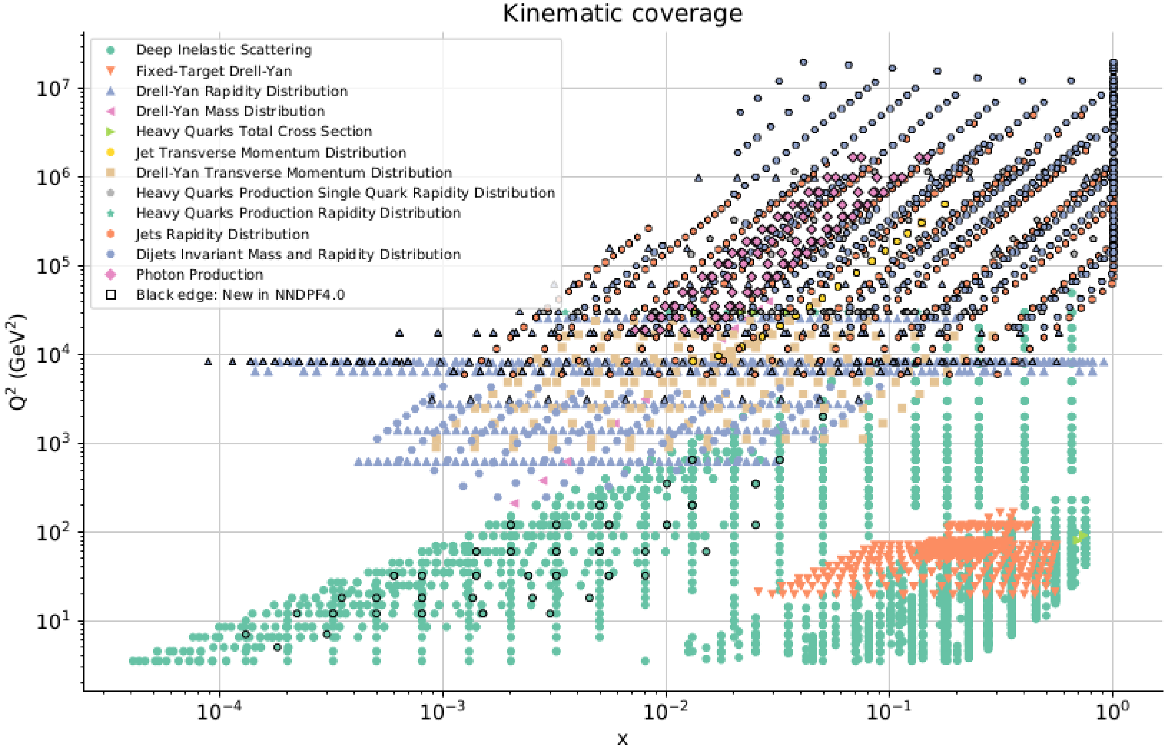
- THE MOMENTUM PROBABILITY DENSITY $xf_i(x)$ IS SHOWN AT TWO DIFFERENT SCALES (LEFT $\Rightarrow$ LOW SCALE; RIGHT $\Rightarrow$ HIGH SCALE)
- AS $x \geq 1$ KINEMATIC CONSTRAINT $f_i(x) = 0$
- "VALENCE" UP AND DOWN: PEAKED AT $x \sim 0.3$; EXPECT $f_x(x) \underset{x\to 1}{\sim} (1-x)_i^\beta$
- "SEA" ANTIQUARK AND GLUON GROW AT SMALL x
- "SINGLET" AND GLUON MIX $\Rightarrow$ ALL PDFs LOOK THE SAME AS $x \to 0$

PDF DETERMINATION

PDF DETERMINATION

DATA → PARTON DISTRIBUTIONS

Experimental data in NNPDF4.0



More than 4000 datapoints!

New processes:

- direct photon
- single top
- dijets
- W+jet
- DIS jet

ISSUES AND TASKS:

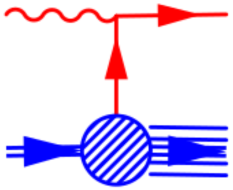
- FROM PHYSICAL OBSERVABLES TO PDFS: SOLVE EVOLUTION EQUATIONS, CONVOLUTE WITH PARTON-LEVEL CROSS-SECTIONS
- DISENTANGLING PDFS: CHOOSE A BASIS OF PDFS ($2N_f$ QUARKS + 1 GLUON) & A SET OF SUITABLE PHYSICAL PROCESSES TO DETERMINE THEM ALL
- PROBABILITY IN THE SPACE OF FUNCTIONS: CHOOSE A STATISTICAL APPROACH (MULTIGAUSSIAN, MONTE CARLO, . . .)
- UNCERTAINTY ON FUNCTIONS: CHOOSE A REGRESSION MODEL

DISENTANGLING PDFs

- CC F_1 AND F_3 IN PRINCIPLE PROVIDE FOUR COMBINATIONS, AND NC F_1 TWO MORE $\Rightarrow$ ALL LIGHT FLAVORS
- $W^\pm$ AND Z PRODUCTION (INCLUDING DOUBLE DIFFERENTIAL: MASS AND RAPIDITY) PROVIDE INDEPENDENT COMBINATIONS
- WHEN PRODUCING ELECTROWEAK FINAL STATES, THE GLUON CAN ONLY BE ACCESSED FROM SCALE DEPENDENCE OR HIGHER ORDERS (DIFFERENTIAL DISTRIBUTIONS)
  …EXCEPT IN HIGGS PRODUCTION!
- JET PRODUCTION GIVES A DIRECT HANDLE ON THE GLUON

FLAVOR SEPARATION (DIS & DY)
LEADING ORDER PARTON CONTENT

DEEP-INELASTIC SCATTERING



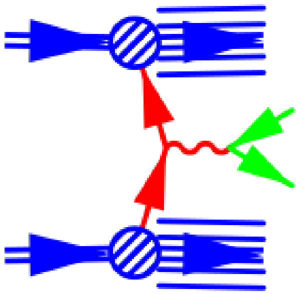
NC	${F_1}^{\gamma} = \sum_i e_i^2\,(q_i + \bar{q}_i)$
NC	${F_1}^{Z,\,\mathrm{int.}} = \sum_i B_i\,(q_i + \bar{q}_i)$
NC	${F_3}^{Z,\,\mathrm{int.}} = \sum_i D_i\,(q_i + \bar{q}_i)$
CC	$F_1^{W^+} = \bar{u} + d + s + \bar{c}$
CC	$-F_3^{W^+}/2 = \bar{u} - d - s + \bar{c}$

ℓ	e	V	A
u,c,t	$+2/3$	$(+1/2 - 4/3\sin^2\theta_W)$	$+1/2$
d,s,b	-1/3	$(-1/2 + 2/3\sin^2\theta_W)$	-1/2
ν	0	$+1/2$	$+1/2$
e,μ, τ	-1	$(-1/2 + 2\sin^2\theta_W)$	-1/2

$B_q(Q^2) = -2e_q V_\ell V_q P_Z + (V_\ell^2 + A_\ell^2)(V_q^2 + A_q^2)P_Z^2;\ D_q(Q^2) = -2e_q A_\ell A_q P_Z + 4V_\ell A_\ell V_q A_q P_Z^2;\ P_Z = Q^2/(Q^2 + M_Z^2)$

$$W^+ \to W^- \Rightarrow u \leftrightarrow d,\ c \leftrightarrow s;\ p \to n \Rightarrow u \leftrightarrow d$$

DRELL-YAN



$$L^{ij}(x_1, x_2) \equiv q_i(x_1, M^2)\bar{q}_j(x_2, M^2)$$

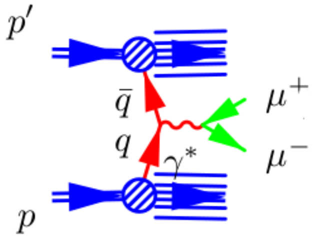
γ	$\frac{d\sigma}{dM^2 dy}(M^2, y) = \frac{4\pi\alpha^2}{9M^2 s}\sum_i e_i^2 L^{ii}(x_1, x_2)$
W	$\frac{d\sigma}{dy} = \frac{\pi G_F M_V^2 \sqrt{2}}{3s}\sum_{i,j}\lvert V_{ij}^{\mathrm{CKM}}\rvert L^{ij}(x_1, x_2)$
Z	$\frac{d\sigma}{dy} = \frac{\pi G_F M_V^2 \sqrt{2}}{3s}\sum_i \left(V_i^2 + A_i^2\right) L^{ij}(x_1, x_2)$

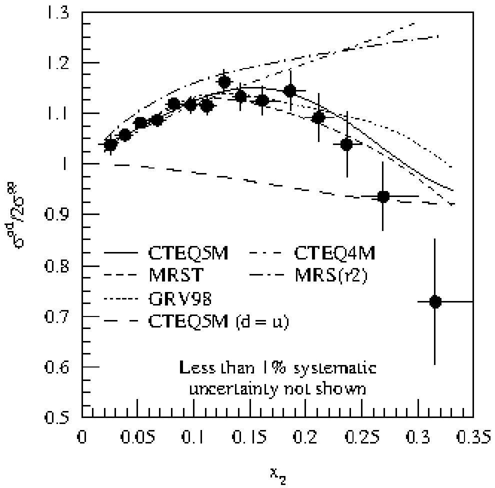
$V_{ij}^{\mathrm{CKM}} \to$ CKM MATRIX ($i = u,\ c\,t,\ j = d,\ s\,b$), $V_{ij}^{\mathrm{CKM}} = 1 + O(\lambda)$; $\lambda = \sin\theta_C \approx 0.22$

FIXED-TARGET DRELL-YAN (TEVATRON)
QUARKS AND ANTIQUARK SEPARATION

BY CHARGE CONJUGATION $\bar{q}_{\bar{P}} = q_p$

DRELL-YAN p/d ASYMMETRY

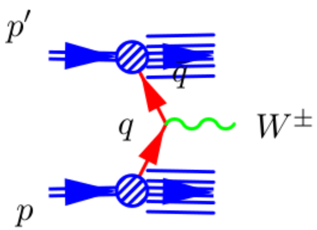


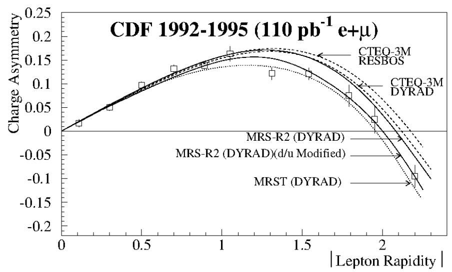


$$\frac{\sigma^{pn}}{\sigma^{pp}} \sim \left.\frac{\frac{4}{9}u^p\bar{d}^p+\frac{1}{9}d^p\bar{u}^p}{\frac{4}{9}u^p\bar{u}^p+\frac{1}{9}d^p\bar{d}^p}\right|_{\text{large } x} \approx \frac{\bar{d}}{\bar{u}}$$

E866 (2001)

$W^{\pm}$ ASYMMETRY





$$\frac{\sigma^{p\bar{p}}_{W+}}{\sigma^{p\bar{p}}_{W-}} = \frac{u^p(x_1)d^p(x_2)+\bar{d}^p(x_1)\bar{u}^p(x_2)}{d^p(x_1)u^p(x_2)+\bar{u}^p(x_1)\bar{d}^p(x_2)} \sim \frac{u^p d^p}{d^p u^p}$$

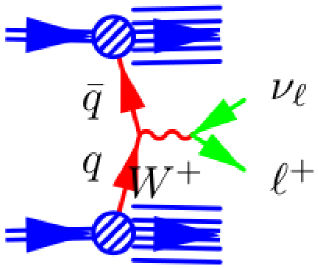
if x_1, x_2 in valence region,
neglecting HQ & Cabibbo suppr.

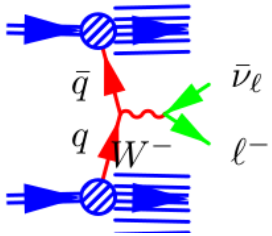
CDF (1998)

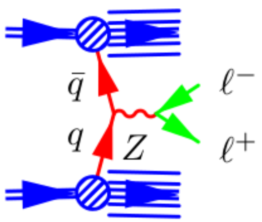
COLLIDER DRELL-YAN (LHC)

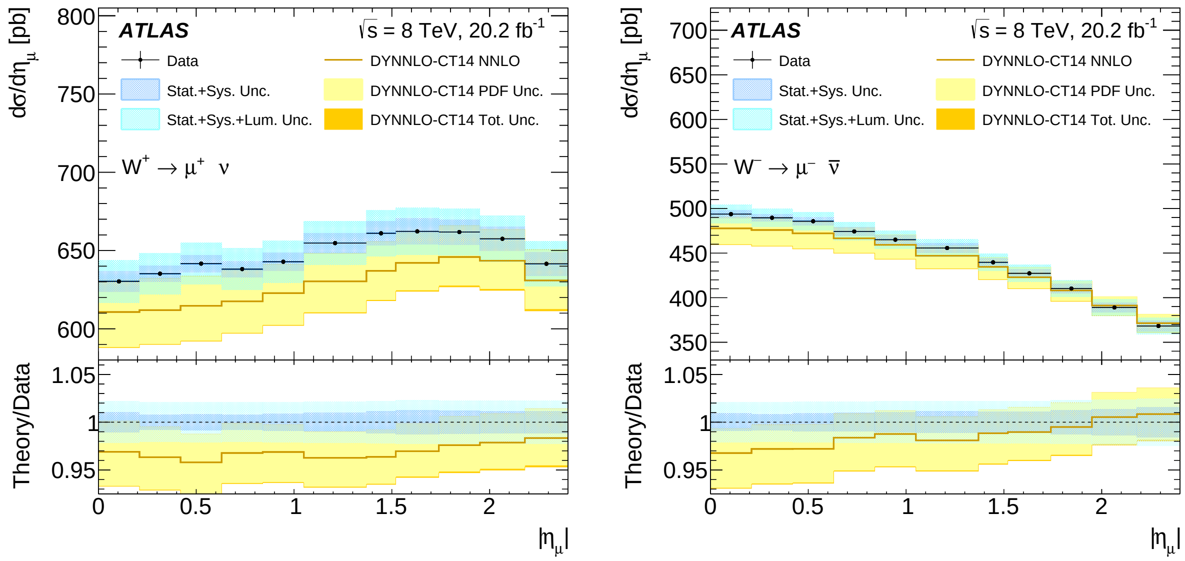
$W^{\pm}$ AND Z PRODUCTION

W AND Z CROSS SECTIONS





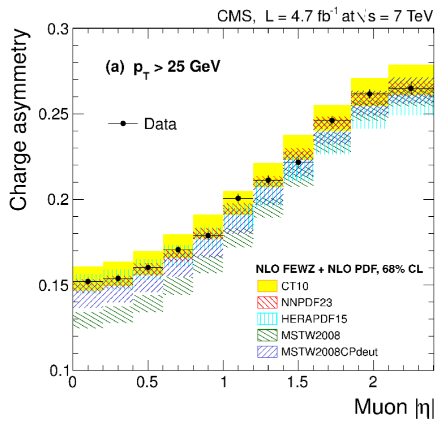




ATLAS (2019)

$\sigma^{p\bar{p}}_{W^+} = u\bar{d} + c\bar{s}$; $\sigma^{p\bar{p}}_{Z} = u\bar{u} + d\bar{d} + s\bar{s}$: STRANGENESS DETERMINED BY COMPARISON

W MUON ASYMMETRY



CMS (2013)

$$\frac{\sigma^{p\bar{p}}_{W^+}}{\sigma^{p\bar{p}}_{W^-}} = \frac{u(x_1)\bar{d}(x_2)+\bar{d}(x_1)u(x_2)}{d(x_1)\bar{u}(x_2)+\bar{u}(x_1)d(x_2)}$$

"VALENCE" $x \Rightarrow$ NEGLECT STRANGENESS

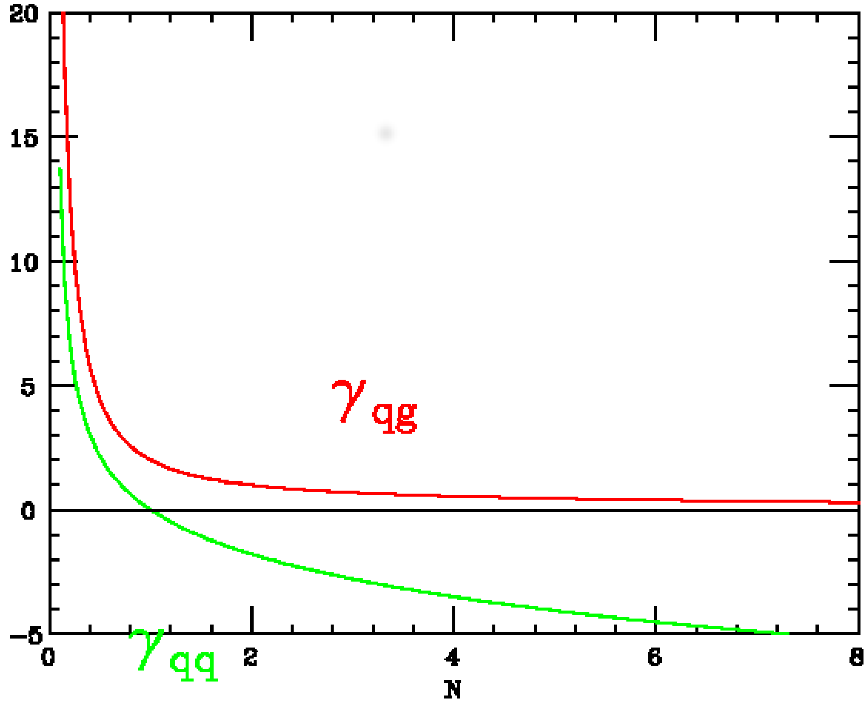
$\Rightarrow$ DETERMINE $\bar{u} - \bar{d}$

THE GLUON FROM DIS

SCALE DEPENDENCE OF FLAVOR SINGLET STRUCTURE FUNCTIONS

$$\frac{d}{dt}F_2^s(N,Q^2) = \frac{\alpha_s(Q^2)}{2\pi}\left[\gamma_{qq}(N)F_2^s + 2\,n_f\gamma_{qg}(N)g(N,Q^2)\right] + O(\alpha_s^2)$$

ANOMALOUS DIMENSIONS



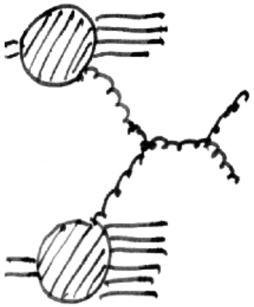
LARGE x GLUON DIFFICULT TO DETERMINE FROM DEEP-INELASTIC SCATTERING

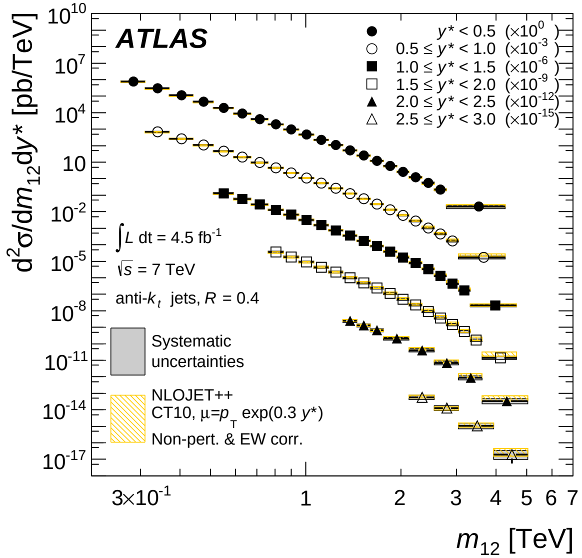
THE GLUON IN HADRONIC COLLISIONS

THE GLUON ONLY INTERACTS THROUGH QCD

JETS

GLUON



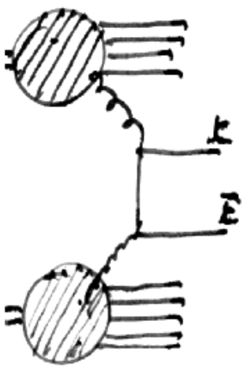


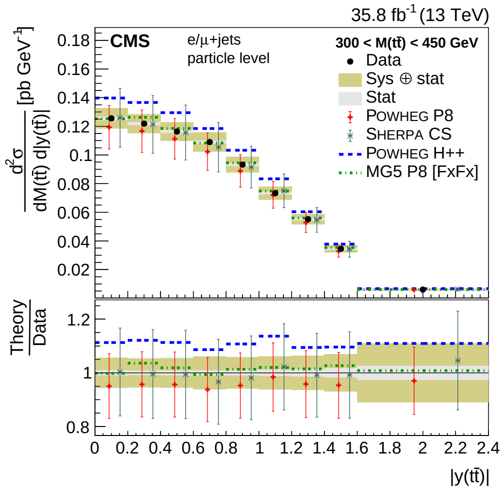
- ONE-JET/DIJET INCLUSIVE USED TO CONSTRAIN THE LARGE x GLUON SINCE TEVATRON
- WIDE KINEMATIC REGION AT LHC

ATLAS (2014)

TOP

GLUON





- WIDE RAPIDITY RANGE: CAN ACCESS WIDE x REGION

CMS (2018)

UNCERTAINTIES

DATA UNCERTAINTIES: COVARIANCE MATRIX

PREDICTIONS VS. DATA: LOSS

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

THE COVARIANCE MATRIX

$$\text{cov}_{ij} = \delta_{ij} s_i^2 + \sum_{\alpha=1}^{N_c} \sigma_{i,\alpha}^{(c)} \sigma_{j,\alpha}^{(c)} + \left(\sum_{\alpha=1}^{N_{\mathcal{L}}} \sigma_{i,\alpha}^{(\mathcal{L})} \sigma_{j,\alpha}^{(\mathcal{L})} \right) D_i D_j$$

- D_i: DATA; T_i: PREDICTION
- s_i: UNCORRELATED STATISTICAL UNCERTAINTY FOR i-TH DATAPOINT
- $\sigma_{i,\alpha}^{(c)}$: α-TH CORRELATED ADDITIVE SYSTEMATICS FOR i-TH DATAPOINT
- $\sigma_{i,\alpha}^{(\mathcal{L})}$: α-TH CORRELATED MULTIPLICATIVE SYSTEMATICS FOR i-TH DATAPOINT

DATA UNCERTAINTIES: NUISANCE PARAMETERS

THE PARAMETERS

$$\chi^2(\{a\},\{\lambda\}) = \sum_{k=1}^{N_{\rm pt}} \frac{1}{s_k^2}\left(D_k - T_k - \sum_{\alpha=1}^{N_\lambda}\beta_{k,\alpha}\lambda_\alpha\right)^2 + \sum_{\alpha=1}^{N_\lambda}\lambda_\alpha^2$$

SYSTEMATIC UNCERTAINTIES

$\beta_{i,\alpha} = \sigma_{i,\alpha}^{(c)}$ for $\alpha = 1,\ldots,N_c$; $\beta_{i,\alpha} = \sigma_{j,\alpha}^{(\mathcal{L})}D_i$ for $\alpha = N_c+1,\ldots,N_{\mathcal{L}}$

BEST-FIT VALUES

$$\lambda_{0\alpha} = \sum_{i=1}^{N_{\rm pt}} \frac{D_i - T_i}{s_i}\sum_{\delta=1}^{N_\lambda}\mathcal{A}^{-1}_{\alpha\delta}\frac{\beta_{i,\delta}}{s_i}$$

REDUCED COVARIANCE MATRIX

$$\mathcal{A}_{\alpha\beta} = \delta_{\alpha\beta} + \sum_{k=1}^{N_{\rm pt}}\frac{\beta_{k,\alpha}\beta_{k,\beta}}{s_k^2}$$

CONSTRUCTION OF THE COVARIANCE MATRIX: INVERSE

$$({\rm cov})^{-1}_{ij} = \left[\frac{\delta_{ij}}{s_i^2} - \sum_{\alpha,\beta=1}^{N_\lambda}\frac{\beta_{i,\alpha}}{s_i^2}\mathcal{A}^{-1}_{\alpha\beta}\frac{\beta_{j,\beta}}{s_j^2}\right],$$

THE COVARIANCE MATRIX

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

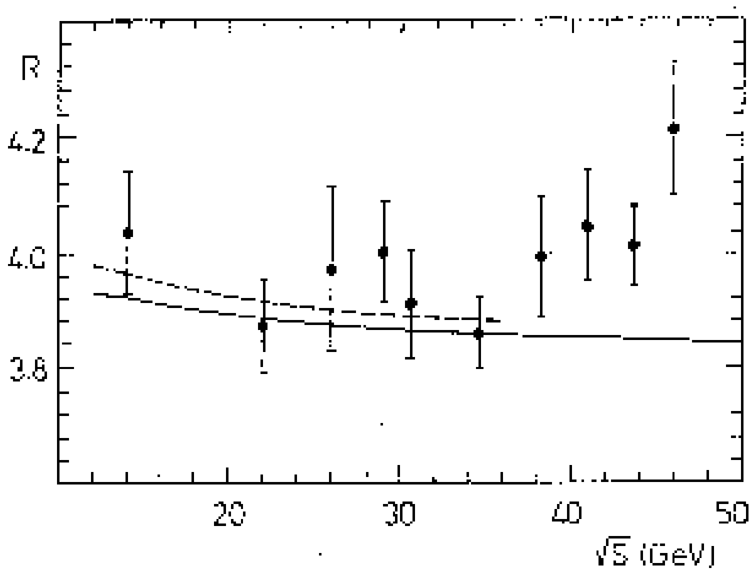
MULTIPLICATIVE UNCERTAINTIES

THE D'AGOSTINI BIAS

- NORMALIZATION UNCERTAINTIES IN COVARIANCE MATRIX
  $(\mathrm{cov})_{ij} = \sigma_{i,n}\sigma_{j,n} D_i D_j$
  $\Rightarrow$ MAXIMUM-LIKELIHOOD RESULT BIASED (d'Agostini, 1994)

- EQUIVALENT TO RESCALING DATA BUT NOT UNCERTAINTIES

$$R = \frac{e^+e^- \to \mathrm{hadrons}}{e^+e^- \to \mu^+\mu^-}$$

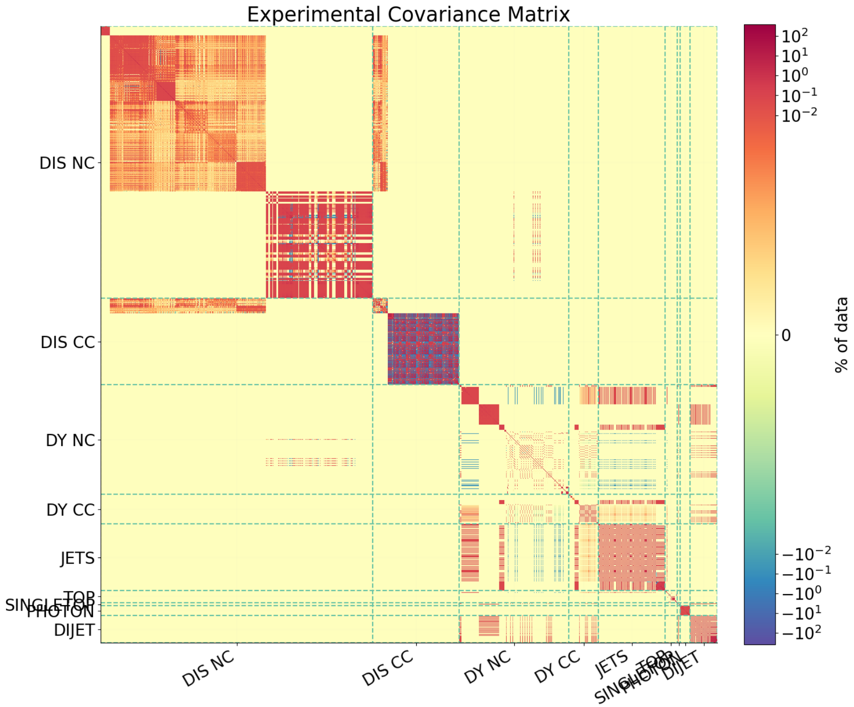


(CELLO collab., 1987)

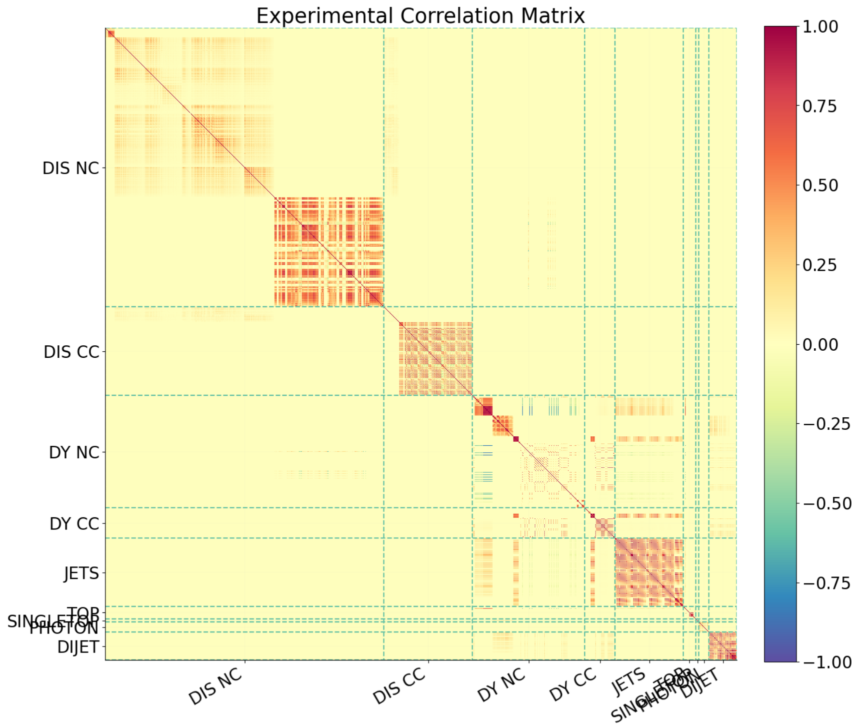
- MUST COMPUTE UNCERTAINTY FROM PREVIOUS THEORY PREDICTION
  RESULT OF PREVIOUS FIT: $(\mathrm{cov})_{ij} = \sigma_{i,n}\sigma_{j,n} T_i^{(0)} T_j^{(0)}$

A LOOK AT THE COVARIANCE MATRIX

COVARIANCE



CORRELATION



MISSING HIGHER ORDER (THEORY) UNCERTAINTIES

- PDFs ARE DETERMINED BY MAXIMIZING THE LIKELIHOOD

$$P = N \exp - \left(\frac{d-t}{2\sigma^2_{exp}} \right)$$

d, t ARE REALLY VECTORS AND $1/\sigma^2$ THE INVERSE COVARIANCE MATRIX

- PROBABILITY OF THE THEORY t BEING CORRECT GIVEN DATA d, WHICH BY BAYES IS

$$P(t|d) \propto P(d|t)P(t)$$

- IF THEORY WAS KNOWN EXACTLY, THEN $P(t) = \delta(t - t^{\text{exact}})$
- IN ACTUAL FACT ONLY SOME PERTURBATIVE RESULT t_p IS EXACTLY KNOWN SO $t^{\text{exact}} = t_p + \Delta_p$, WHERE Δ_p INCLUDES MHO
- ASSUMING Δ TO BE GAUSSIANLY DISTRIBUTED, WITH UNCERTAINTY σ_{th} AND INTEGRATING OUT

$$P = N \exp \left[\frac{d-t_p}{2\left(\sigma^2_{exp} + \sigma^2_{th}\right)} \right]$$

- THEORETICAL UNCERTAINTY ADDED IN QUADRATURE, PROPAGATES INTO PDF UNCERTAINTY UPON MINIMZATION
- CAN COMPUTE PDF UNCERTAINTY GIVEN MHOU

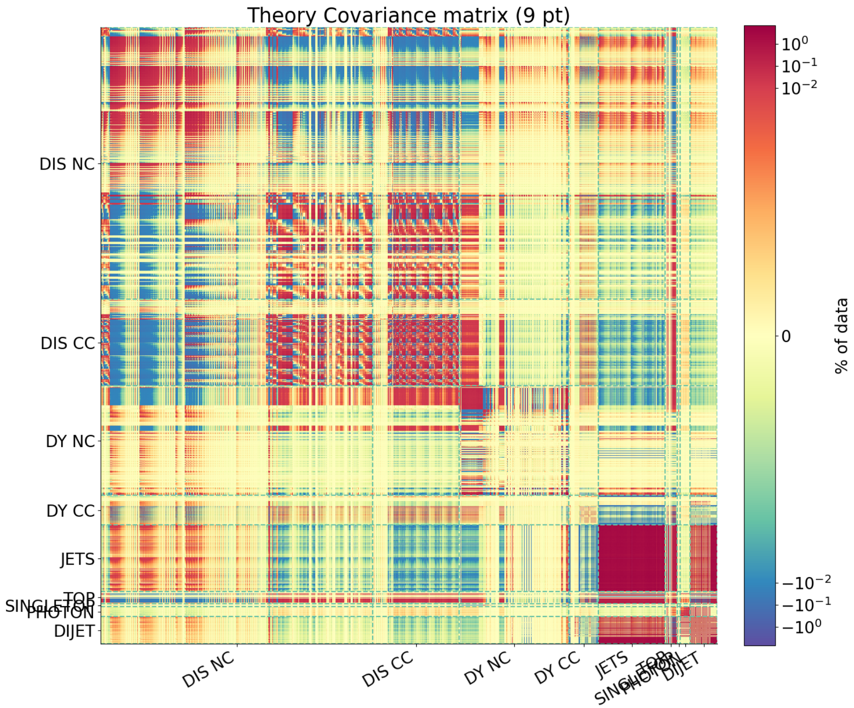
THEORY COVARIANCE MATRICES

THEORY COVARIANCE MATRIX: processes i, j, scale choice $\{\mu^{(k)}\}$, default $\{\mu_0\}$

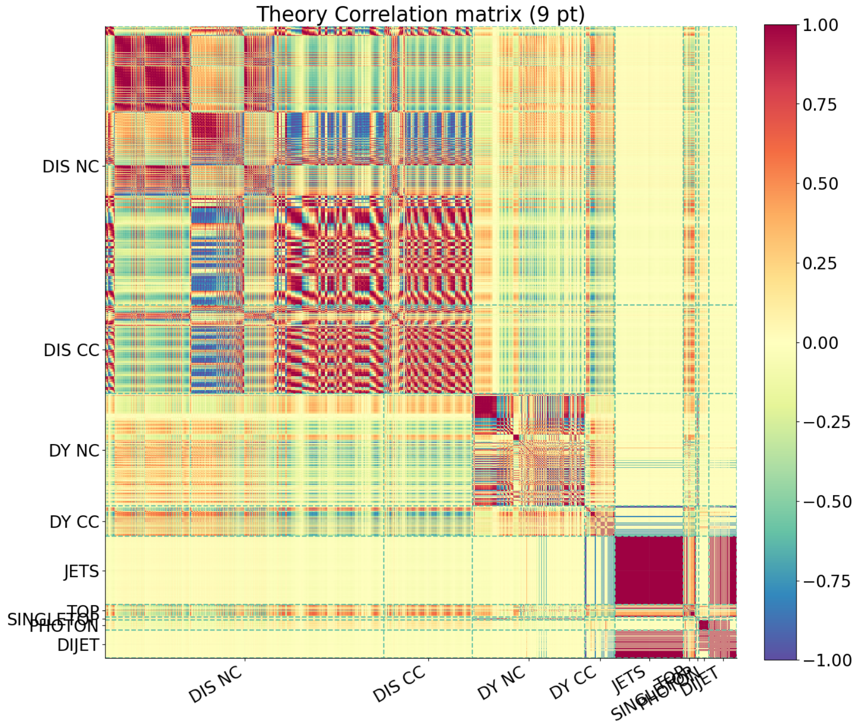
$$\sigma_{i,j} = \frac{1}{N} \sum_k \left(\sigma_i[\{\mu^{(k)}\}] - \sigma_i[\{\mu_0\}] \right) \left(\sigma_j[\{\mu^{(k)}\}] - \sigma_j[\{\mu_0\}] \right)$$

- SINGLE PROCESS: k RUNS OVER COMMON SET OF SCALE CHOICES
- MANY PROCESSES:
  - UNCORRELATED RENORMALIZATION: DIFFERENT FOR DIFFERENT HARD PROCESSES
  - CORRELATED FACTORIZATION: MHOU OF PERTURBATIVE EVOLUTION UNIVERSAL

COVARIANCE



CORRELATION



MODEL-DEPENDENT REGRESSION

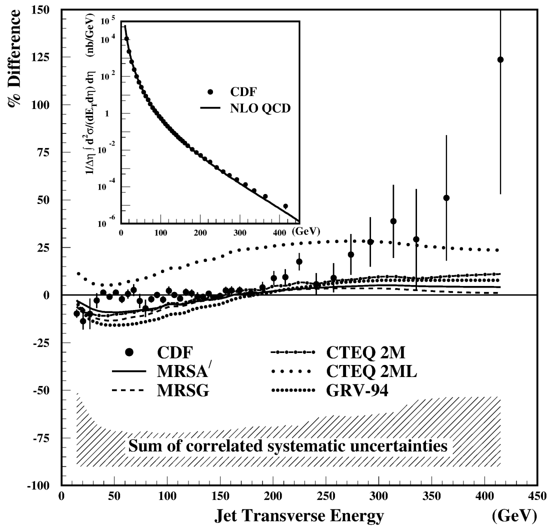
A MODEL-DEPENDENT (HESSIAN) APPROACH

- CHOOSE A FIXED FUNCTIONAL FORM
  - SINCE 1973, PHYSICALLY MOTIVATED ANSATZ $f_i(x, Q_0^2) = x^\alpha (1-x)^\beta g_i(x)$;
    $g_i(x)$ POLYNOMIAL IN x OR $\sqrt{x}$
  - MMHT 2015:
    * BASIS FUNCTIONS g; $u_v = u - \bar{u}$; $d_v = d - \bar{d}$; $S = 2(\bar{u} + \bar{d}) + s + \bar{s}$; $s_+ = s + \bar{s}$; $\Delta = \bar{d} - \bar{u}$;
      $s_- = s - \bar{s}$.
    * FOR ALL BUT Δ s_-, $g \Rightarrow xf_i(x, Q_0^2) = Ax^\alpha(1-x)^\beta \left(1 + \sum_{i=1}^4 a_i T_i(y(x))\right)$;
      T_i CHEBYSHEV POLYNOMIALS, $y = 1 - 2\sqrt{x} \leftrightarrow$ MUST MAP $x = [0,1]$ INTO $y = [-1,1]$;
      $T_i(-1) = T_i(1) = 1$
    * GLUON $xg(x, Q_0^2) = Ax^\alpha(1-x)^\beta \left(1 + \sum_{i=1}^2 a_i T_i(y(x))\right) + A' xT\alpha'(1-x)^{\beta'}$
    * SEA ASYMMETRY $x\Delta(x, Q_0^2) = Ax^\alpha(1-x)^\beta(1 + \gamma x + \epsilon x^2)$
    * STRANGENESS ASYMMETRY $x\Delta(x, Q_0^2) = Ax^\alpha(1-x)^\beta(1 - x/x_0)$
    * 41 PARAMETERS, 4 FIXED BY SUM RULES
    * 12 PARMS FIXED AT BEST FIT, REMAINING 25 USED FOR COVARIANCE MATRIX
      $\Rightarrow$ INCREASED TO 30 IN MSHT 2019
- EVOLVE TO DESIRED SCALE & COMPUTE PHYSICAL OBSERVABLES
- DETERMINE BEST-FIT VALUES OF PARAMETERS
- DETERMINE ERROR BY PROPAGATION OF ERROR ON PARMS. $\Delta\chi^2 = 1$('HESSIAN METHOD');
  PARM. SCANS ALSO POSSIBLE ('LAGR. MULTIPLIER METHOD')

THE PROBLEM OF MODEL DEPENDENCE:

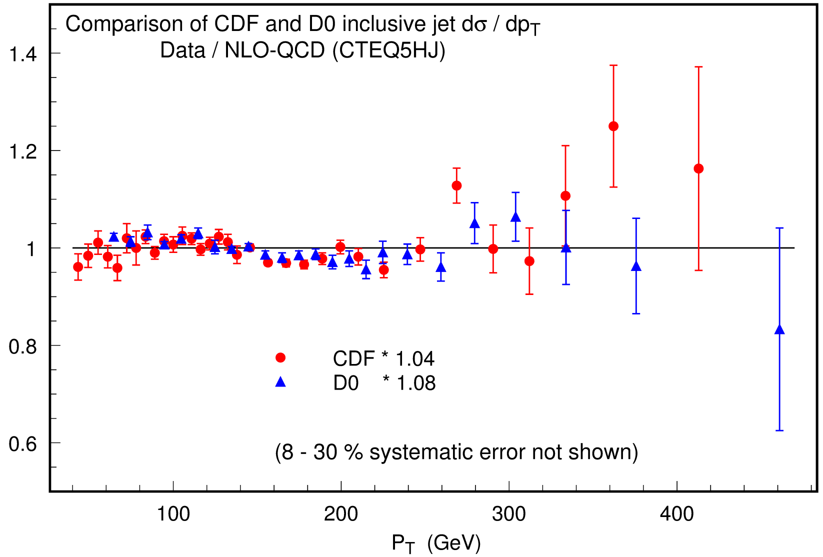
A DISCOVERY THAT WASN'T

- DISCREPANCY BETWEEN QCD CALCULATION AND CDF JET DATA (1995)
- EVIDENCE FOR QUARK COMPOSITENESS?
- RESULT STRONGLY DEPENDS ON GLUON AT $x \gtrsim 0.1$
- PDF MUST VANISH AT $x = 0$, BUT (THEN) NO DATA FOR $x \gtrsim 0.05$!



DISCREPANCY REMOVED IF JET DATA USED FOR GLUON DETERMINATION

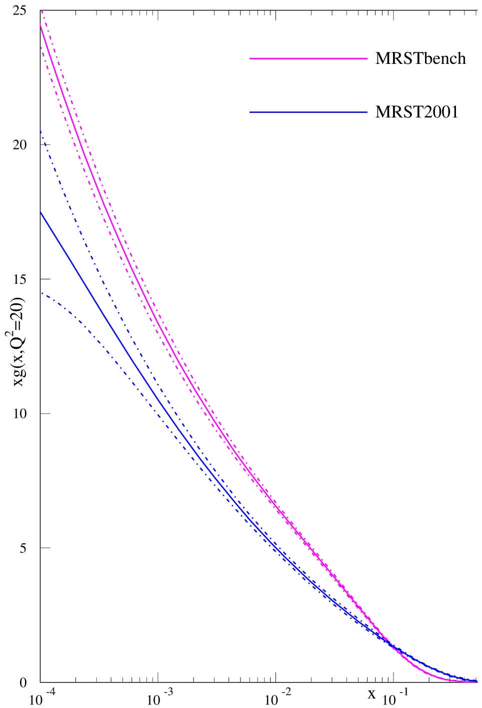
NEW CTEQ GLUON (1998)



MODEL-DEPENDENT UNCERTAINTIES
THE HERA-LHC BENCHMARK PUZZLE

- RESTRICTED AND VERY CONSISTENT DATASET USED
- RESULTS COMPARED TO THEN-BEST RESULT FROM FULL DATASET

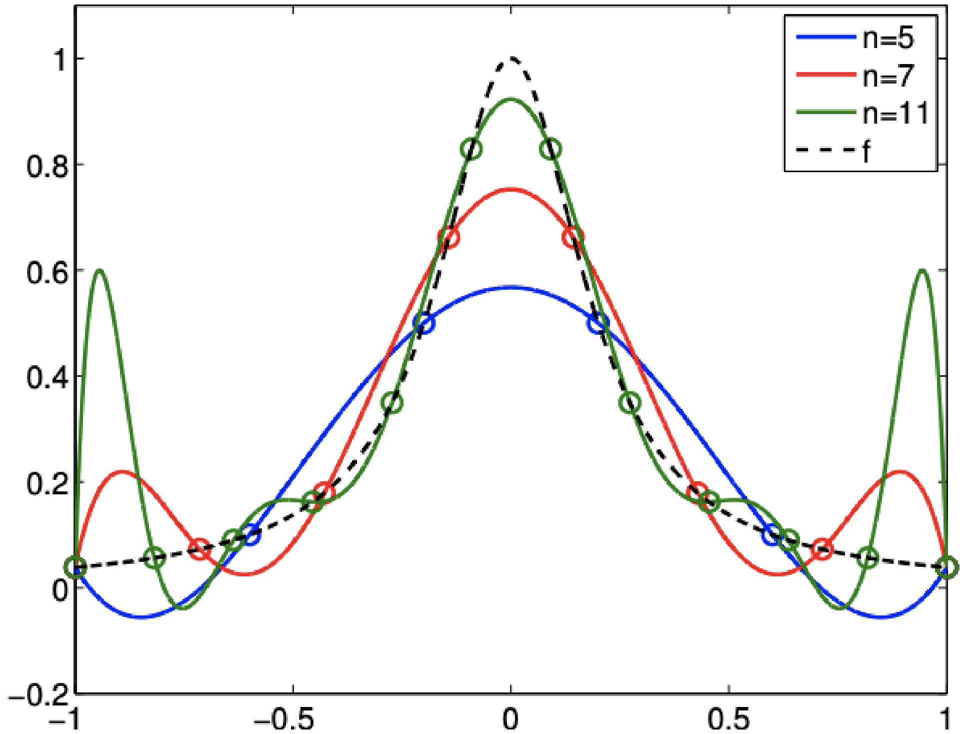
BENCHMARK VS DEFAULT GLUON



"...the partons extracted using a very limited data set are completely incompatible, even allowing for the uncertainties, with those obtained from a global fit with an identical treatment of errors...The comparison illustrates the problems in determining the true uncertainty on parton distributions." (R.Thorne, HERALHC, 2005)

REGRESSION THROUGH POLYNOMIALS

THE RUNGE PHENOMENON

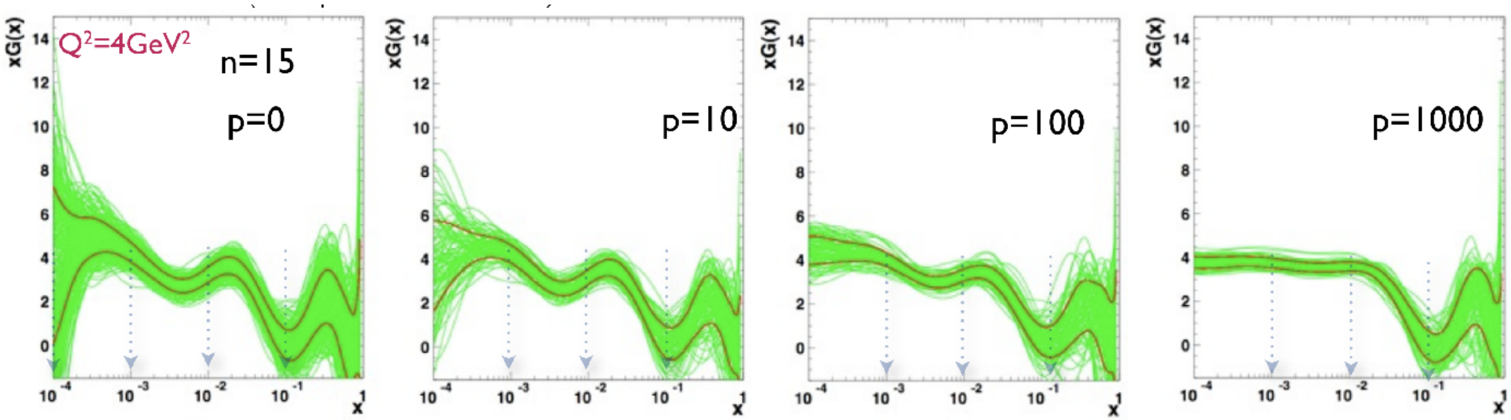


- INTERPOLATE WITH A POLYNOMIAL A FUNCTION SAMPLED AT EQUALLY SPACED POINTS
- RUNGE/AGNESI FUNCTION: $\frac{1}{1+x^2}$ SAMPLED AT $x_i = \frac{2i}{n} - 1,\ i = 1, \ldots, n$
- OSCILLATIONS INCREASE AS THE DEGREE OF THE POLYNOMIAL INCREASES
- ALLEVIATED WITH SUITABLE CHOICE OF POLYNOMIALS AND SAMPLING POINTS (CHEBYSHEV NODES)

REGRESSION THROUGH POLYNOMIALS

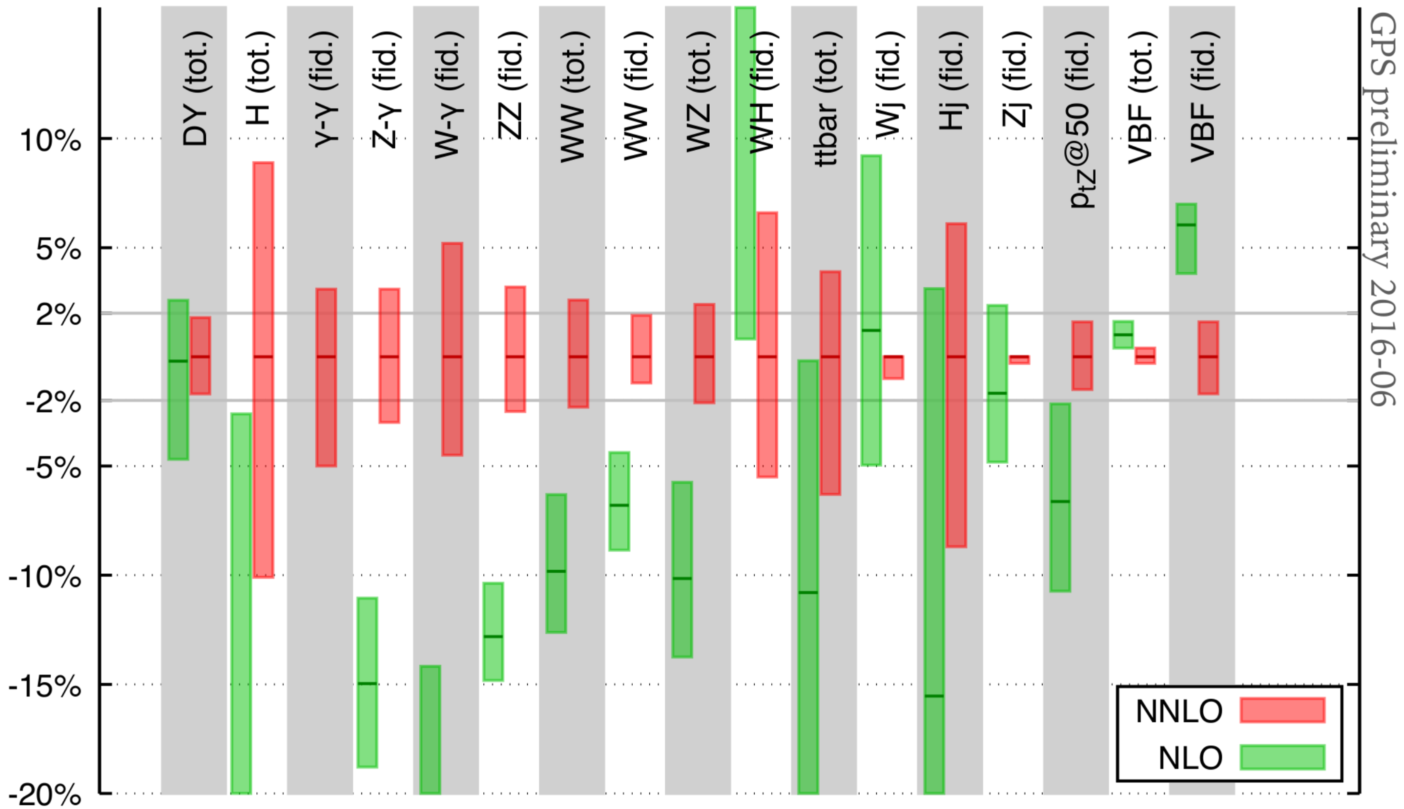
CHEBYSHEV AND LENGTH PENALTY

- OLD IDEA FOR PDF MODELING (PARISI, SOURLAS, 1978):
  EXPAND PDFS OVER BASIS OF ORTHOGONAL POLYNOMIALS
- GLAZOV, RADESCU, 2009: SYSTEMATIC MONTE CARLO APPROACH
- LENGTH PENALTY STABILIZATION:
  CONTRIBUTION TO χ^2 PROPORTIONAL TO THE ARCLENGTH WITH WEIGHT p
- RESULTS STRONGLY DEPENDENT ON ARBITRARY CHOICE OF p



EXTRAS

VALIDATION OF THEORY UNCERTAINTIES
NLO SCALE VARIATION VS ACTUAL NNLO CORRECTION



(G. Salam, 2016)

- SCALE VARIATION CANNOT PREDICT OPENING OF NEW CHANNELS OR RESUMMATION
- ISSUE KNOWN IN MOST CASES

VALIDATION OF THEORY UNCERTAINTIES

NLO SCALE VARIATION VS ACTUAL NNLO CORRECTION

OBSERVABLES INCLUDED IN NNPDF4.0 PDF DETERMINATION