NNPDF
STUDIES ON PDF UNCERTAINTIES

STEFANO FORTE

MILAN UNIVERSITY & INFN

FOR THE COLLABORATION: R. D. BALL, L. DEL DEBBIO,

S.F., A. GUFFANTI, J. I. LATORRE, J. ROJO, M. UBIALI

UNIVERSITÀ DEGLI STUDI DI MILANO

DIPARTIMENTO DI FISICA

PDF4LHC WORKSHOP

DESY, OCTOBER 23, 2009

SOME QUESTIONS:

- ARE EXPERIMENTAL UNCERTAINTIES SIZABLY UNDERESTIMATED?

  ARE THERE SIGNIFICANT DATA INCOMPATIBILITIES?

- WHERE DOES THE UNCERTAINTY ON PDFS COME FROM?

  IS IT RELATED TO PARTON PARAMETRIZATION?

- DOES THE TREATMENT OF CORRELATED UNCERTAINTIES HAVE AN IMPACT?

SOME QUESTIONS:

- ARE EXPERIMENTAL UNCERTAINTIES SIZABLY UNDERESTIMATED?

  ARE THERE SIGNIFICANT DATA INCOMPATIBILITIES?

- WHERE DOES THE UNCERTAINTY ON PDFS COME FROM?

  IS IT RELATED TO PARTON PARAMETRIZATION?

- DOES THE TREATMENT OF CORRELATED UNCERTAINTIES HAVE AN IMPACT?

WILL BE ADDRESSED USING THE NNPDF METHODOLOGY;

ALL STUDIES BASED ON PUBLISHED NNPDF1.2 FIT

RELEVANT NNPDF FEATURES

A REMINDER

MONTE CARLO

- PDFs ARE FITTED TO DATA REPLICAS
- REPLICAS FLUCTUATE ABOUT CENTRAL DATA:

$$F_{i,p}^{(art)(k)} = S_{p,N}^{(k)} F_{i,p}^{\exp} \left(1 + r_p^{(k)} \sigma_p^{\rm stat} + \sum_{j=1}^{N_{\rm sys}} r_{p,j}^{(k)} \sigma_{p,j}^{\rm sys}\right)$$

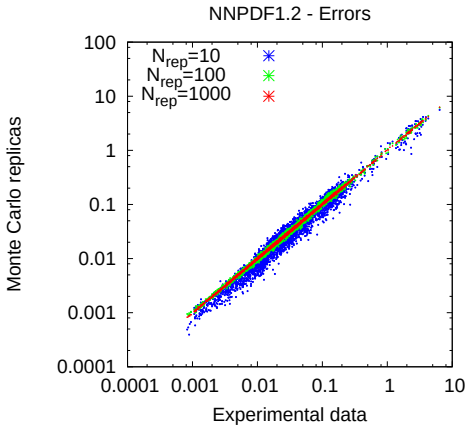
- SIZE OF FLUCTUATION $\leftrightarrow$ DATA UNCERTAINTY

  SAME AS FLUCTUATION OF CENTRAL DATA

  ABOUT "TRUE" VALUE

REPLICA STANDARD DEV.

VS. UNCERTAINTIES

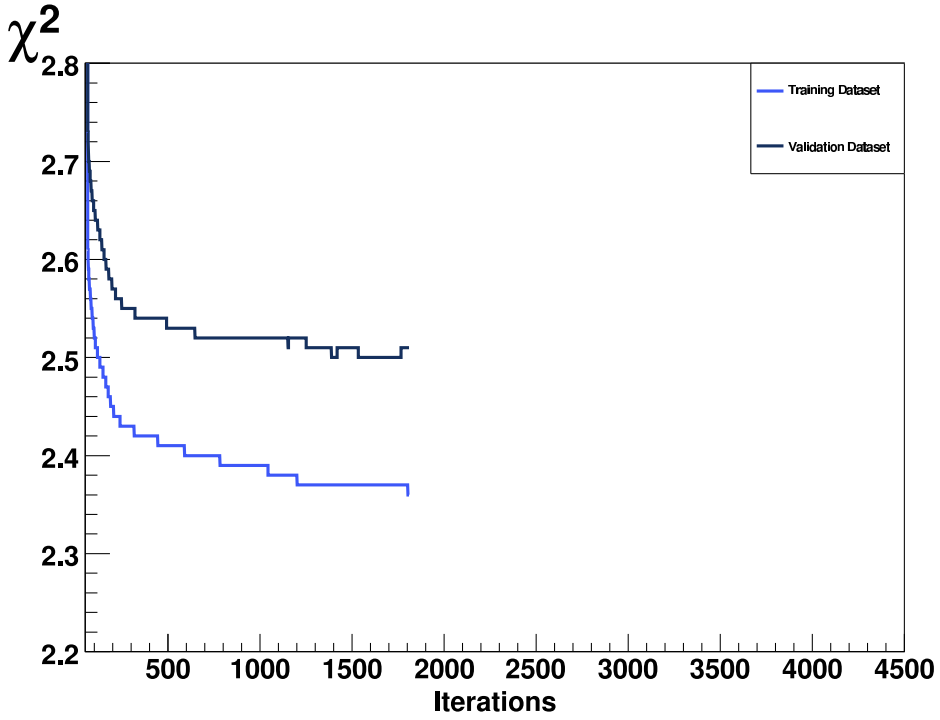


RELEVANT NNPDF FEATURES II
CROSS-VALIDATION

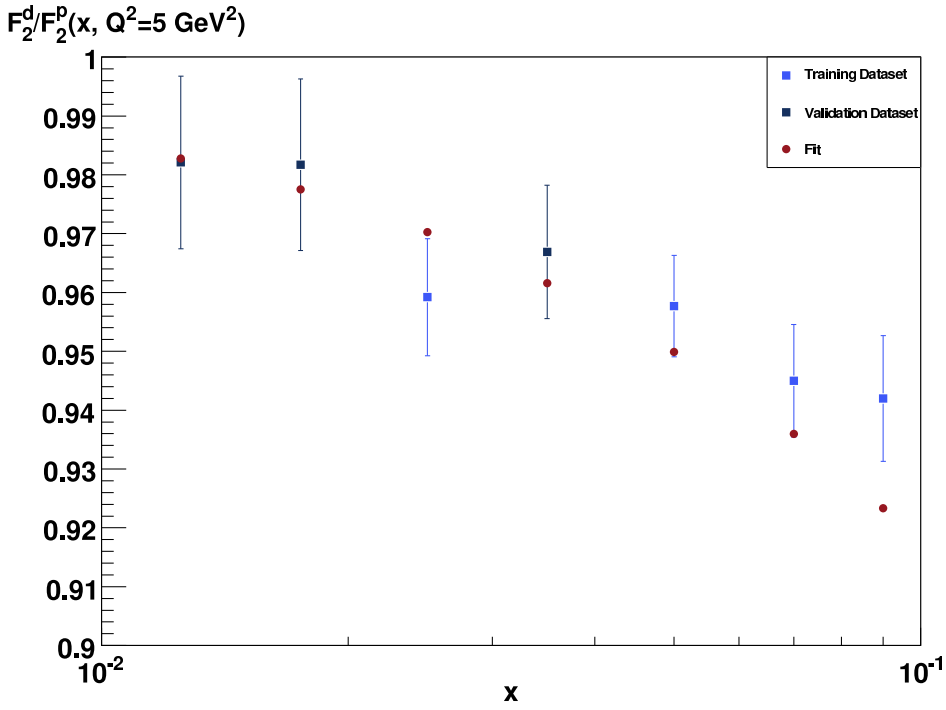
- REPLICAS ARE FITTED TO A DATA SUBSET
- A DIFFERENT SUBSET OF DATA USE FOR EACH REPLICA
-

OPTIMAL FITTING

χ^2



FIT TO DATA

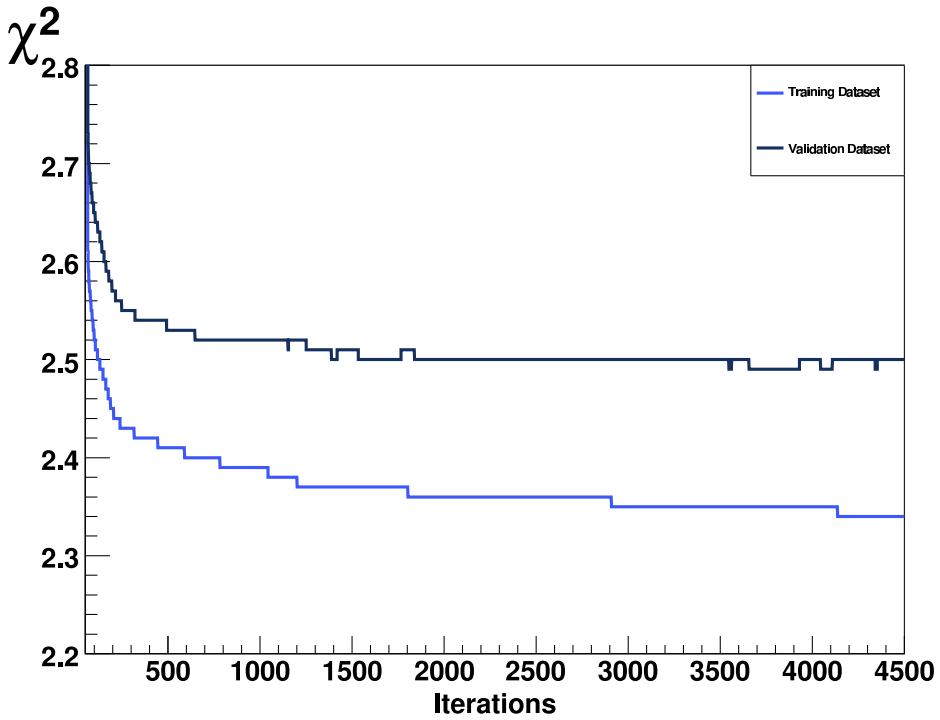


RELEVANT NNPDF FEATURES II
CROSS-VALIDATION

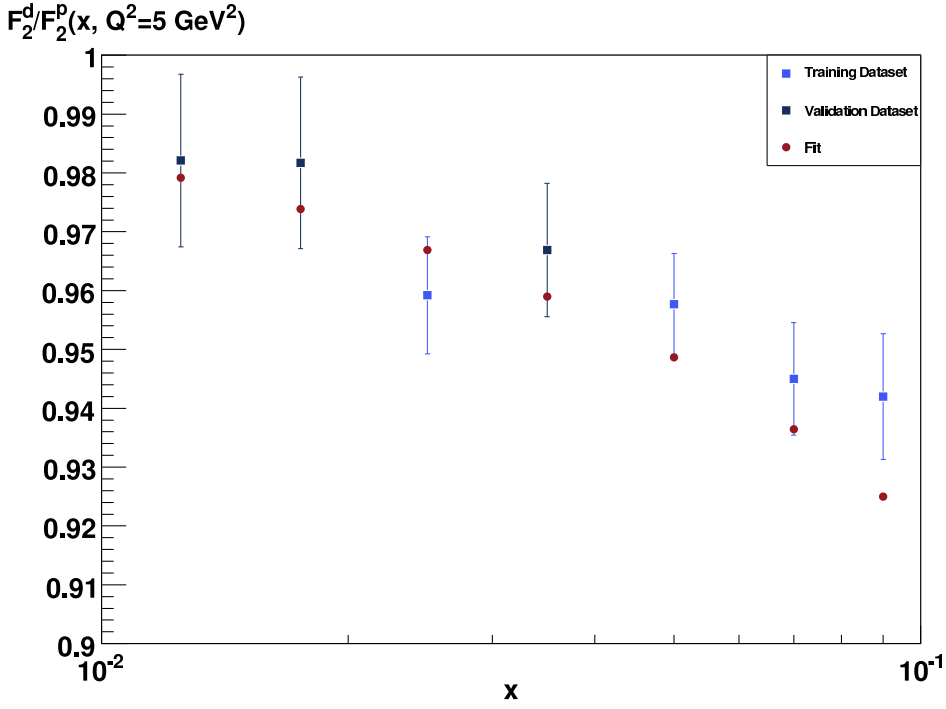
- REPLICAS ARE FITTED TO A DATA SUBSET
- A DIFFERENT SUBSET OF DATA USE FOR EACH REPLICA
- THE BEST FIT IS NOT AT THE MINIMUM OF THE χ^2

OVERFITTING

χ^2



FIT TO DATA



IDEAS

Thanks to J. Pumplin

- FIT TO REPLICAS VS. FIT TO DATA PARTITIONS $\Leftrightarrow$
  $\Leftrightarrow$FLUCTUATION OF DATA (TRUE) VS. FLUCTUATION OF REPLICAS (NOMINAL)
- FIT TO PARTITIONS VS. FIT TO A SINGLE PARTITION $\Leftrightarrow$
  $\Leftrightarrow$ UNCERTAINTY DUE TO DATA VS. UNCERTAINTY DUE TO OTHER SOURCES
- OPTIMAL FIT VS. OVERLEARNING FIT $\Leftrightarrow$
  $\Leftrightarrow$ UNDERLYING LAW VS. STATISTICAL NOISE

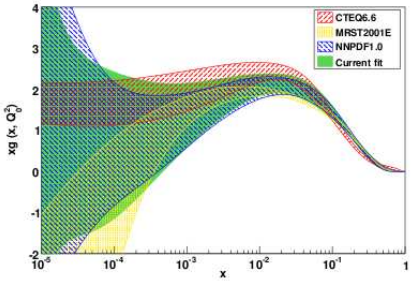
WHERE IS THE UNCERTAINTY COMING FROM?

FIT TO REPLICAS VS RANDOM SUBSET OF CENTRAL VAL.S

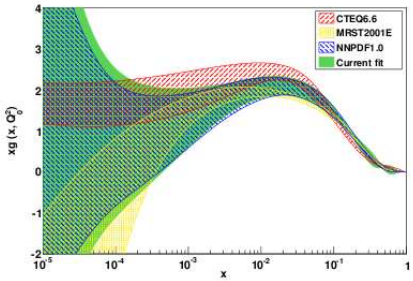
	REPLICAS	CENTRAL V.
χ^2	1.32	1.32
$\langle\chi^2\rangle_{\rm rep}$	2.79 ± 0.24	1.65 ± 0.20
$\langle\sigma^{\rm dat}\rangle$	0.039	0.035

GLUE

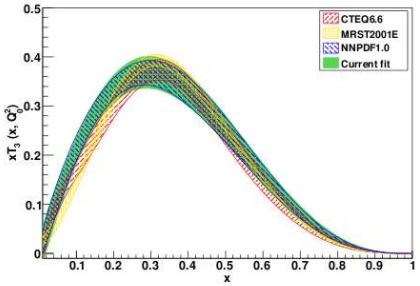
replicas

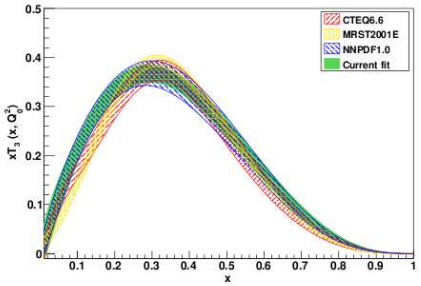


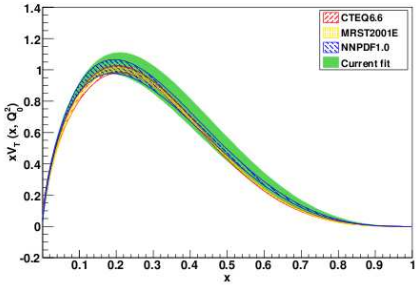
c. vals.

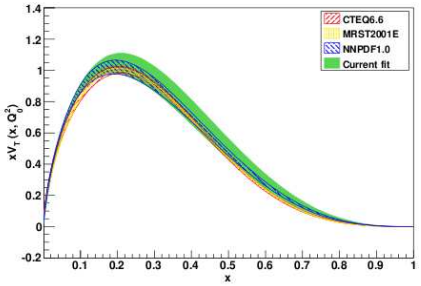


LIGHT QUARKS

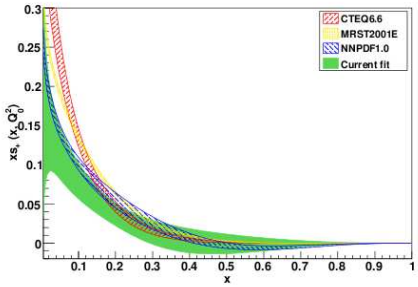


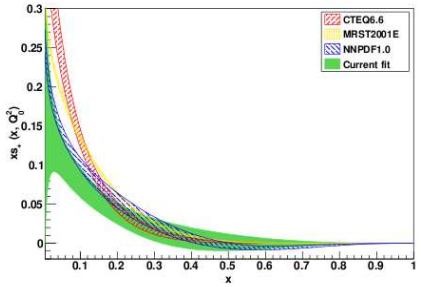






STRANGE





- QUALITY OF FIT &PDFS UNCHANGED
- REDUCTION OF $\langle\chi^2\rangle_{\rm rep}$ BY FACTOR $\sim 2 \Rightarrow$ FLUCTUATIONS ABOUT TRUE VALUE HALVED
- UNCERTAINTY ON DATA ONLY REDUCED BY 1.1 $\Rightarrow$ EXPT. UNCERTAINTIES UNDERESTIMATED OR UNDERLYING INCOMPRESSIBLE UNCERTAINTY

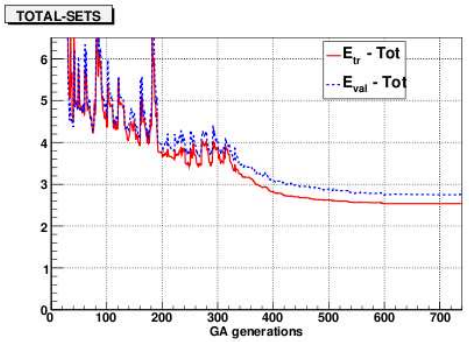
WHERE IS THE UNCERTAINTY COMING FROM?

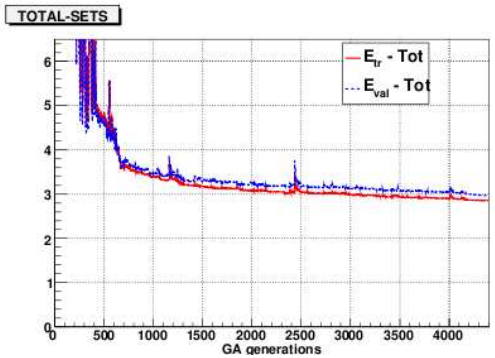
CENTRAL VALUES: VARYING PARTITION VS FIXED PARTITION

	REPLICAS	CENTRAL VALUE	FIXED PARTITION
χ^2	1.32	1.32	$\sim$1.3
$\langle\chi^2\rangle_{\rm rep}$	2.79 ± 0.24	1.65 ± 0.20	$\sim 1.6 \pm 0.2$
$\langle\sigma^{\rm dat}\rangle$	0.039	0.035	$\sim$0.03

fixed partition results obtained averaging over 5 different choices of partition (100 replicas each); more partitions needed for accurate results

- QUALITY OF FIT UNCHANGED
- $\langle\chi^2\rangle_{\rm rep}$ UNCHANGED $\Rightarrow$ CENTRAL FIT UNCHANGED
- UNCERTAINTY ON PREDICTION (I.E. ON PDFS) REDUCED

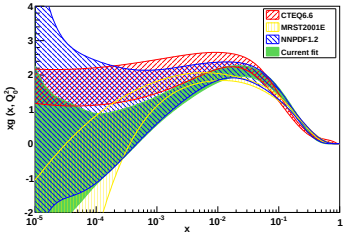




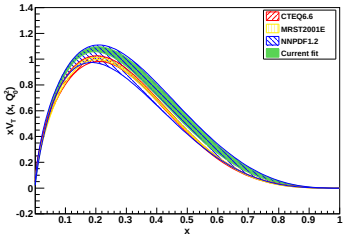
FUNCTIONAL UNCERTAINTY

- MORE THAN HALF OF UNCERTAINTY DUE TO "FUNCTIONAL FORM": $\langle\sigma^{\rm dat}\rangle =\sim 0.3$ SMALLER FOR HERA DATA
- REMAINING UNCERTAINTY ROUGHLY SCALES WITH DATA UNCERTAINTY: $\langle\sigma^{\rm dat}\rangle =\sim 0.005$ CENT.; $\langle\sigma^{\rm dat}\rangle =\sim 0.009$ REP.

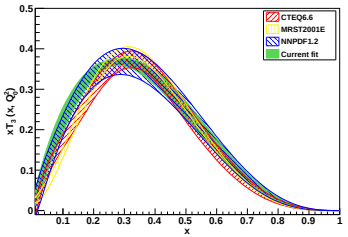
GLUE



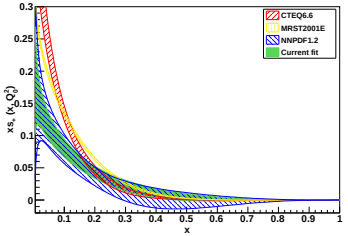
VALENCE



TRIPLET



STRANGE



ARE WE CONSTRAINED BY THE FUNCTIONAL FORM?

REMOVE STOPPING: OVERLEARNING FIT

PERFORM A FIT WITH A FIXED, VERY LARGE NUMBER OF GA GENERATIONS:
25000 gens. (AVERAGE 1000 gens. FOR STANDARD FIT)

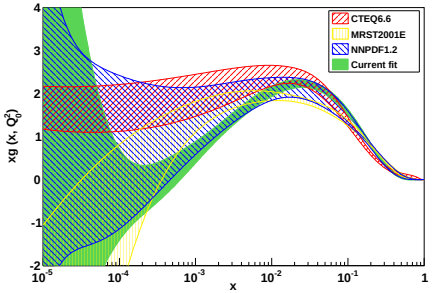
	STANDARD STOPPING			FIXED LONG	
	REPLICAS	CENTRAL VALUE	FIXED PARTITION	REPLICAS	CENTRAL VALUE
χ^2	1.32	1.32	$\sim$1.3	1.18	1.19
$\langle\chi^2\rangle_{\rm rep}$	2.79 ± 0.24	1.65 ± 0.20	$\sim 1.6 \pm 0.2$	2.43 ± 0.13	1.29 ± 0.06
$\langle\chi^2_{\rm tr}\rangle_{\rm rep}$	2.76	1.59	$\sim$1.6	2.40	1.27
$\langle\chi^2_{\rm val}\rangle_{\rm rep}$	2.80	1.61	$\sim$1.6	2.47	1.30
$\langle\sigma^{\rm dat}\rangle$	0.039	0.035	$\sim$0.03	0.032	0.019

χ^2 OF THE GLOBAL FIT DECREASES A LOT!

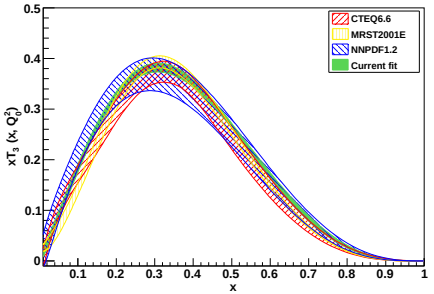
IS IT REALLY OVERLEARNING?

- PERCENTAGE DIFFERENCE BETWEEN VALIDATION AND TRAINING $\langle\chi^2\rangle_{\rm rep}$ MORE THAN DOUBLED (FROM 1.5% TO 3%)
  (NOTE 1650 DATA POINTS EACH)
- SOME PDFS HAVE FUNNY SHAPES
- REDUCTION OF $\langle\sigma^{\rm dat}\rangle$ BY FACTOR $1.7 > \sqrt{2}$
  WHEN GOING FROM REPLICAS TO CENTRAL VALUES

GLUON



TRIPLET



ARE WE CONSTRAINED BY THE FUNCTIONAL FORM?

REMOVE STOPPING: OVERLEARNING FIT

PERFORM A FIT WITH A FIXED, VERY LARGE NUMBER OF GA GENERATIONS:
25000 gens. (AVERAGE 1000 gens. FOR STANDARD FIT)

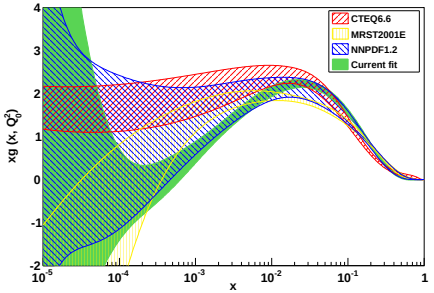
	STANDARD STOPPING			FIXED LONG	
	REPLICAS	CENTRAL VALUE	FIXED PARTITION	REPLICAS	CENTRAL VALUE
χ^2	1.32	1.32	$\sim$1.3	1.18	1.19
$\langle\chi^2\rangle_{\rm rep}$	2.79 ± 0.24	1.65 ± 0.20	$\sim 1.6 \pm 0.2$	2.43 ± 0.13	1.29 ± 0.06
$\langle\chi^2_{\rm tr}\rangle_{\rm rep}$	2.76	1.59	$\sim$1.6	2.40	1.27
$\langle\chi^2_{\rm val}\rangle_{\rm rep}$	2.80	1.61	$\sim$1.6	2.47	1.30
$\langle\sigma^{\rm dat}\rangle$	0.039	0.035	$\sim$0.03	0.032	0.019

χ^2 OF THE GLOBAL FIT DECREASES A LOT!

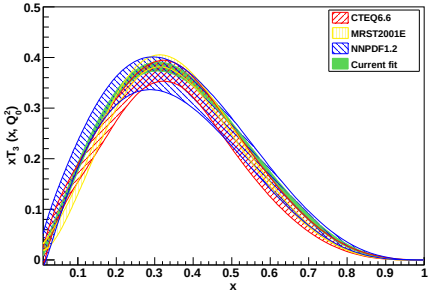
IS IT REALLY OVERLEARNING?

- PERCENTAGE DIFFERENCE BETWEEN VALIDATION AND TRAINING $\langle\chi^2\rangle_{\rm rep}$ MORE THAN DOUBLED (FROM 1.5% TO 3%)
  (NOTE 1650 DATA POINTS EACH)
- SOME PDFS HAVE FUNNY SHAPES
- REDUCTION OF $\langle\sigma^{\rm dat}\rangle$ BY FACTOR $1.7 > \sqrt{2}$
  WHEN GOING FROM REPLICAS TO CENTRAL VALUES
- AMOUNT OF OVERLEARNING SMALL, $\Leftrightarrow \langle\chi^2\rangle_{\rm rep}$ DOUBLES WHEN GOING FROM CENTRAL VALS. TO REPLICAS,
  SHOULD REMAIN UNCHANGED FOR EXTREME OVERLEARNING

GLUON



TRIPLET



YES!

WHERE IS THE UNCERTAINTY COMING FROM?

WHEN THE BEST FIT IS NOT AT THE MINIMUM

	STANDARD STOPPING			FIXED LONG	
	REPLICAS	CENTRAL VALUE	FIXED PARTITION	REPLICAS	CENTRAL VALUE
χ^2	1.32	1.32	1.35	1.18	1.19
$\langle\chi^2\rangle_{\rm rep}$	2.79 ± 0.24	1.65 ± 0.20	1.60 ± 0.19	2.43 ± 0.13	1.29 ± 0.06
$\langle\sigma^{\rm dat}\rangle$	0.39	0.35	0.28	0.32	0.19

- FIT QUALITY:
  - "FUNCTIONAL" UNCERTAINTY SUPPRESSED IN OVERLEARNING FITS:
    $\Rightarrow \langle\sigma^{\rm dat}\rangle \approx 0.2 \Rightarrow$ "DATA" UNCERTAINTY
  - FLUCTUATION OF $\langle\chi^2\rangle_{\rm rep}$ FOR OVERLEARNING FIT STATISTICAL:
    $$\sigma = \sqrt{\frac{2}{N_{\rm dat}}} \approx 0.05$$
  - FLUCTUATION OF $\langle\chi^2\rangle_{\rm rep}$ IN STANDARD FIT MUCH LARGER:
    CONTROLLED BY DISTANCE FROM THE MINIMUM
    IF $\Delta\chi^2 = 1$ DUE TO UNDERLYING PARM AT $\chi^2_{\rm min}$, THEN ONE SIGMA VARIATION AROUND $\chi^2_0 > \chi^2_{\rm min}$ EQUALS $\sqrt{\chi^2_0 - \chi^2_{\rm min}}$
- DATA INCONSISTENCY: FOR STANDARD FIT, VALUE OF $\chi^2 = 1.3 > 1$
  $\Rightarrow$ ERRORS UNDERESTIMATED BY 30%

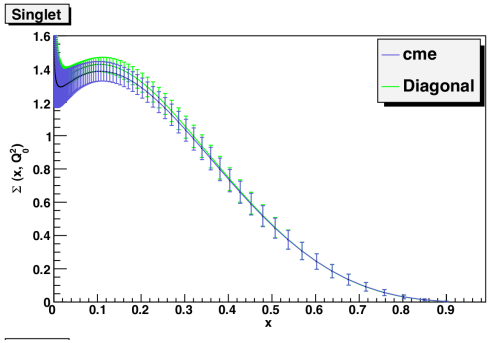
THE IMPACT OF CORRELATED UNCERTAINTIES

REPEAT THE FIT NEGLECTING ALL CORRELATIONS (A.Donati)

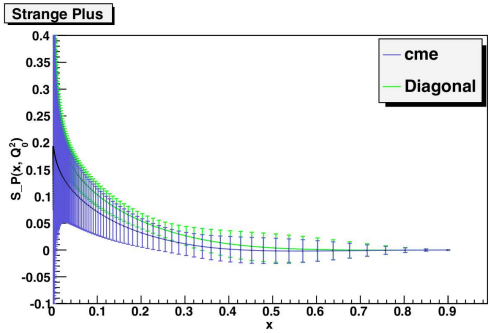
		CME fit		Diagonal fit	
Experiment	Set	$\chi^2_{\rm diag}$	$\chi^2_{\rm cme}$	$\chi^2_{\rm diag}$	$\chi^2_{\rm cme}$
TOT (all exp)		0.988	1.323	0.844	1.321
NMC-pd		1.965	1.457	1.167	1.155
NMC		1.006	1.659	1.078	1.76
SLAC		0.836	1.185	1.008	1.406
	SLACp	1.018	1.307	1.132	1.525
	SLACd	0.651	0.912	0.882	1.275
BCDMS		0.777	1.646	0.552	1.604
	BCDMSp	0.873	1.808	0.617	1.703
	BCDMSd	0.648	1.296	0.465	1.23
ZEUS		0.770	1.055	0.742	1.048
	Z97lowQ2	0.474	1.294	0.434	1.367
	Z97NC	0.718	1.125	0.669	1.106
	Z97CC	0.912	0.800	1.021	0.894
	Z02NC	0.798	0.767	0.763	0.733
	Z02CC	0.619	0.592	0.593	0.569
	Z03NC	0.975	1.104	0.907	1.012
	Z03CC	1.131	1.001	1.259	1.115
H1		1.020	1.053	0.997	1.028
	H197mb	0.861	1.298	0.877	1.33
	H197lwQ2	0.666	0.948	0.774	0.97
	H197NC	1.071	0.903	0.986	0.852
	H197CC	0.758	0.764	0.831	0.824
	H199NC	1.229	1.109	1.171	1.068
	H199CC	0.621	0.646	0.644	0.668
	H199NChy	0.333	0.361	0.326	0.353
	H100NC	1.208	1.172	1.120	1.102
	H100CC	1.122	1.013	1.311	1.146
CHORUS		1.018	1.380	0.745	1.392
	CHORUSnu	1.082	1.449	0.628	1.403
	CHORUSnb	0.954	1.178	0.861	1.254
FLH108		0.984	1.729	0.946	1.7
NTVDMN		0.869	0.692	1.094	0.984
	NTVnuDMN	1.061	0.763	0.445	0.421
	NTVnbDMN	0.667	0.660	1.774	1.618
ZEUS-H2		1.392	1.509	1.373	1.512
	Z06NC	1.691	1.495	1.667	1.472
	Z06CC	0.664	1.230	0.659	1.252

- DIAGONAL χ^2 OF DIAGONAL FIT MUCH LOWER, CORREL. χ^2 OF TWO FITS UNCHANGED
- DIAGONAL FIT REWEIGHTS EXPERIMENTS
  $\Rightarrow$ EXPTS WITH LARGER SYST. (FIXED TARGET) GET SMALLER WEIGHT
- VALENCE & STRANGE PDFS AFFECTED AT THE $\frac{1}{4}\sigma$ LEVEL

SINGLET



STRANGE



SUMMARY

- A LARGE FRACTION OF THE UNCERTAINTY COMES FROM THE FREEDOM TO CHOOSE THE FUNCTIONAL FORM
  FLUCTUATIONS OF FIT QUALITY DOMINATED BY LACK OF KNOWLEDGE OF THE "TRUE" UNDERLYING FUNCTIONAL FORM
- SOME DATA INCOMPATIBILITY (UNDERESTIMATION OF DATA UNCERTAINTY), BUT SMALL EFFECT
  ABOUT 30% ON AVERAGE, CONCENTRATED ON LIMITED NUMBER OF DATA POINTS
- INCLUSION OF CORRELATED SYSTEMATICS HAS A SMALL BUT NON-NEGLIGIBLE EFFECT