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 5, 2023

III: PROPER LEARNING

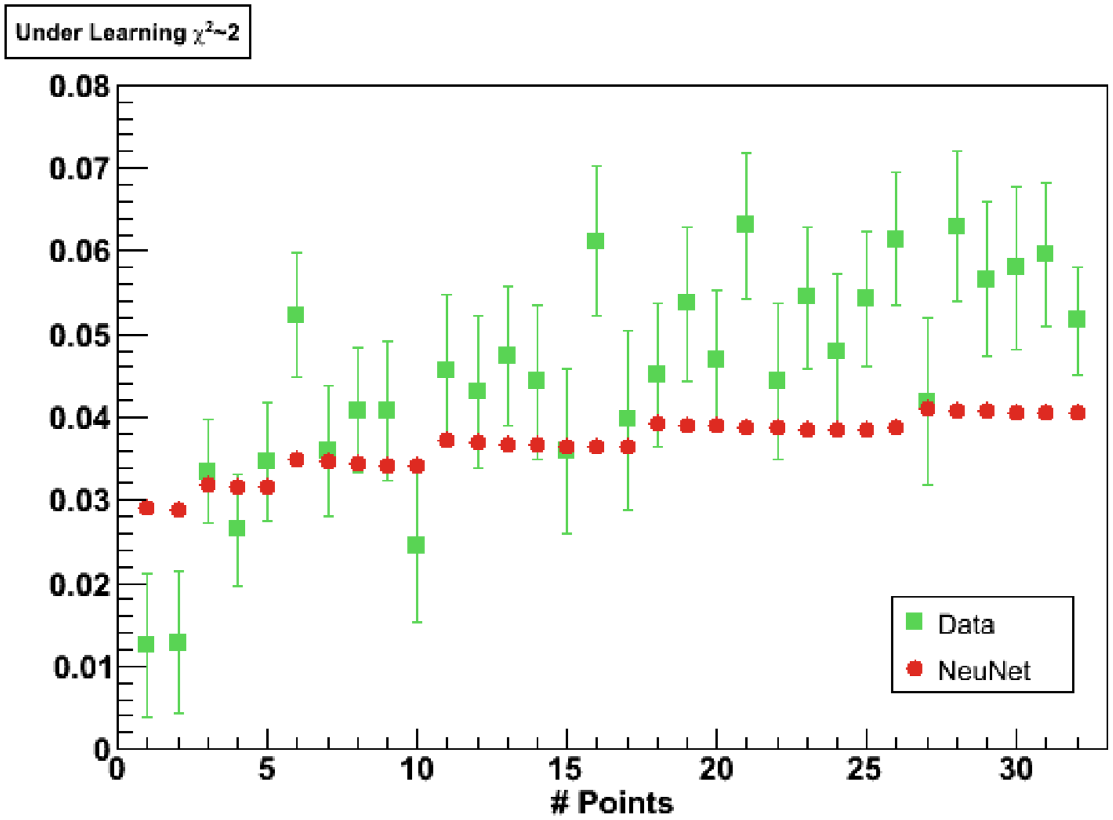
- CROSS-VALIDATION
  - NEURAL LEARNING
  - TRAINING AND VALIDATION
  - STOPPING
- HYPEROPTIMIZATION
  - HYPERPARAMETER OPTIMIZATION
  - OVERFITTING AND OVERFITTING METRICS
- GENERALIZATION
  - THE TEST SET METHOD
  - K-FOLDS

CROSS-VALIDATION

LEARNING

- COMPLEXITY INCREASES WITH DECREASING LOSS
- UNTIL LEARNING NOISE
- WHEN SHOULD ONE STOP?

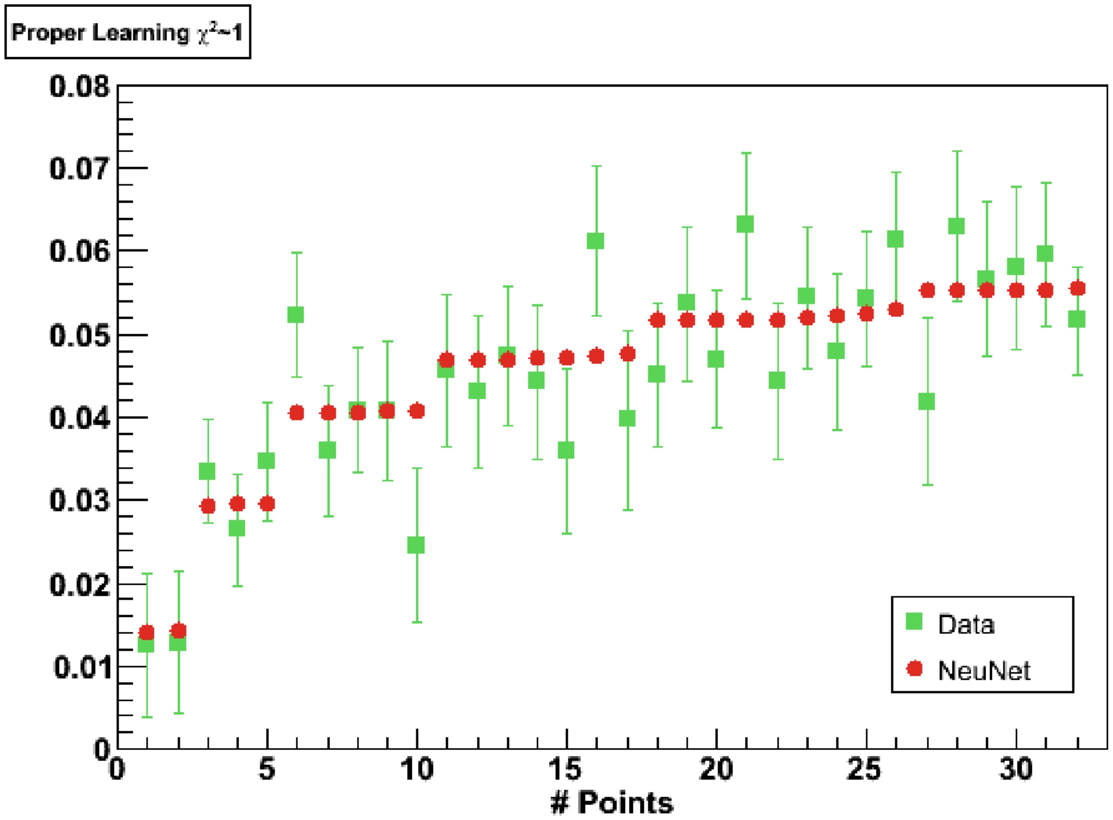
UNDERLEARNING



LEARNING

- COMPLEXITY INCREASES WITH DECREASING LOSS
- UNTIL LEARNING NOISE
- WHEN SHOULD ONE STOP?

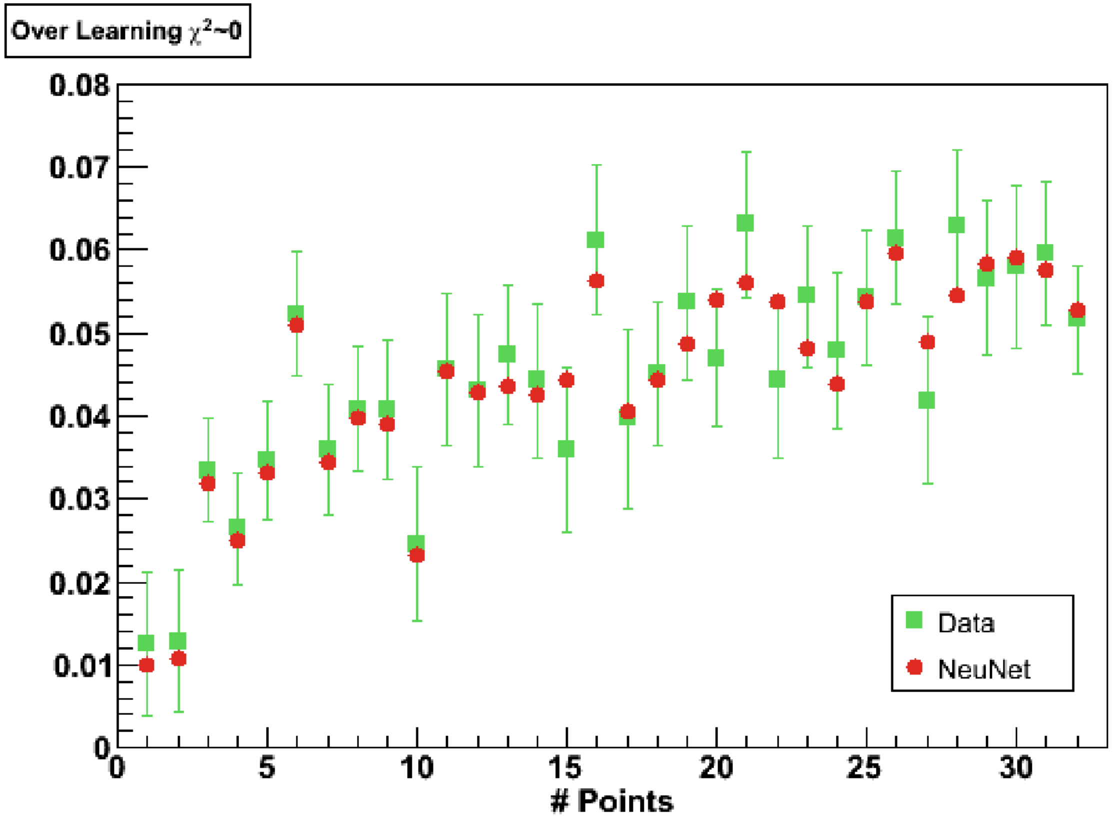
PROPER LEARNING



LEARNING

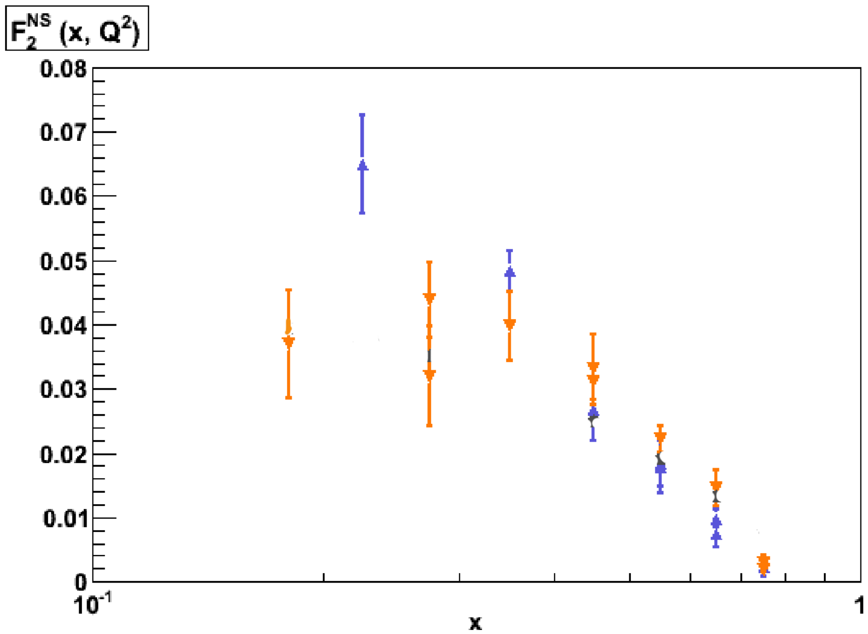
- COMPLEXITY INCREASES WITH DECREASING LOSS
- UNTIL LEARNING NOISE
- WHEN SHOULD ONE STOP?

OVERLEARNING



OPTIMAL LEARNING: CROSS-VALIDATION

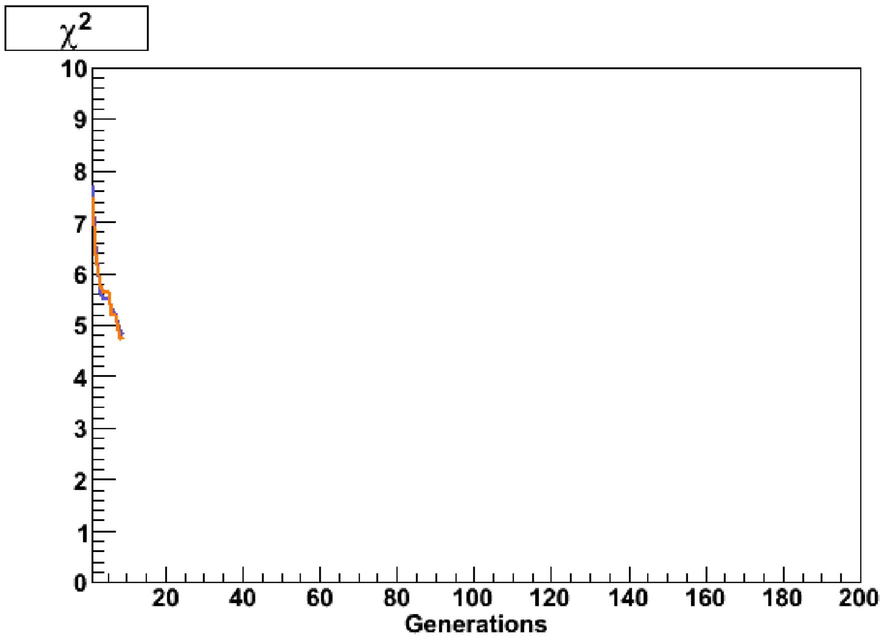
- DIVIDE THE DATA IN TWO SETS: TRAINING AND VALIDATION
- MINIMIZE THE LOSS OF THE DATA IN THE TRAINING SET
- AT EACH ITERATION, COMPUTE THE LOSS FOR THE DATA IN THE VALIDATION SET (NOT USED FOR TRAINING)
- WHEN THE VALIDATION LOSS STOPS DECREASING, STOP THE TRAINING

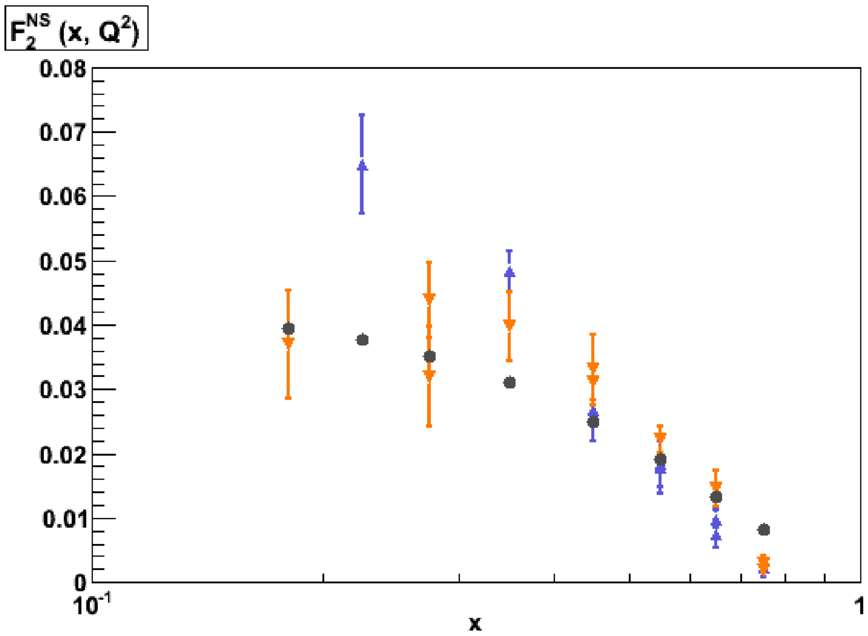


OPTIMAL LEARNING: CROSS-VALIDATION

- DIVIDE THE DATA IN TWO SETS: TRAINING AND VALIDATION
- MINIMIZE THE LOSS OF THE DATA IN THE TRAINING SET
- AT EACH ITERATION, COMPUTE THE LOSS FOR THE DATA IN THE VALIDATION SET (NOT USED FOR TRAINING)
- WHEN THE VALIDATION LOSS STOPS DECREASING, STOP THE TRAINING

GO!

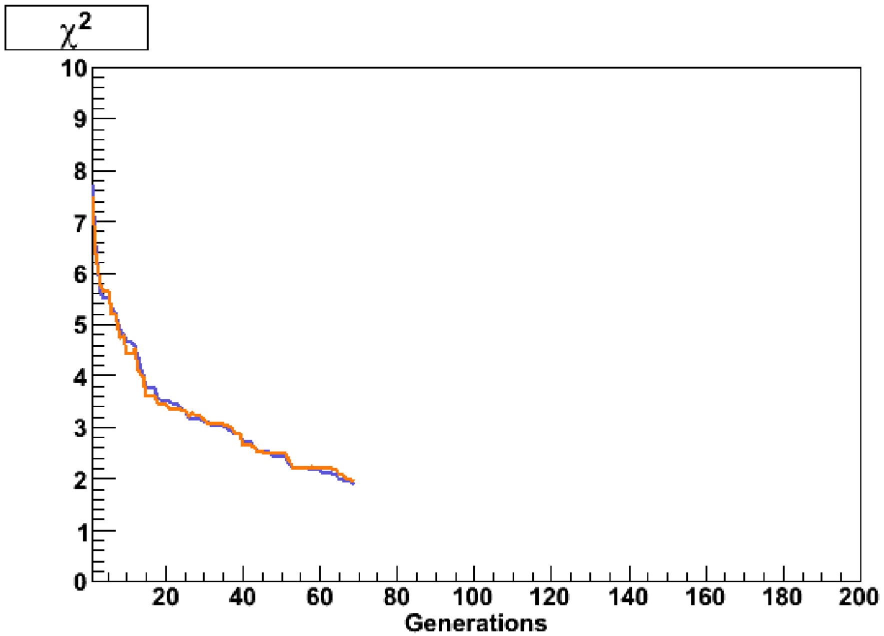


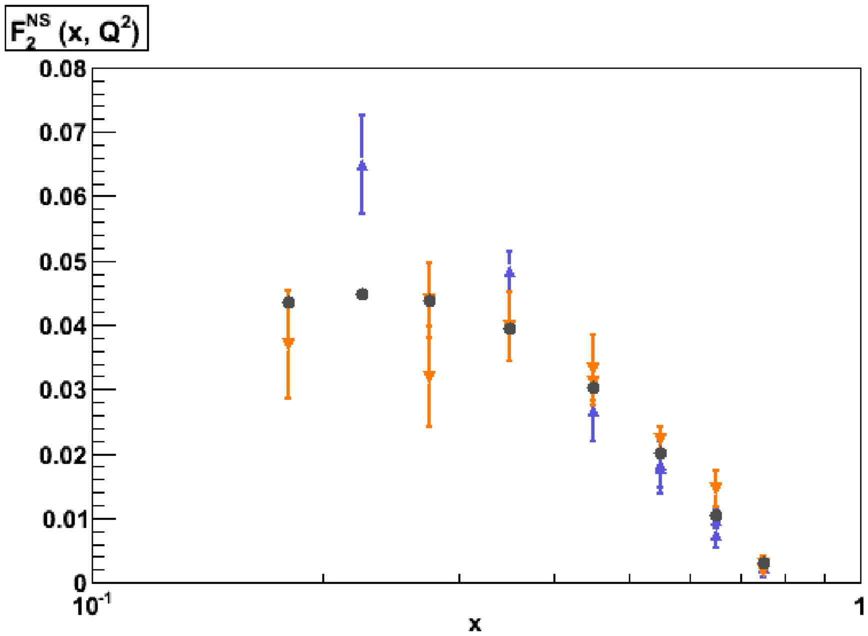


OPTIMAL FIT: CROSS-VALIDATION

- DIVIDE THE DATA IN TWO SETS: TRAINING AND VALIDATION
- MINIMIZE THE LOSS OF THE DATA IN THE TRAINING SET
- AT EACH ITERATION, COMPUTE THE LOSS FOR THE DATA IN THE VALIDATION SET (NOT USED FOR FITTING)
- WHEN THE VALIDATION LOSS STOPS DECREASING, STOP THE FIT

STOP!



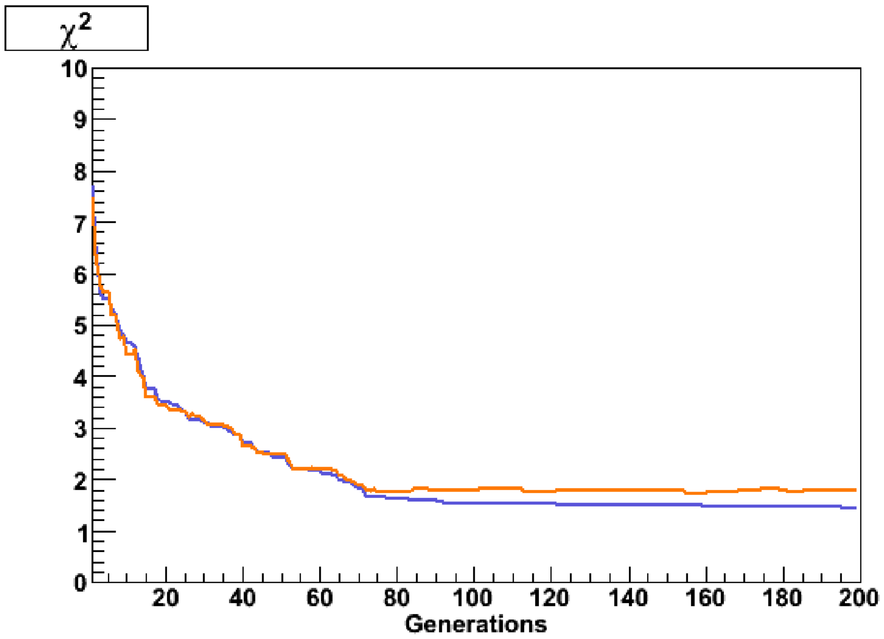


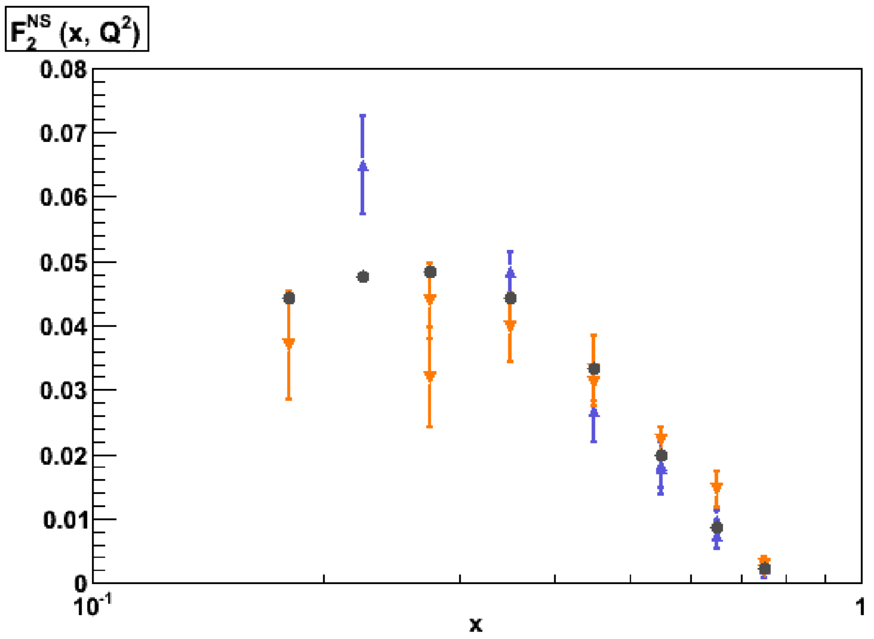
OPTIMAL FIT: CROSS-VALIDATION

GENETIC MINIMIZATION:
AT EACH GENERATION, χ^2 EITHER UNCHANGED OR DECREASING

- DIVIDE THE DATA IN TWO SETS: TRAINING AND VALIDATION
- MINIMIZE THE χ^2 OF THE DATA IN THE TRAINING SET
- AT EACH ITERATION, COMPUTE THE χ^2 FOR THE DATA IN THE VALIDATION SET (NOT USED FOR FITTING)
- WHEN THE VALIDATION χ^2 STOPS DECREASING, STOP THE FIT

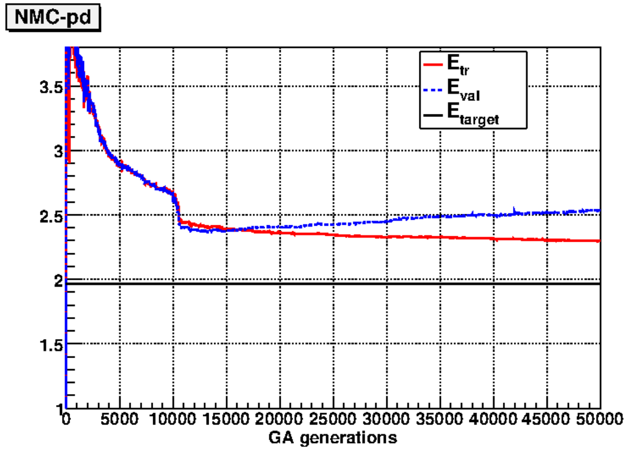
TOO LATE!





STOPPING

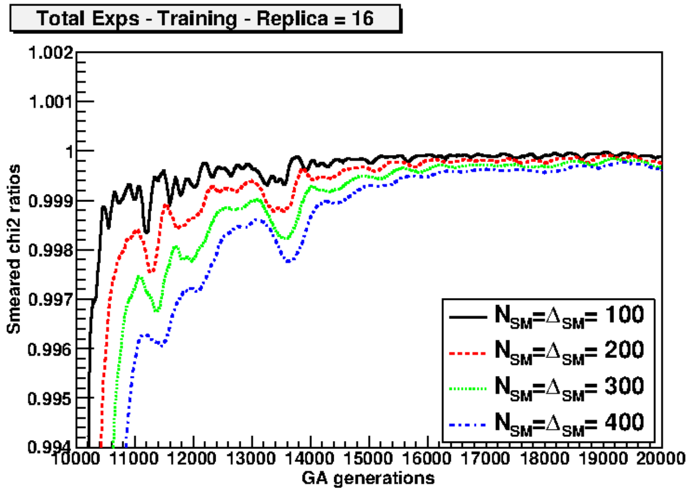
TRAINING/VALIDATION LOSS WITH NO STOPPING



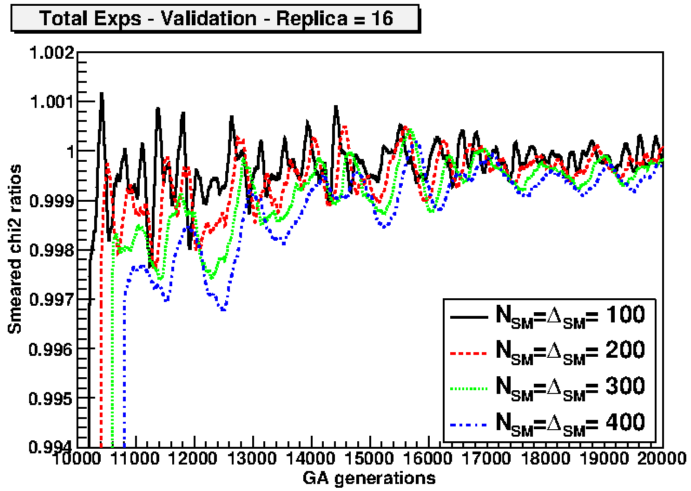
THRESHOLD STOPPING

- define tr/val ratios $r_{\rm tr} \equiv \frac{\langle E_{\rm tr}(i)\rangle}{\langle E_{\rm tr}(i-\Delta_{\rm smear})\rangle}$, $r_{\rm val} \equiv \frac{\langle E_{\rm val}(i)\rangle}{\langle E_{\rm val}(i-\Delta_{\rm smear})\rangle}$
  WITH MOVING-AVERAGED LOSS $\langle E_{\rm tr,val}(i)\rangle \equiv \frac{1}{N_{\rm smear}} \sum_{l=i-N_{\rm smear}+1}^{i} E_{\rm tr,val}(l)$
- STOP IF $r_{\rm tr} > 1-\delta_{\rm tr}; \quad r_{\rm val} > 1+\delta_{\rm val}$
  (training does not decrease too much, validation increases)

AVERAGED TRAINING LOSS

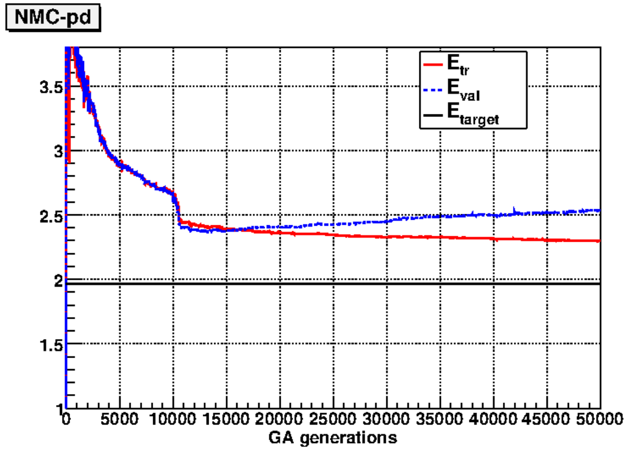


AVERAGED VALIDATION LOSS



STOPPING

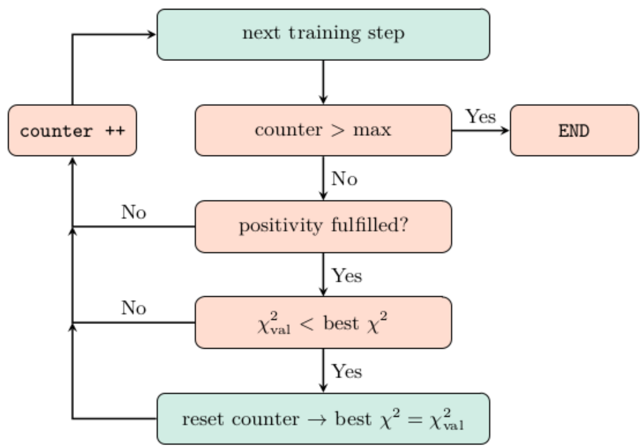
TRAINING/VALIDATION LOSS WITH NO STOPPING



LOOKBACK STOPPING

- NO (INFINITE) PATIENCE:
  - TRAIN FOR MAX $N_{\rm max}$ GENERATIONS
- FINITE PATIENCE
  - VALIDATION LOSS NOT DECREASING
    $\Rightarrow$ KEEP TRAINING FOR $N_{\rm patience}$ GENERATIONS
- GO BACK& STOP AT ABSOLUTE MINIMUM OF VALIDATION LOSS

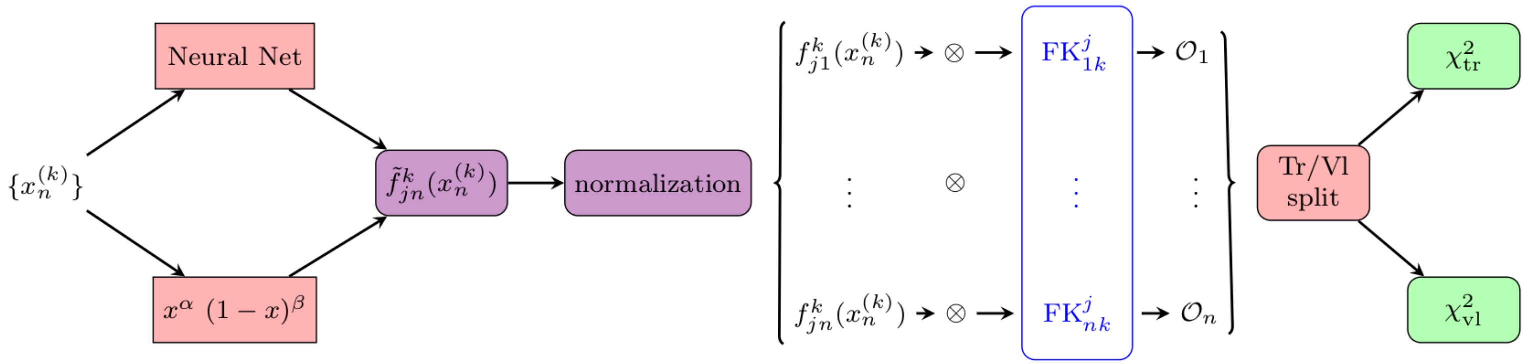
THE PATIENCE ALGORITHM



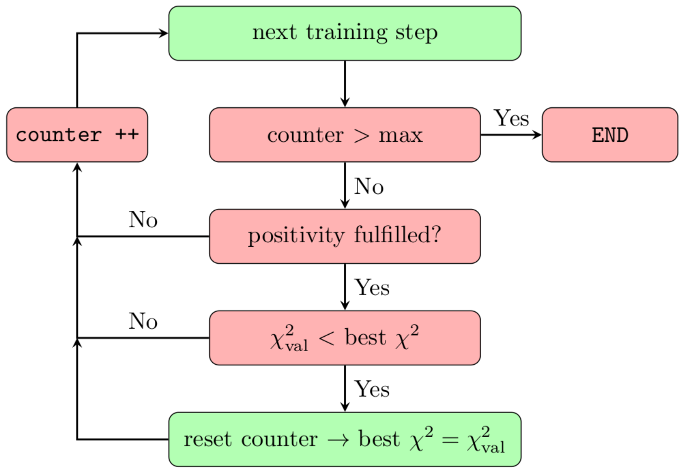
HYPEROPTIMIZATION

THE ALGORITHM

CROSS-VALIDATION



STOPPING



THE HYPERPARAMETERS

MODEL	MINIMIZATION
Number of layers	Optimizer
Size of each layer	Initializer
Activation functions	Learning rate
Initial positivity	Clipnorm
Initial integrability	Maximum number of epochs
	Stopping Patience

HYPERPARAMETER SELECTION

GAUSSIAN PROCESS INTERPOLATION

- VIEW FUNCTION $f(x_i)$ AS VECTOR $\vec{y}$ WITH COMPONENTS $y_i = f(x_i)$
- ASSUME y_i DISTRIBN. MULTIGAUSSIAN: $p(y_i) = \exp \frac{1}{2}(y_i - y_i^0) C_{ij} (y_j y_j^0)$
- ASSUME 0-TH ORDER COVARIANCE MATRIX GIVEN BY KERNEL DEFINED FOR ALL x: $C_{ij} = K(x_i, x_j)$
  E.G. $K(x, x') = \theta_o \exp - \left[\frac{\theta_1}{2} (x - x')^2\right] + \theta_2 + \theta_3 x x'$
- COMBINED GAUSSIAN C_{ij} BASED ON OBSERVED $y_i \Rightarrow$ MULTIGAUSSIAN WITH $C_{ij} = K(x_i, x_j) + \mathrm{cov}_{ij}$, cov_{ij} EXPT COVARIANCE MATRIX
- DETERMINE POSTERIOR (CONDITIONAL) GAUSSIAN FOR UNOBSERVED x_i

GOAL: MINIMIZE LOSS IN PARAMETER SPACE

- SAMPLE LOSS FOR A SET OF HYPERPARAMETER VALUES
- INTERPOLATE LOSS USING GAUSSIAN PROCESS
- LOOK FOR POINTS WITH MAXIMAL EXPECTED GAIN
  $\Rightarrow$ CLOSE TO MIN OF INTERPOLATED LOSS, OR WITH LARGE UNCERTAINTY
- SAMPLE AGAIN

THE APPLICATION OF BAYESIAN METHODS FOR SEEKING THE EXTREMUM

J. Mockus, V. Tiešis, A. Žilinskas
Institute of Mathematics and Cybernetics
Academy of Sciences of the Lithuanian SSR
Vilnius

The purpose of this paper is to describe how the Bayesian approach can be applied to the global optimization of multiextremal functions. The function to be minimized is considered as a realization of some stochastic function. The optimization technique based upon the minimization of the expected deviation from the extremum is called Bayesian. The implementation of Bayesian methods is considered.

The results of the application to the minimization of some standard test functions are given.

INTRODUCTION

Many well known methods for seeking the extremum have been developed on the basis of quadratic approximation. In some problems of global optimization the function to be minimized can be considered as a realization of some stochastic function. The optimization technique based upon the minimization of the expected deviation from the extremum is called Bayesian.

The description of such methods is given in [1, 2, 3]. However, to make this paper reasonably complete a brief definition of the Bayesian methods will be given.

DEFINITION OF BAYESIAN METHODS

Assume the function to be minimized is a realization $f(x, \omega)$ of some stochastic function $f(x)$, where $x \in A \subset R^n$ and $\omega \in \Omega$ is some fixed but unknown index.

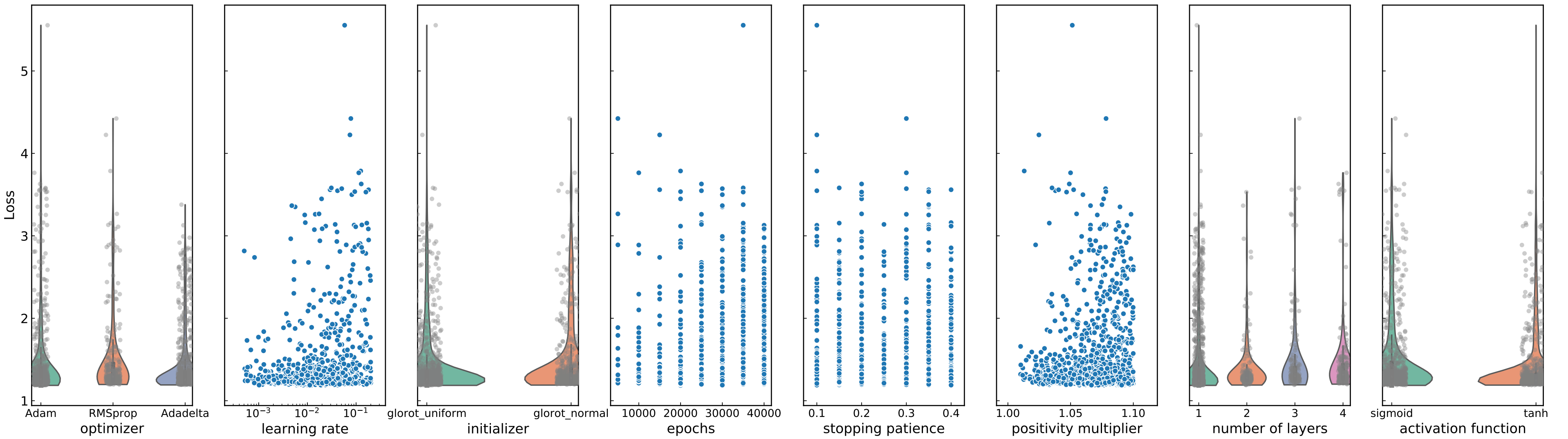
The probability distribution P on Ω is defined by the equalities:

$$F_{x_1, \dots, x_m}(y_1, \dots, y_m) = P\left\{ \omega : f(x_1, \omega) < y_1, \dots, f(x_m, \omega) < y_m \right\} \qquad (1)$$

where P is a priori probability of an event:

$$\left\{ \omega : f(x_1, \omega) < y_1, \dots, f(x_m, \omega) < y_m \right\} \qquad (2)$$

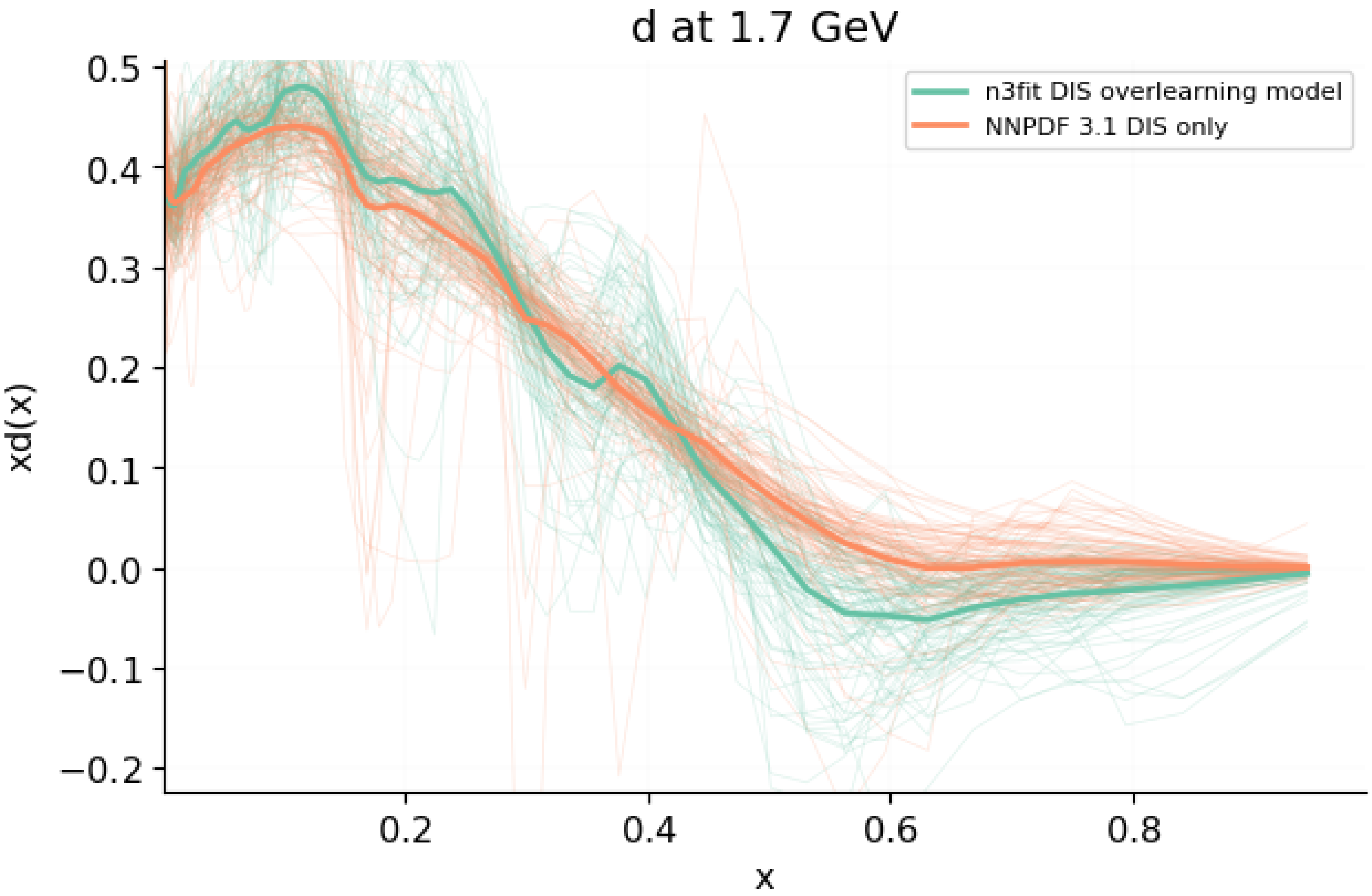
HYPEROPTIMIZATION SCAN



- BAYESIAN SCAN OF PARAMETER SPACE
- OPTIMIZE LOSS: VALIDATION χ^2

RESULTS: OVERFITTING!

DOWN QUARK: HYPEROPTIMIZED VS. HAND-PICKED



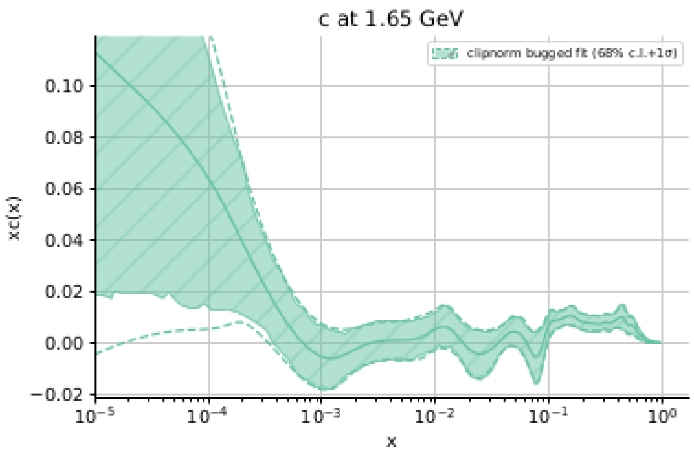
- HAND-PICKED: WIGGLES: FINITE SIZE ⇒ WILL GO AWAY AS N_{rep} GROWS
- HYPEROPT: WIGGLY PDFS ⇔ OVERFITTING ⇒ WILL NOT GO AWAY ($\chi^2_{\mathrm{train}} \ll \chi^2_{\mathrm{valid}}$ EVEN THOUGH VALIDATION LOSS MNIMIZED)

OBJECTIVE?
VALIDATION: OVERFITTING METRIC

- TEST VALIDATION $\chi^2_{\mathrm{val}'}$
  - DIFFERENT FLUCTUATED VALIDATION DATA
  - BUT KEEP SAME TRAINING-VALIDATION SPLIT
- COMPUTE AVERAGE OVER REPLICAS $\langle\chi^2_{\mathrm{val}'}\rangle$ &
  DETERMINE DIFFERENCE TO STANDARD VALIDATION χ^2_{val}
  OVERFITNESS: $\mathcal{R}_O = \chi^2_{\mathrm{val}} - \langle\chi^2_{\mathrm{val}'}\rangle$
- NEGATIVE OVERFITNESS $\mathcal{R}_O \Rightarrow$ OVERFIT

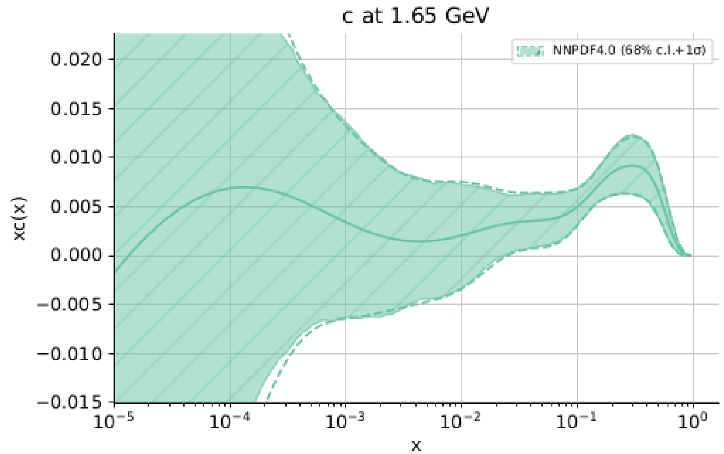
CHARM PDF

OVERFIT (NO CLIPNORM)



$$\mathcal{R}_O = -0.024 \pm 0.012$$

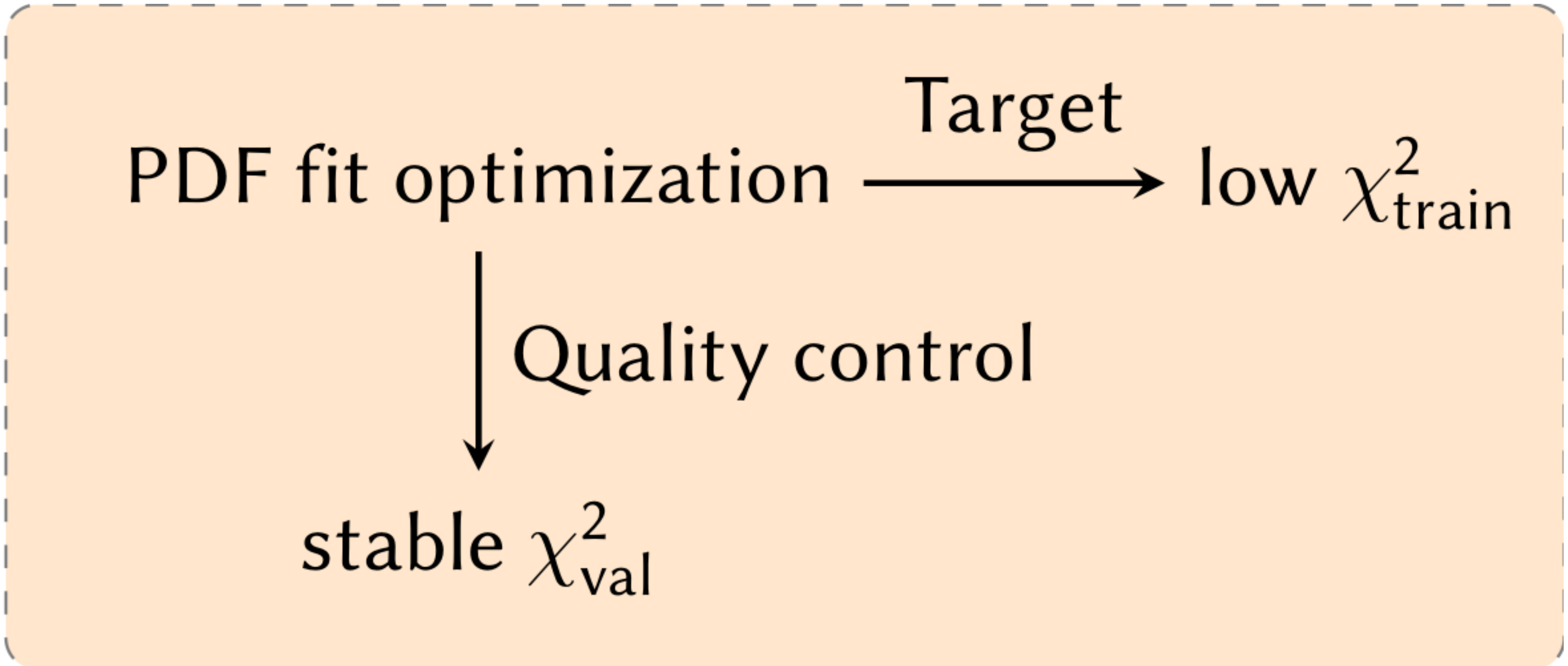
PROPER FIT (NNPDF4.0)



$$\mathcal{R}_O = -0.001 \pm 0.013$$

WHAT HAPPENED?

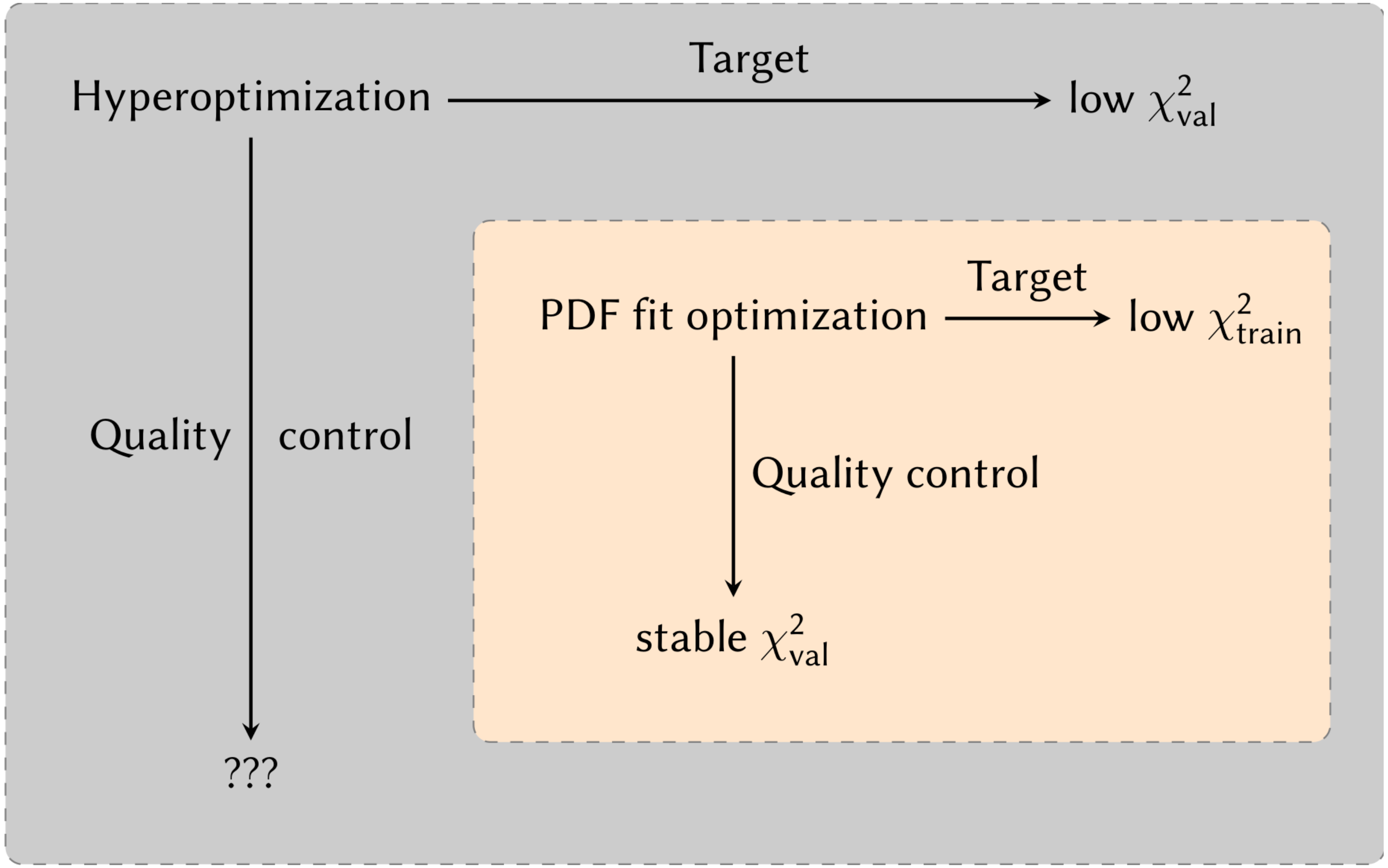
OPTIMIZATION



CROSS-VALIDATION SELECTS THE OPTIMAL MINIMUM

WHAT HAPPENED?

HYPEROPTIMIZATION

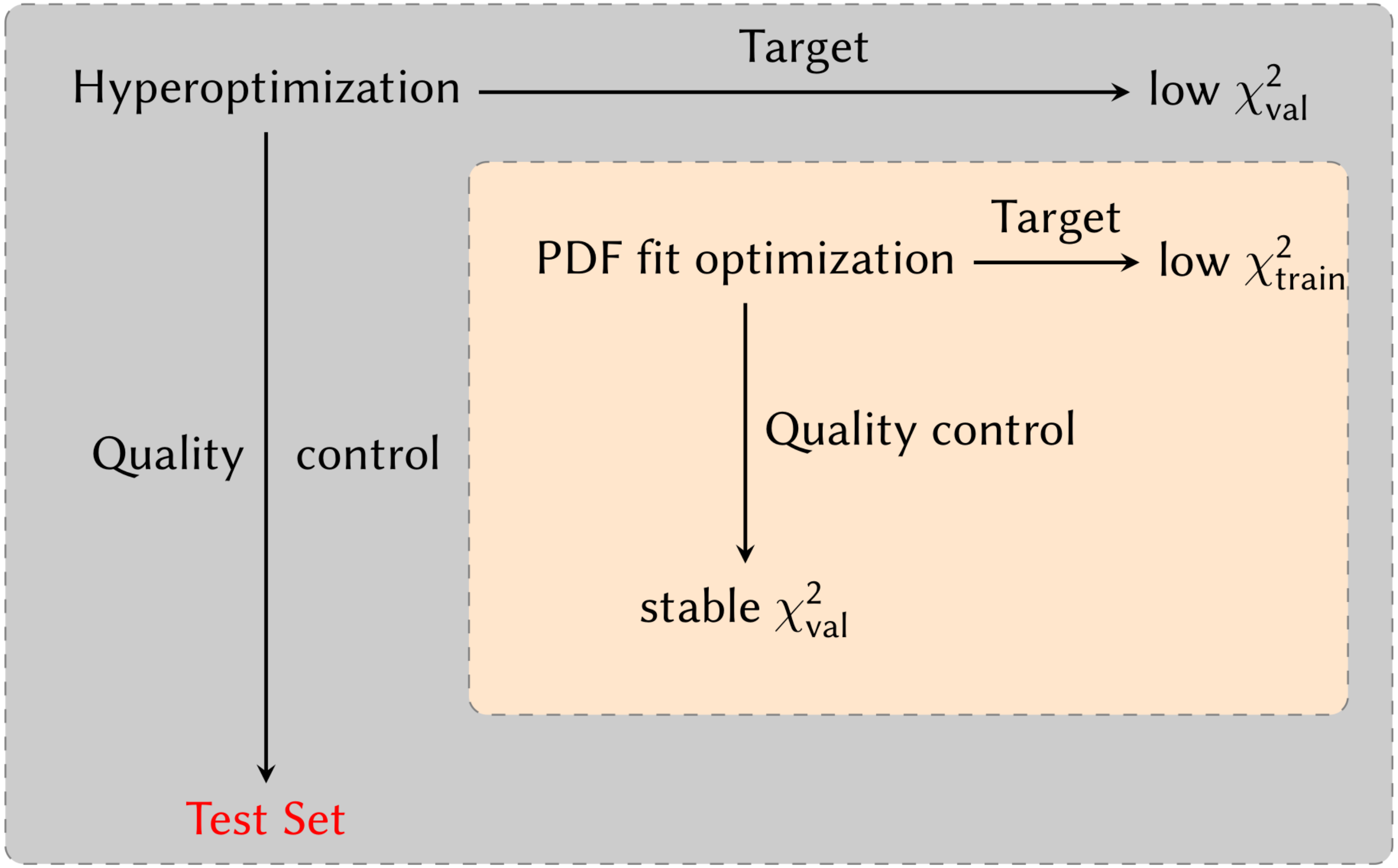


WE ARE MISSING A SELECTION CRITERION

GENERALIZATION

THE SOLUTION

THE TEST SET



COMPARE TO A A TEST SET $\Rightarrow$ NEW DATA PREVIOUSLY NOT USED AT ALL
TESTS GENERALIZATION POWER

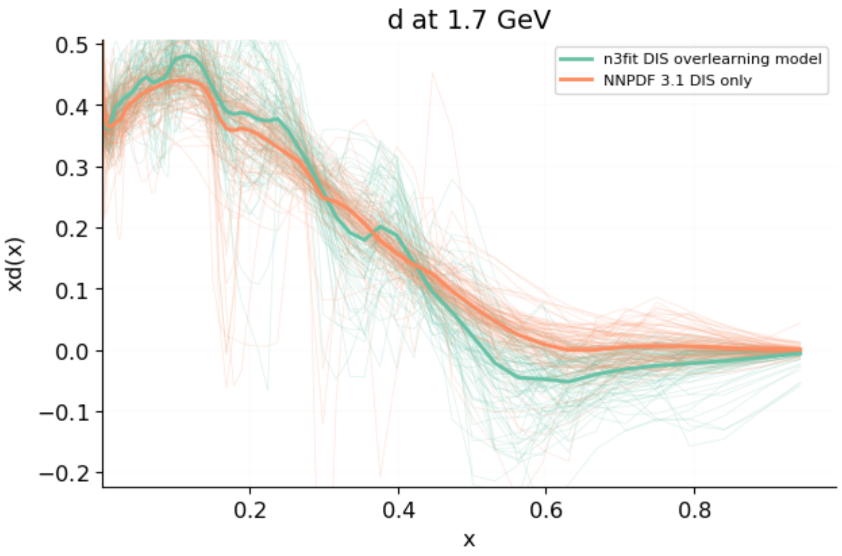
TEST SET RESULS

- COMPLETELY UNCORRELATED TEST SET (JETS, FOR DIS-ONLY DATASET)
- OPTIMIZE ON WEIGHTED AVERAGE OF VALIDATION AND TEST
  $\Rightarrow$ NO OVERLEARNING

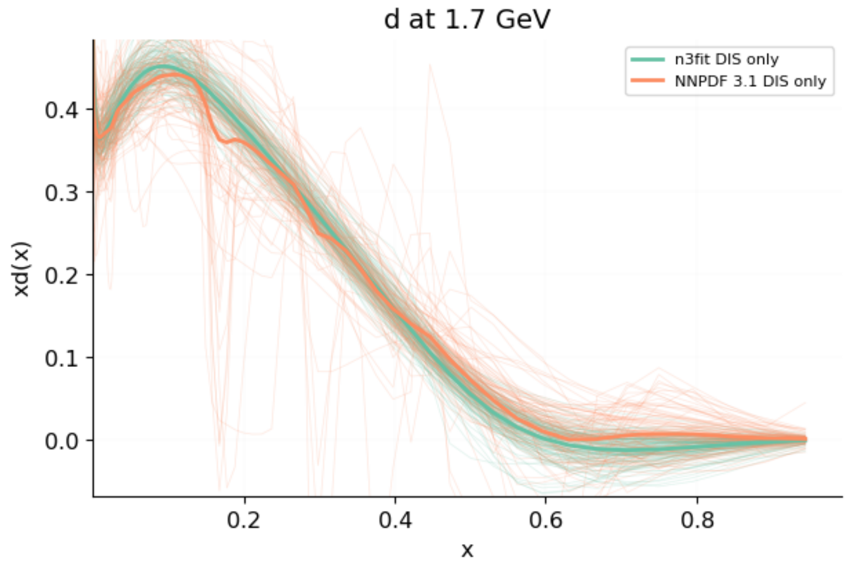
HYPEROPTIMIZED PDFs

DOWN QUARK

OVERFIT VS HAND-PICKED



TEST-SET VS HAND-PICKED



- IS THE TEST SET REALLY INDEPENDENT?
- IS IT GENERAL ENOUGH?

K-FOLDS

THE BASIC IDEA:

- DIVIDE THE DATA INTO n REPRESENTATIVE SUBSETS
  EACH CONTAINING PROCESS TYPES, KINEMATIC RANGE OF FULL SET
- TRAIN $n-1$ SETS AND USE n-TH SET AS TEST
  $\Rightarrow n$ VALUES OF $\chi^2_{\text{test, i}}$

Fold 1		
CHORUS σ^{ν}_{CC}	HERA I+II inc NC e^+p 920 GeV	BCDMS p
LHCb Z 940 pb	ATLAS W, Z 7 TeV 2010	CMS Z p_T 8 TeV (p^{ll}_T, y_{ll})
DY E605 $\sigma^p_{\rm DY}$	CMS Drell-Yan 2D 7 TeV 2011	CMS 3D dijets 8 TeV
ATLAS single-$\bar{t}$ y (normalised)	ATLAS single top R_t 7 TeV	CMS $t\bar{t}$ rapidity $y_{t\bar{t}}$
CMS single top R_t 8 TeV		

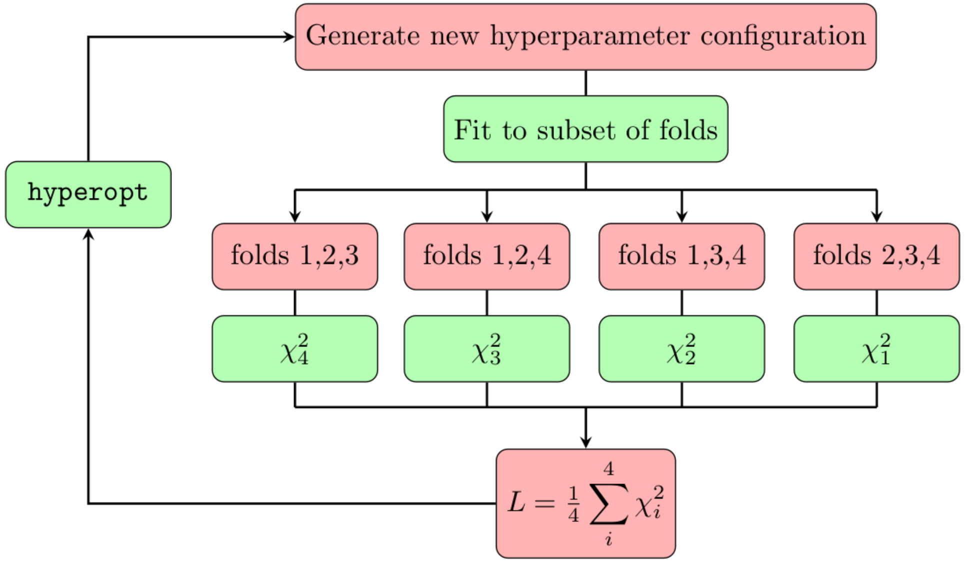
Fold 2		
HERA I+II inc CC e^-p	HERA I+II inc NC e^+p 460 GeV	HERA comb. $\sigma^{\rm red}_{b\bar{b}}$
NMC p	NuTeV $\sigma^{\bar{\nu}}_c$	LHCb $Z \to ee$ 2 fb
CMS W asymmetry 840 pb	ATLAS Z p_T 8 TeV (p^{ll}_T, M_{ll})	D0 $W \to \mu\nu$ asymmetry
DY E886 $\sigma^p_{\rm DY}$	ATLAS direct photon 13 TeV	ATLAS dijets 7 TeV, R=0.6
ATLAS single antitop y (normalised)	CMS $\sigma^{\rm tot}_{tt}$	CMS single top $\sigma_t + \sigma_{\bar{t}}$ 7 TeV

Fold 3		
HERA I+II inc CC e^+p	HERA I+II inc NC e^+p 575 GeV	NMC d/p
NuTeV σ^{ν}_c	LHCb $W, Z \to \mu$ 7 TeV	LHCb $Z \to ee$
ATLAS W, Z 7 TeV 2011 Central selection	ATLAS W^++jet 8 TeV	ATLAS HM DY 7 TeV
CMS W asymmetry 4.7 fb	DYE 866 $\sigma^d_{\rm DY}/\sigma^p_{\rm DY}$	CDF Z rapidity (new)
ATLAS $\sigma^{\rm tot}_{tt}$	ATLAS single top y_t (normalised)	CMS $\sigma^{\rm tot}_{tt}$ 5 TeV
CMS $t\bar{t}$ double diff. $(m_{t\bar{t}}, y_t)$		

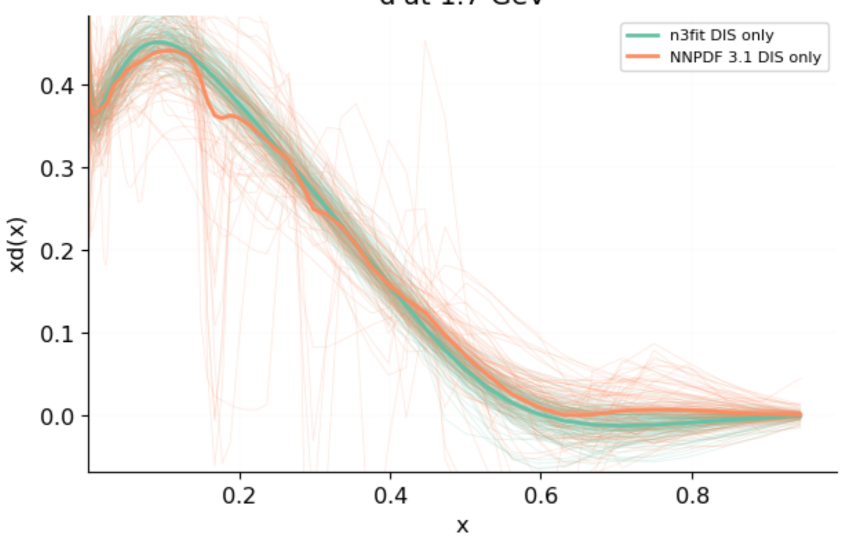
Fold 4		
CHORUS $\sigma^{\bar{\nu}}_{CC}$	HERA I+II inc NC e^+p 820 GeV	LHCb $W, Z \to \mu$ 8 TeV
LHCb $Z \to \mu\mu$	ATLAS W, Z 7 TeV 2011 Fwd	ATLAS W^-+jet 8 TeV
ATLAS low-mass DY 2011	ATLAS Z p_T 8 TeV (p^{ll}_T, y_{ll})	CMS W rapidity 8 TeV
D0 Z rapidity	CMS dijets 7 TeV	ATLAS single top y_t (normalised)
ATLAS single top R_t 13 TeV	CMS single top R_t 13 TeV	

K-FOLD VALIDATION

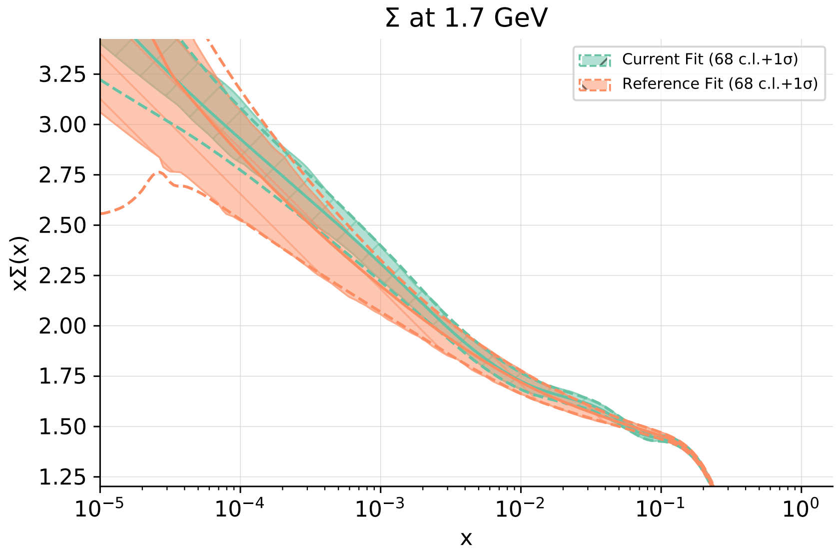
LOSS: AVERAGE χ^2 OF NON-FITTED FOLDS



TEST-SET VS HAND-PICKED



K-FOLD VS. TEST-SET



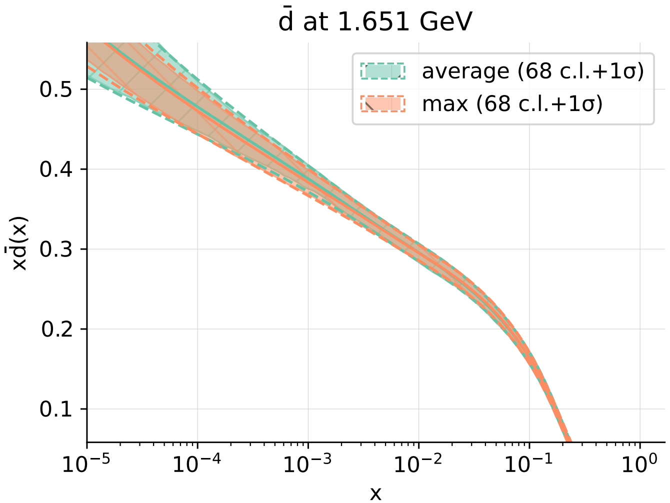
K-FOLD VALIDATION: RESULTS AND STABILITY

HYPEROPTIMIZED PARAMETERS

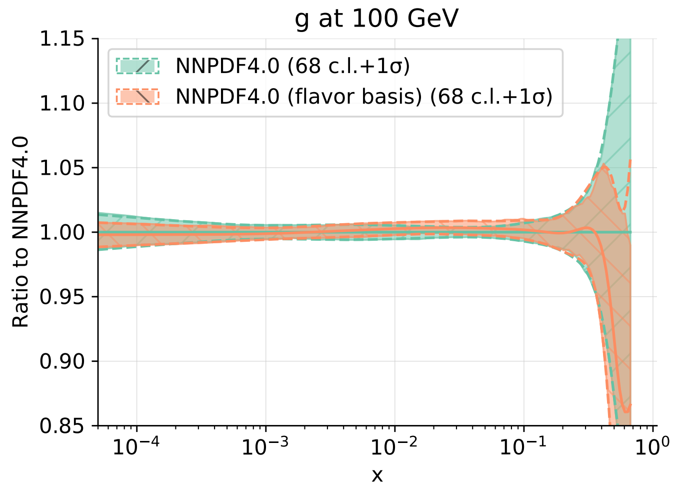
Parameter	NNPDF4.0	L as in Eq. (3.21)	Flavour basis Eq. (3.2)
Architecture	25-20-8	70-50-8	7-26-27-8
Activation function	hyperbolic tangent	hyperbolic tangent	sigmoid
Initializer	`glorot_normal`	`glorot_uniform`	`glorot_normal`
Optimizer	`Nadam`	`Adadelta`	`Nadam`
Clipnorm	6.0×10^{-6}	5.2×10^{-2}	2.3×10^{-5}
Learning rate	2.6×10^{-3}	2.5×10^{-1}	2.6×10^{-3}
Maximum # epochs	17×10^{3}	45×10^{3}	45×10^{3}
Stopping patience	10% of max epochs	12% of max epochs	16% of max epochs
Initial positivity $\Lambda^{(\mathrm{pos})}$	185	106	2
Initial integrability $\Lambda^{(\mathrm{int})}$	10	10	10

- DIFFERENT CHOICES OF LOSS: $L = \frac{1}{n_{\mathrm{fold}}}\sum_{k=1}^{n_{\mathrm{fold}}} \chi_k^2$ vs. $L = \max\left(\chi_1^2, \chi_2^2, \chi_3^2, \ldots, \chi_{n_{\mathrm{fold}}}^2\right)$
- PDF FLAVOR VS. EVOLUTION BASIS

CHANGE OF LOSS (ANTIDOWN)



CHANGE OF BASIS (GLUON)



GENERALIZATION

WIKIPEDIA
The Free Encyclopedia

Search Wikipedia

Machine learning

Article Talk

From Wikipedia, the free encyclopedia

ıtents [hide]

ɔ)

Generalization [edit]

The difference between optimization and machine learning arises from the goal of generalization: while optimization algorithms can minimize the loss on a training set, machine learning is concerned with minimizing the loss on unseen samples. Characterizing the generalization of various learning algorithms is an active topic of current research, especially for deep learning algorithms.