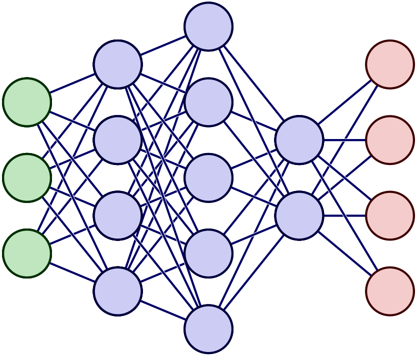
Training NNs for PDF Determinations



L Del Debbio

Higgs Centre for Theoretical Physics
University of Edinburgh

work in collaboration with

A Chiefa, R Kenway, T Giani, A Candido, G Petrillo, M Wilson

see also

- Idd et al (2022)
- Idd et al (2024)
- new, *preliminary* results

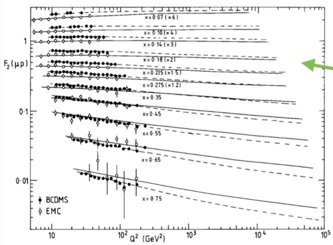
precision physics at the LHC

theoretical description of protons in terms of PDFs

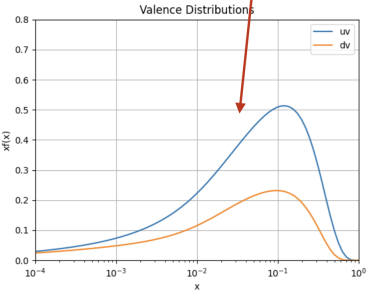
$$\text{e.g. for DIS}\,, \quad T_I = \int dx\, C_I(x) f(x)$$

- requires precise and accurate determinations of the PDFs from data
- universality of PDFs: predictions for new processes
- typical example of an inverse problem
- NNPDF methodology: focus on the training process of NNs
- useful for LGTs: extraction of spectral densities/trivializing maps (normalizing flows)

inverse problem



$$y_I = \int dx\, C_I(x) f(x)$$

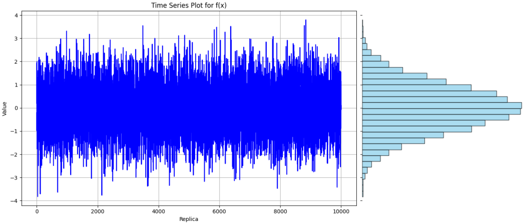


[NNPDF4.0]

parametrization of $f_i(x)$ for $x \in [0,1]$: bias vs variance of the results

Bayesian framework

- promote f to a stochastic process, $f(x)$ are stochastic variables
- **choose** a prior distribution $p(f)$ - prior knowledge about f
- probability distributions are represented by ensembles of *replicas*

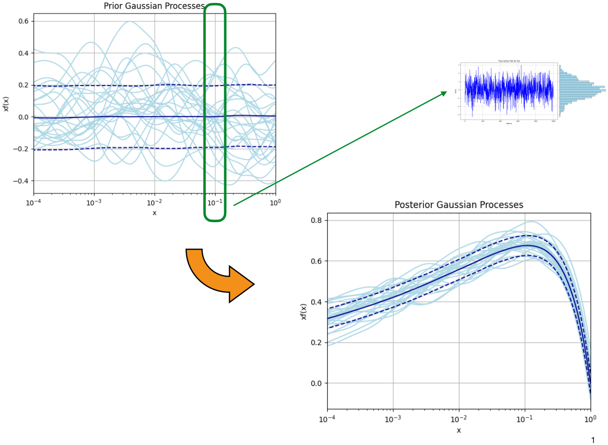


- observables are computed from averages over replicas

$$\bar{O} = E_p\left[O(f)\right] = \int df\, p(f)\, O(f) = \frac{1}{N_{\mathrm{rep}}} \sum_{k=1}^{N_{\mathrm{rep}}} O\left(f^{(k)}\right)$$

solving the inverse problem: posterior distribution

- use the data to infer a posterior probability $\tilde{p}(f)$

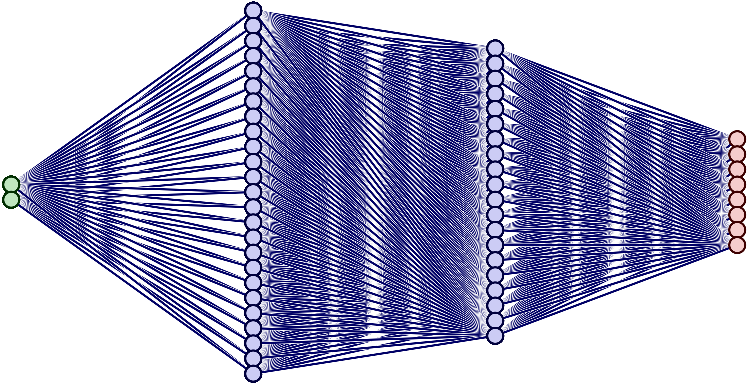


$$\bar{f}(x) = E_{\tilde{p}}\left[f(x)\right]\,, \quad \delta f(x) = E_{\tilde{p}}\left[\left(f(x) - \bar{f}(x)\right)^2\right]$$

NNPDF paradigm

- parametrize the unknown functions f using neural networks
- study a vector of values $f_\alpha = f(x_\alpha)$
- initialize N_{rep} NN replicas
- train each NN on a replica of the dataset
- training/validation split to determine the stopping condition
- the ensemble of trained NNs provides the posterior distribution

NNPDF parametrization - conventions



$$\rho_i^{(\ell)} = \rho(\phi_i^{(\ell)})\,, \quad \phi_i^{(\ell)} = \sum_{j=1}^{n_{\ell-1}} w_{ij}^{(\ell)} \rho_j^{(\ell-1)} + b_i^{(\ell)}$$

$$\boxed{f(x) = \phi^{(L)}(x;\theta)}$$

finite-dimensional, sufficiently large to be unbiased

NN prior distribution

parameters θ are initialized using a Glorot-Normal distribution

initialize weights and biases using Gaussians

$$\langle b_i^{(\ell)} \rangle = 0 \,, \quad \langle b_{i_1}^{(\ell)} b_{i_2}^{(\ell)} \rangle = \delta_{i_1 i_2} C_b^{(\ell)}$$

$$\langle w_{ij}^{(\ell)} \rangle = 0 \,, \quad \langle w_{i_1 j_1}^{(\ell)} w_{i_2 j_2}^{(\ell)} \rangle = \delta_{i_1 i_2} \delta_{j_1 j_2} \frac{C_w^{(\ell)}}{n_{\ell-1}}$$

parameters/functions duality

$$p(\phi^{(\ell)}) = \int [dw\, p(w)]\, [db\, p(b)]\, \prod_{i,\alpha} \delta \left(\phi_{i\alpha}^{(\ell)} - \sum_j w_{ij}^{(\ell)} \rho \left(\phi_{j\alpha}^{(\ell-1)} \right) - b_i^{(\ell)} \right)$$

EFT approach

symmetry and $1/n$ counting yield the probability distribution for ϕ

$$
\begin{aligned}
p(\phi^{(\ell)}) &= \frac{1}{Z}\,\exp\left[-S(\phi^{(\ell)})\right] \\
&= \frac{1}{Z}\,\exp\Bigg[-\frac{1}{2}\gamma^{(\ell)}_{\alpha_1\alpha_2}\phi^{(\ell)}_{\alpha_1}\cdot\phi^{(\ell)}_{\alpha_2} \\
&\qquad\qquad -\frac{1}{8n_{\ell-1}}\gamma^{(\ell)}_{\alpha_1\alpha_2,\alpha_3\alpha_4}\phi^{(\ell)}_{\alpha_1}\cdot\phi^{(\ell)}_{\alpha_2}\,\phi^{(\ell)}_{\alpha_3}\cdot\phi^{(\ell)}_{\alpha_4} + O(1/n^2_{\ell-1})\Bigg]
\end{aligned}
$$

correlators can be computed using Feynman diagrams

$$
\begin{aligned}
\langle\phi^{(\ell)}_{i_1,\alpha_1}\phi^{(\ell)}_{i_2,\alpha_2}\rangle &= \delta_{i_1 i_2}K^{(\ell)}_{\alpha_1\alpha_2} + O(1/n_{\ell-1}) \\
\langle\phi^{(\ell)}_{i_1,\alpha_1}\phi^{(\ell)}_{i_2,\alpha_2}\phi^{(\ell)}_{i_3,\alpha_3}\phi^{(\ell)}_{i_4,\alpha_4}\rangle_c &= O(1/n_{\ell-1})
\end{aligned}
$$

ϕ are approximately Gaussian processes, corrections are $O(1/n)$

training

for all parametrizations and for a quadratic loss

$$\mathcal{L}_t = \frac{1}{2}\left(y - T[f_t]\right)^T C_Y^{-1} \underbrace{\left(y - T[f_t]\right)}_{\epsilon_t}$$

gradient descent

$$\begin{aligned} \frac{d}{dt}\theta_\mu &= -\nabla_\mu \mathcal{L} \\ \frac{d}{dt} f_t &= (\nabla_\mu f_t) \frac{d}{dt}\theta_\mu = \Theta_t \left(\frac{\partial T}{\partial f}\right)_t C_Y^{-1} \epsilon_t \end{aligned}$$

where

$$\Theta_t = (\nabla_\mu f_t)(\nabla_\mu f_t)^T$$

is the Neural Tangent Kernel (NTK)

linear data & NN parametrization

for linear data

$$T = (\mathrm{FK}) f \quad \Longrightarrow \quad \left(\frac{\partial T}{\partial f}\right) = (\mathrm{FK})$$

for wide neural networks

$$\boxed{\Theta_t = \Theta + O(1/n) \quad \text{(lazy training)}}$$

hence we obtain a linear equation for f_t

$$\begin{aligned} \frac{d}{dt} f_t &= \Theta (\mathrm{FK})^T C_Y^{-1} \left(y - (\mathrm{FK}) f_t\right) \\ &= -\Theta M f_t + b \end{aligned}$$

where $M = (\mathrm{FK})^T C_Y^{-1} (\mathrm{FK})$

flat directions & NTK spectrum

eigenvalues of the NTK

$$\Theta z^{(k)} = \lambda^{(k)} z^{(k)} \,, \quad f_{t,k} = \left(z^{(k)}, f_t \right)$$

evolution equations in the NTK eigenbasis

$$\frac{d}{dt} f_{t,k} = \lambda^{(k)} \left(z^{(k)}, (\mathrm{FK})^T C_Y^{-1} \left(y - (\mathrm{FK}) f_t \right) \right)$$

NTK kernel: directions that do NOT evolve during training

$$f_{t,\parallel} = f_{0,\parallel}$$

no bias, but irreducible noise dictated by the prior

analytic solution

for each replica:

$$\boxed{f_t = U(t)f_0 + V(t)Y}$$

and therefore we have analytic expressions for

$$\bar{f}_t = \mathbb{E}[U(t)f_0] + \mathbb{E}[V(t)Y]$$

$$\mathrm{Cov}[f_t, f_t^T] = C_t^{(00)} + C_t^{(0Y)} + C_t^{(YY)}$$

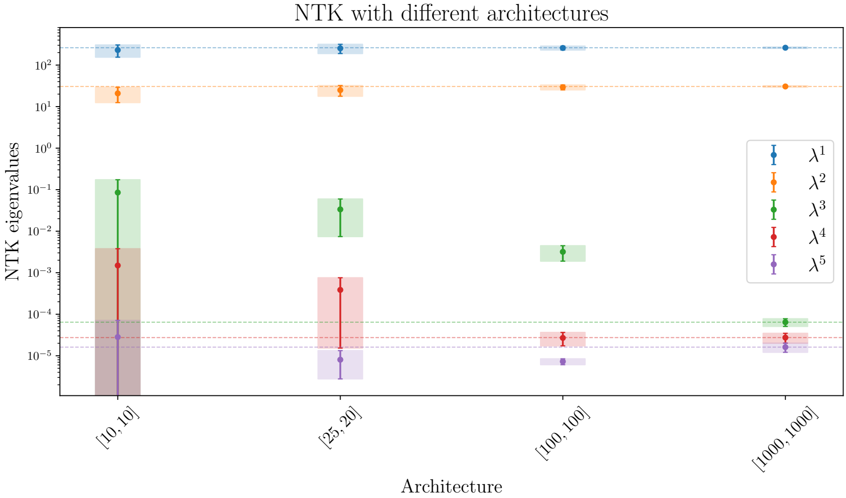
phenomenology

- for each replica $k = 1, \ldots, N_{\mathrm{rep}}$
- initialize a NN
- generate synthetic data using a known f^{in}

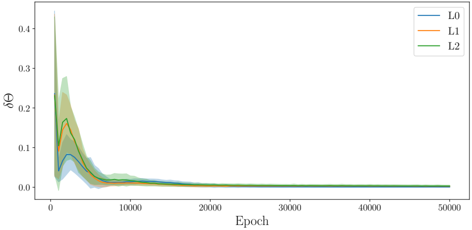
$$Y^{(k)} = \underbrace{(\mathrm{FK}) f^{\mathrm{in}}}_{L_0} + \underbrace{\eta}_{L_1} + \underbrace{\xi^{(k)}}_{L_2}$$

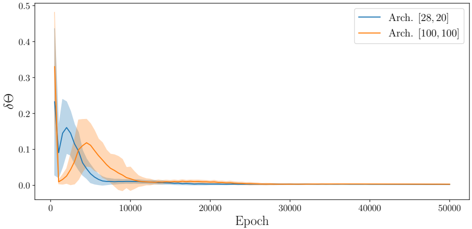
- train numerically and compare with the analytic expressions

NTK at initialization

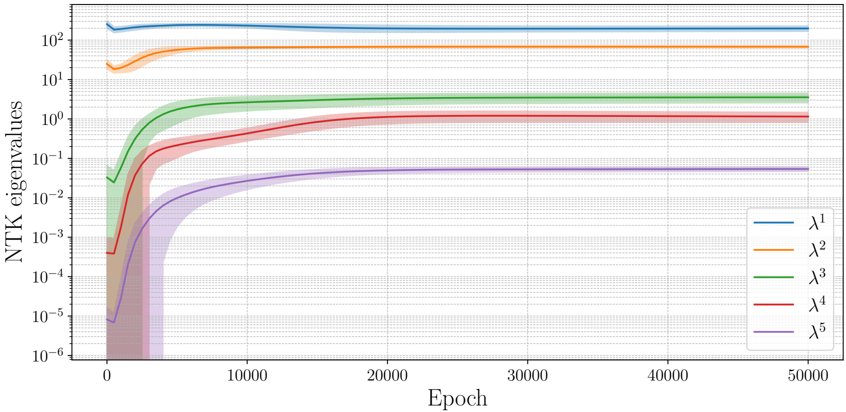


NTK during training

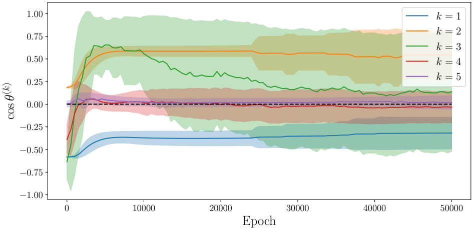


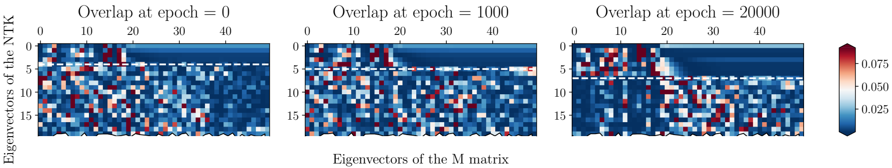


NTK alignment

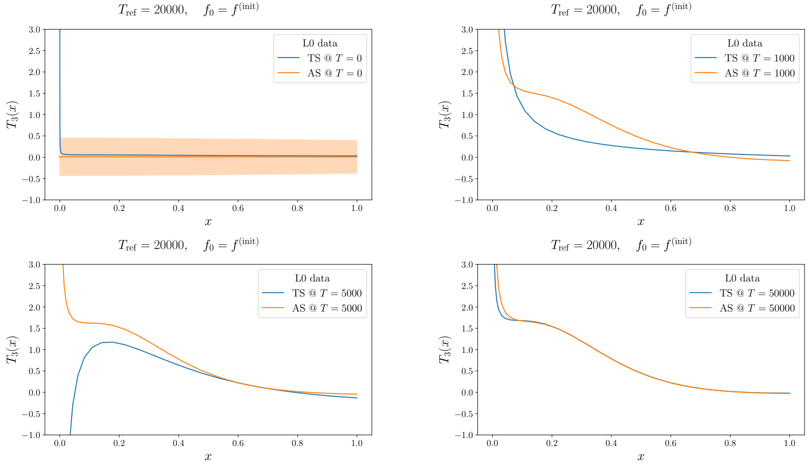


NTK alignment

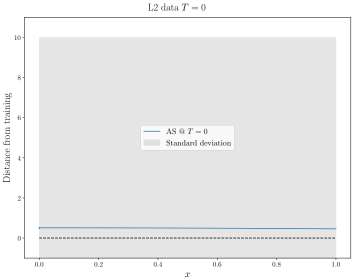


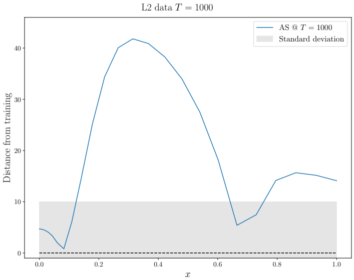


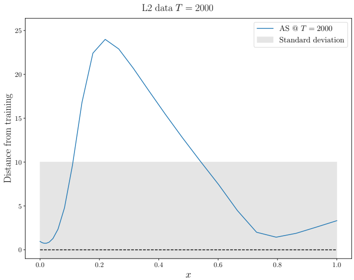
evolution with frozen NTK

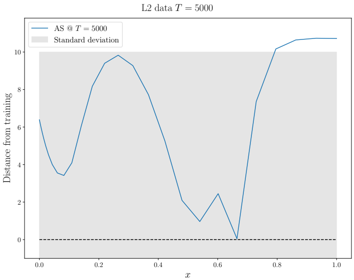


distances

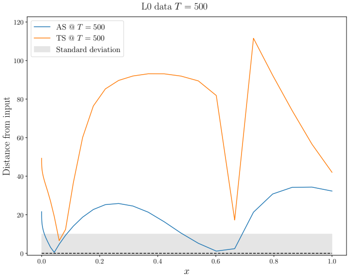

L2 data $T = 0$
AS @ $T = 0$
Standard deviation
Distance from training
x

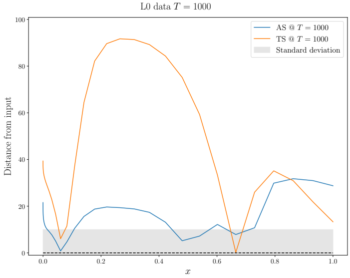

L2 data $T = 1000$
AS @ $T = 1000$
Standard deviation
Distance from training
x

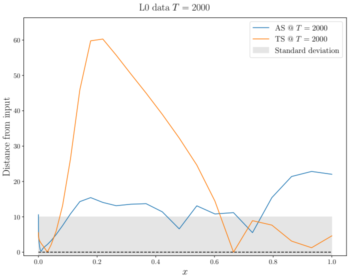

L2 data $T = 2000$
AS @ $T = 2000$
Standard deviation
Distance from training
x

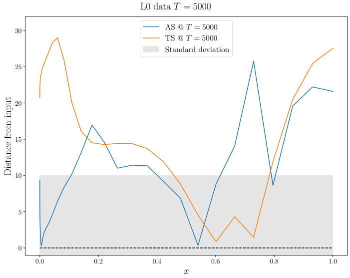

L2 data $T = 5000$
AS @ $T = 5000$
Standard deviation
Distance from training
x

distances

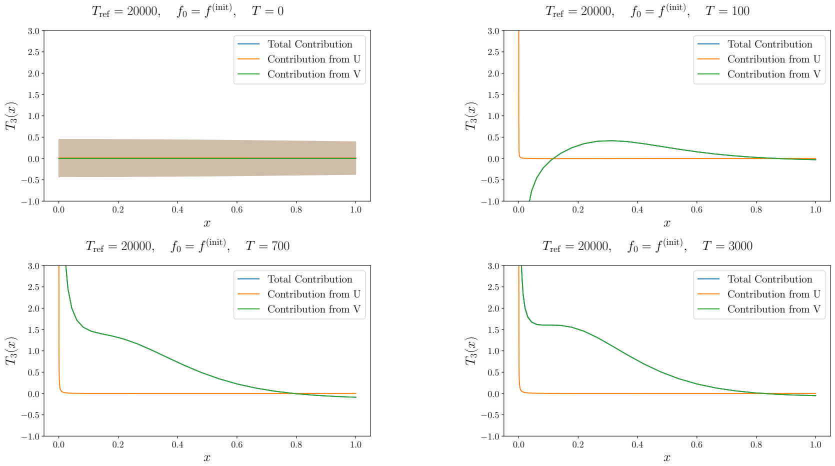

L0 data $T = 500$
AS @ $T = 500$
TS @ $T = 500$
Standard deviation
Distance from input
x


L0 data $T = 1000$
AS @ $T = 1000$
TS @ $T = 1000$
Standard deviation
Distance from input
x


L0 data $T = 2000$
AS @ $T = 2000$
TS @ $T = 2000$
Standard deviation
Distance from input
x


L0 data $T = 5000$
AS @ $T = 5000$
TS @ $T = 5000$
Standard deviation
Distance from input
x

U and V contributions



outlook

- PDFs are a central ingredient for LHC analyses
- Bayesian approach is convenient to solve inverse problems
- all hypotheses are clearly spelled in the prior
- analytical control of the training process is key to quantify the bias/variance budget
- same methods can be applied to normalizing flows?