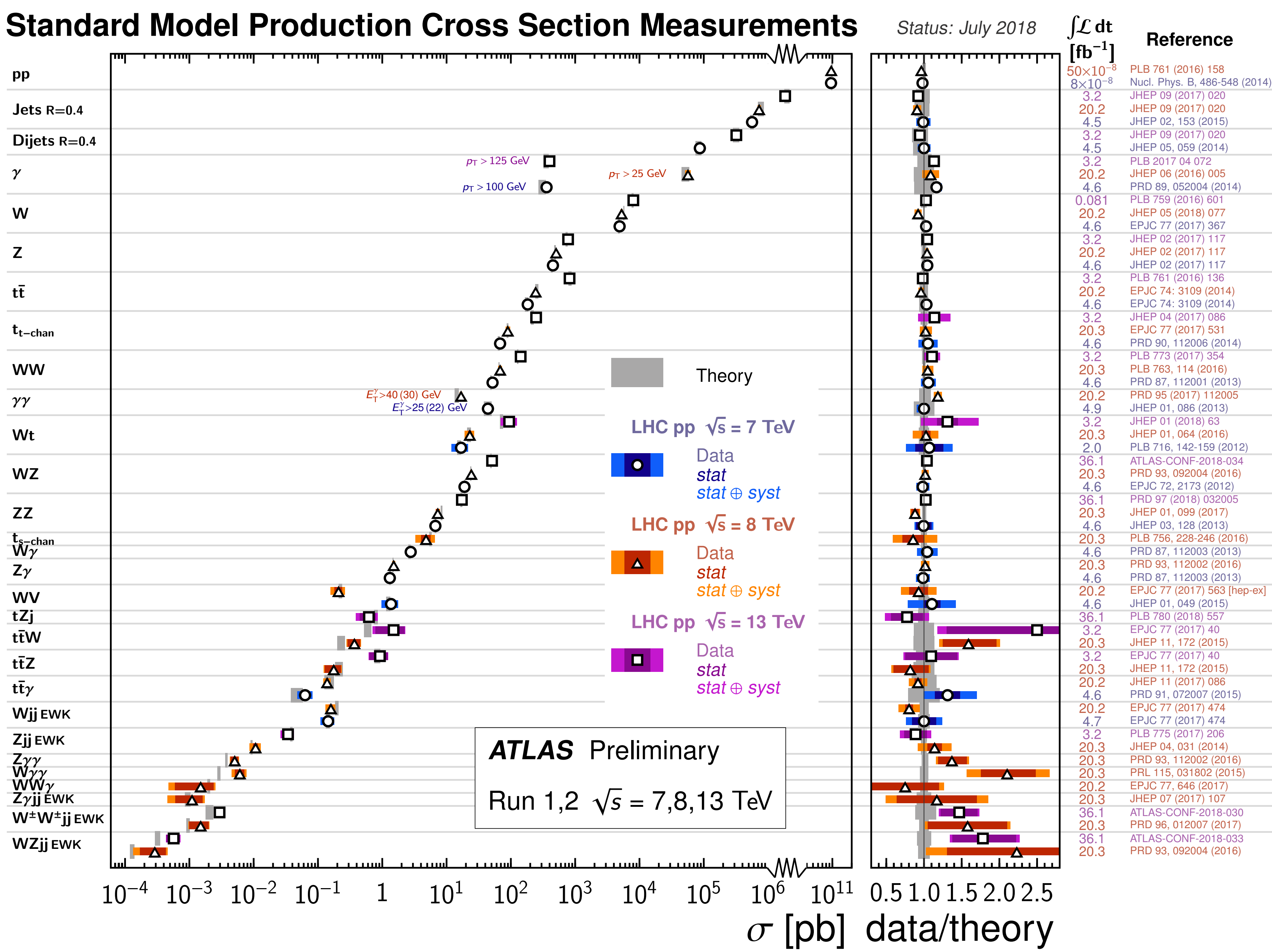
Bayesian Approach to Inverse Problems

L Del Debbio

Higgs Centre for Theoretical Physics
The University of Edinburgh

Precision Frontier at the LHC

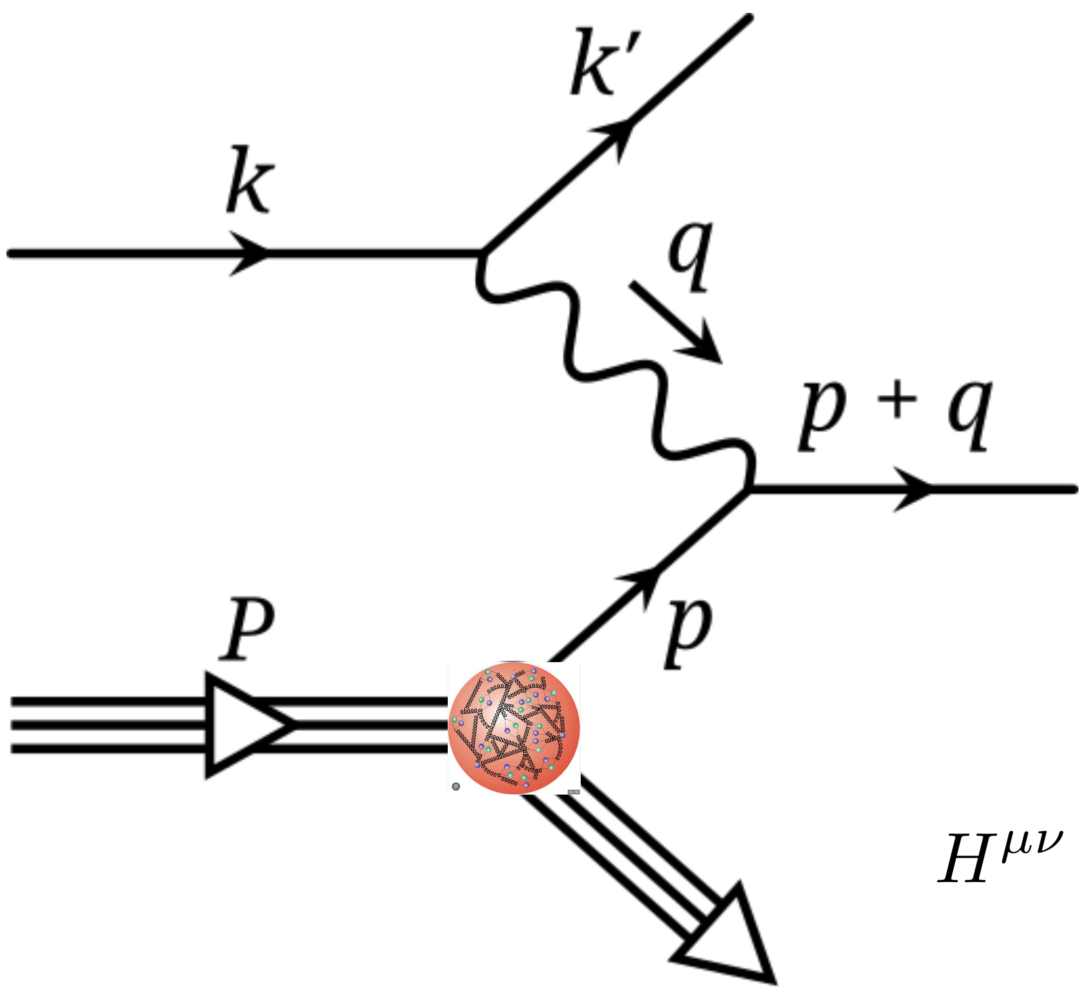


PDFs in the LHC era

- Parton Distribution Functions (PDFs) describe the nonperturbative structure of nucleons
- Predictions of any physics observable at the LHC require a “precise” knowledge of PDFs
- PDFs are amongst the dominant uncertainties for the determination of SM parameters (W mass, EW mixing angle, Higgs production, alphas)
- … and we cannot neglect the correlations between PDFs and these parameters [Forte & Kassabov 20]
- Discovery of new physics depends critically on this knowledge - what is 5 sigma?

Deep Inelastic Scattering

- lepton-nucleon scattering



$$d\Gamma \propto L_{\mu\nu} H^{\mu\nu}\, d\Phi$$

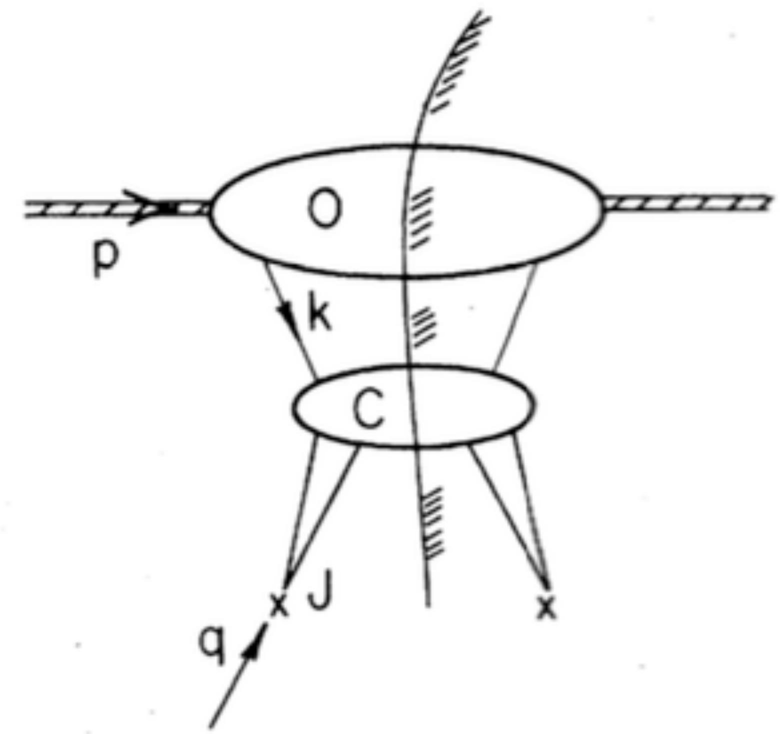
$$H^{\mu\nu} = \int d^D y\, e^{iq\cdot y}\, \langle P | J^\mu(y) J^\nu(0) | P \rangle$$

non perturbative physics

$$H^{\mu\nu} = F_1(x, Q^2)\left(\frac{q^\mu q^\nu}{q^2} - g^{\mu\nu}\right) + F_2(x, Q^2)\dots$$

measured by experiments

Factorization & PDFs



measured by experiments

computed in perturbation theory

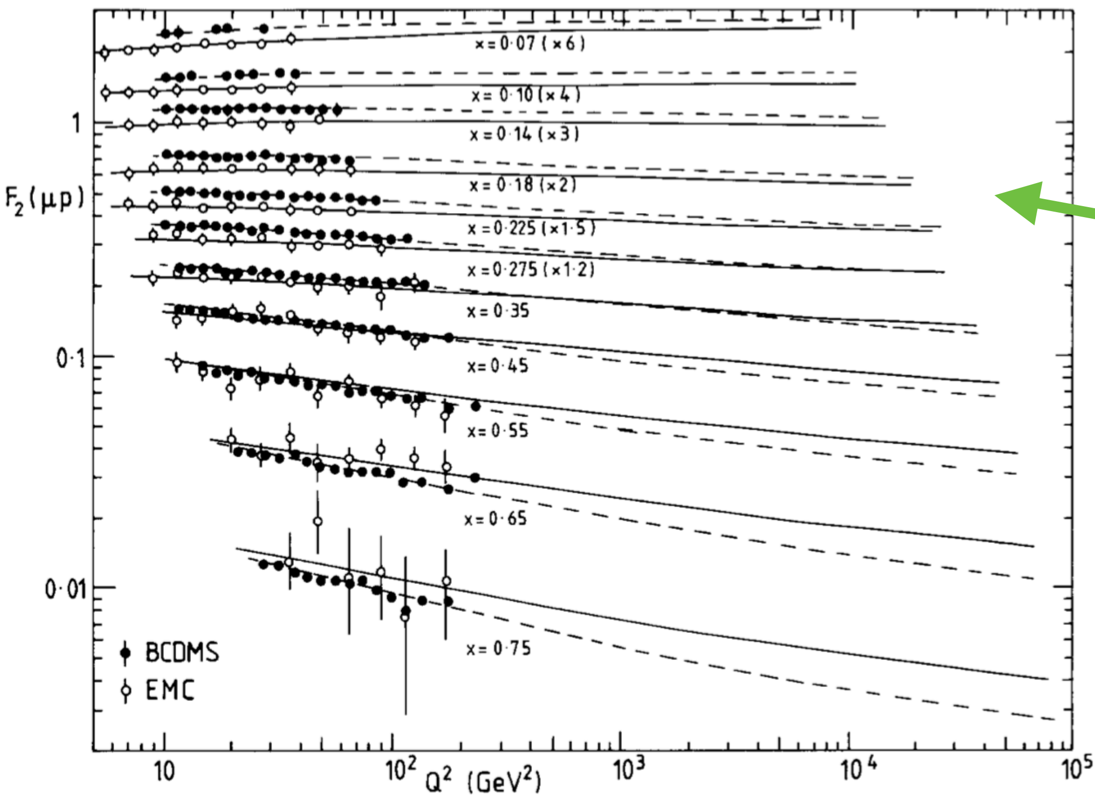
$$F_i(x,Q^2) = \int_x^1 \frac{d\xi}{\xi} C_i(\xi, Q^2, \mu^2) f_R(\frac{x}{\xi}, \mu^2) + \mathcal{O}\left(\frac{1}{Q^2}\right)$$

$$f(x) = \int \frac{dz^-}{2\pi} \exp\left(i(xP^+)z^-\right) \langle P|\bar\psi\left(-\frac{z^-}{2}\right)\Gamma\lambda_A \mathcal{U}\psi\left(\frac{z^-}{2}\right)|P\rangle$$

universal: encodes the non perturbative dynamics of the nucleon

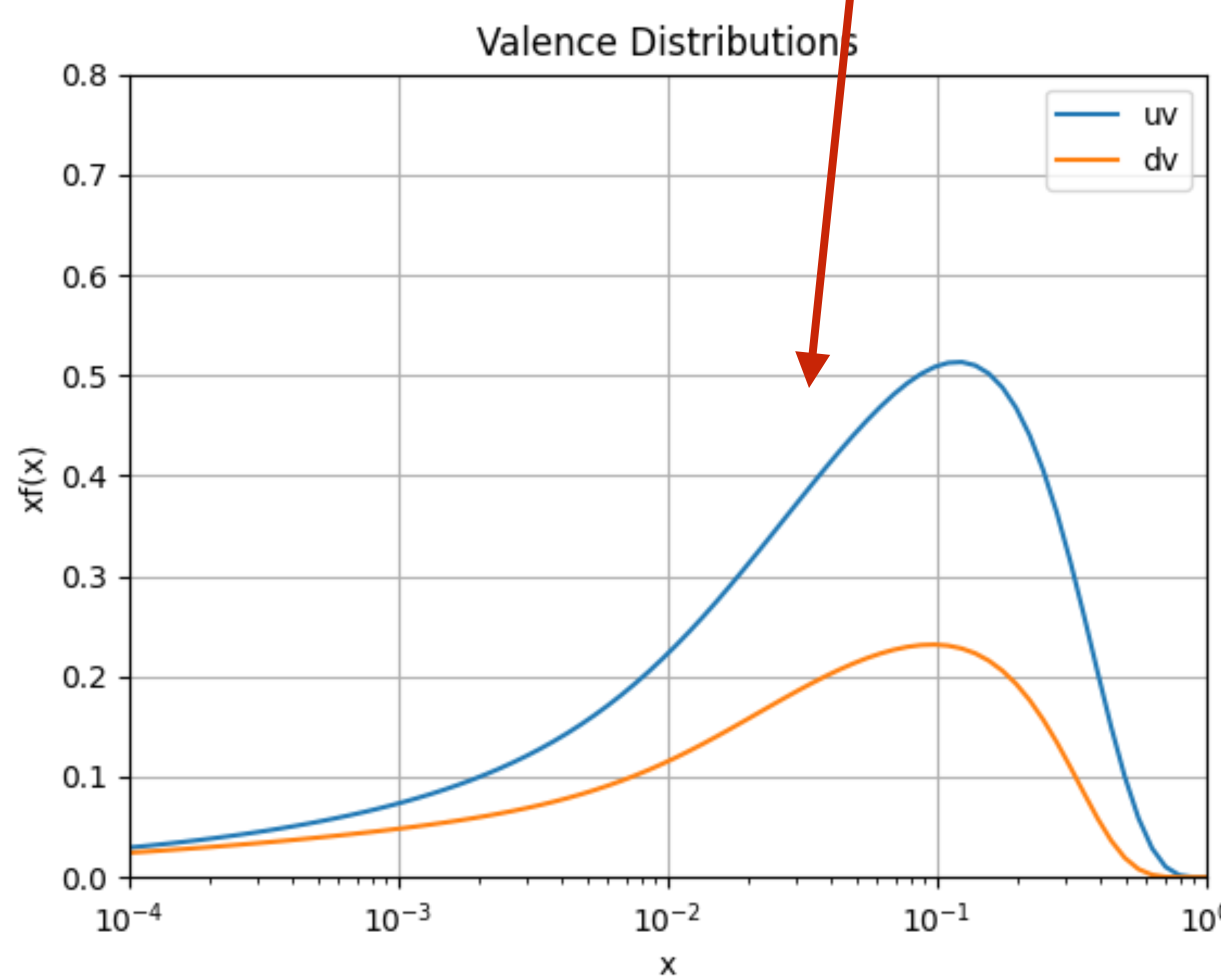
- Lattice pseudo-PDF, quasi-PDF, Ioffe-time correlators are just observables

PDF Determination & Inverse Problems



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

[NNPDF4.0]



- Trying to determine a (continuous) function

- … using a finite set of experimental data

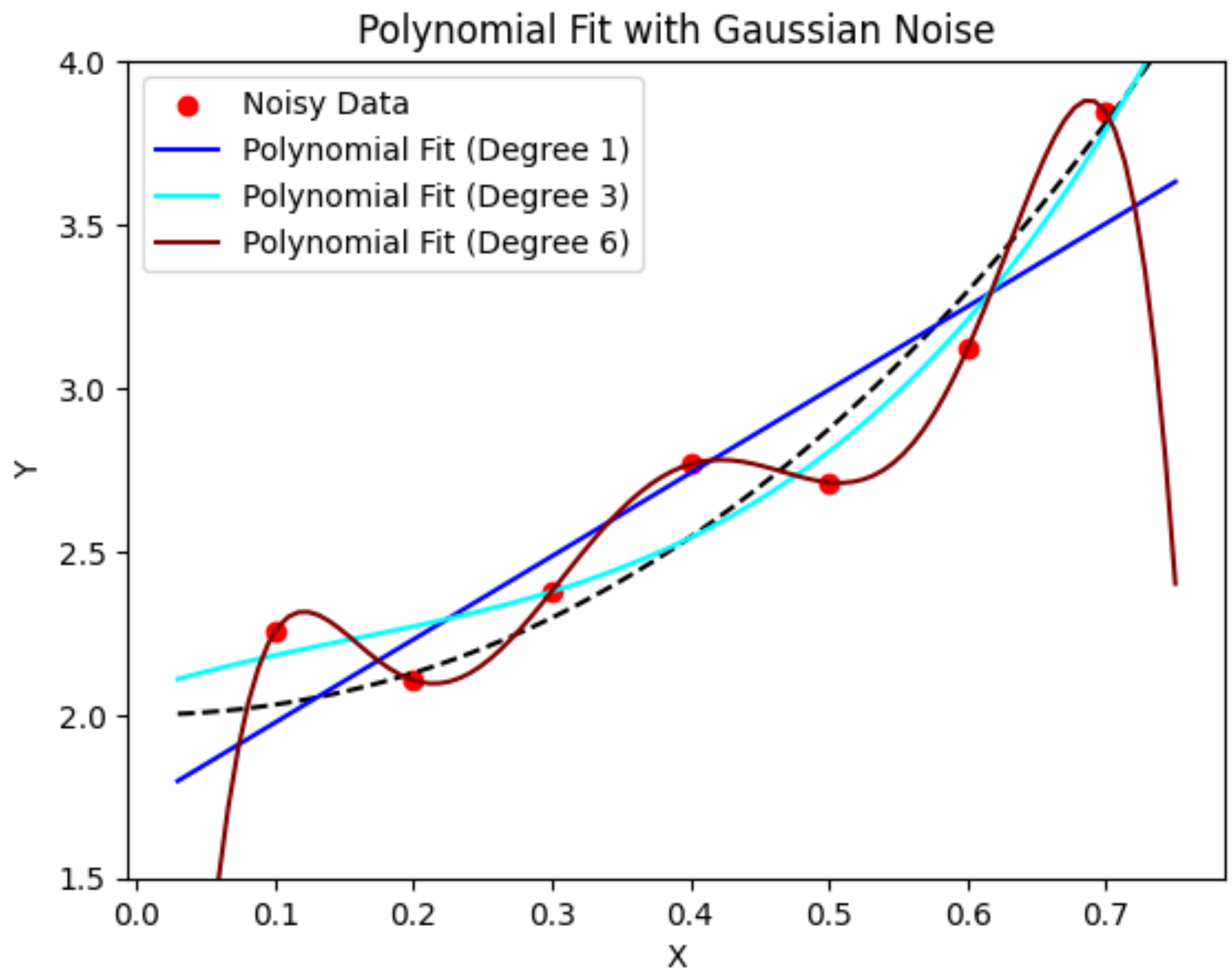
ill-posed problem

- Solution depends on the assumptions that are made

- Central values and covariances are both affected

- Bias/Variance trade-off

- Overfitting/underfitting, robust extrapolations



Bayesian Approach

- f is promoted to be a *stochastic process*
- $f(x)$ for $x \in \mathcal{I}$ is a set of stochastic variables
- for any given $\mathbf{f}$, where $f_i = f(x_i)$, we have a prior $p(\mathbf{f})$
- all a priori knowledge about f is encoded in p (more later)
- posterior distribution obtained from Bayes theorem

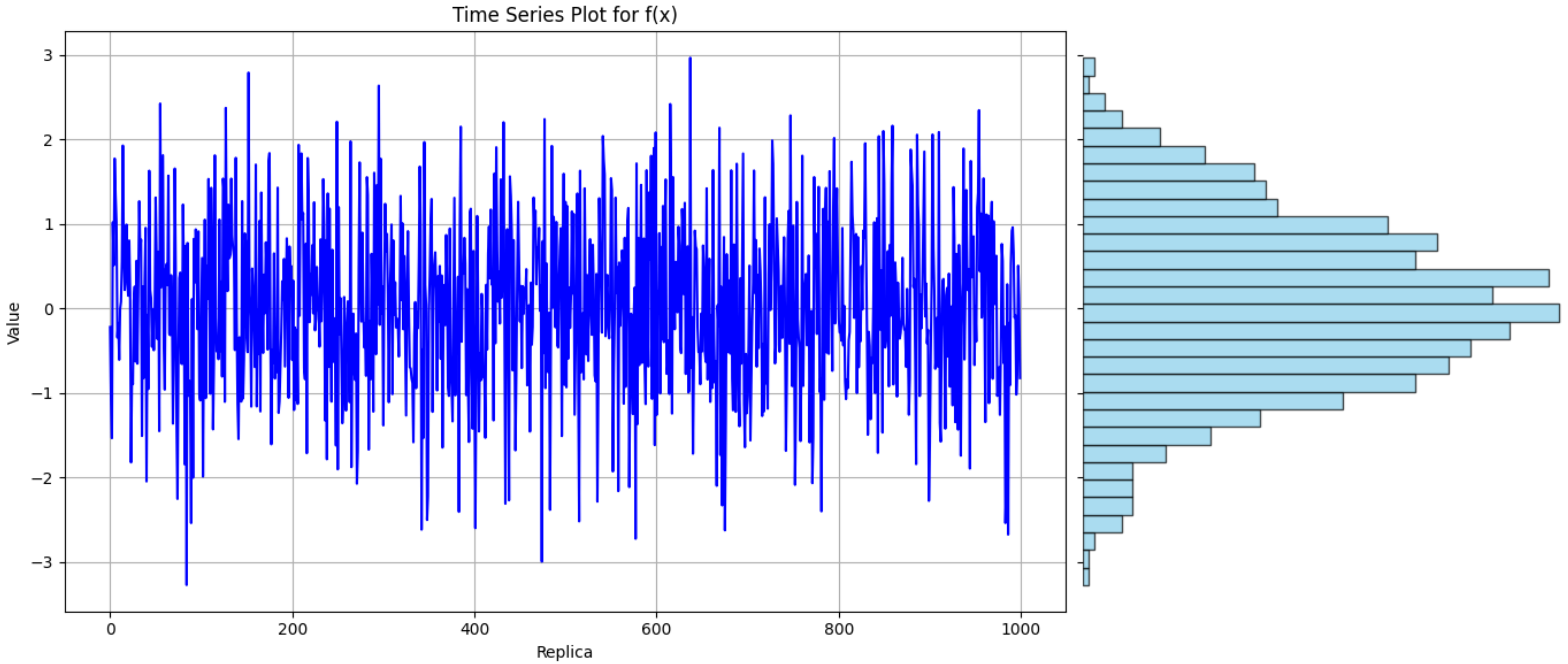
$$\tilde{p}(\mathbf{f}) = p(\mathbf{f}|y) = \frac{p(y|\mathbf{f})p(\mathbf{f})}{p(y)}$$

- knowledge about the solution is encoded in the posterior, eg

$$\text{central value} : E_{\tilde{p}}[\mathbf{f}]$$

$$\text{covariance} : \mathrm{Cov}_{\tilde{p}}[\mathbf{f}, \mathbf{f}']$$

- Probability distributions are represented by ensembles of *replicas*



Gaussian Processes

GPs are a specific kind of stochastic process

$$f \sim \mathcal{GP}\left(m, k\right) \ ,$$

where

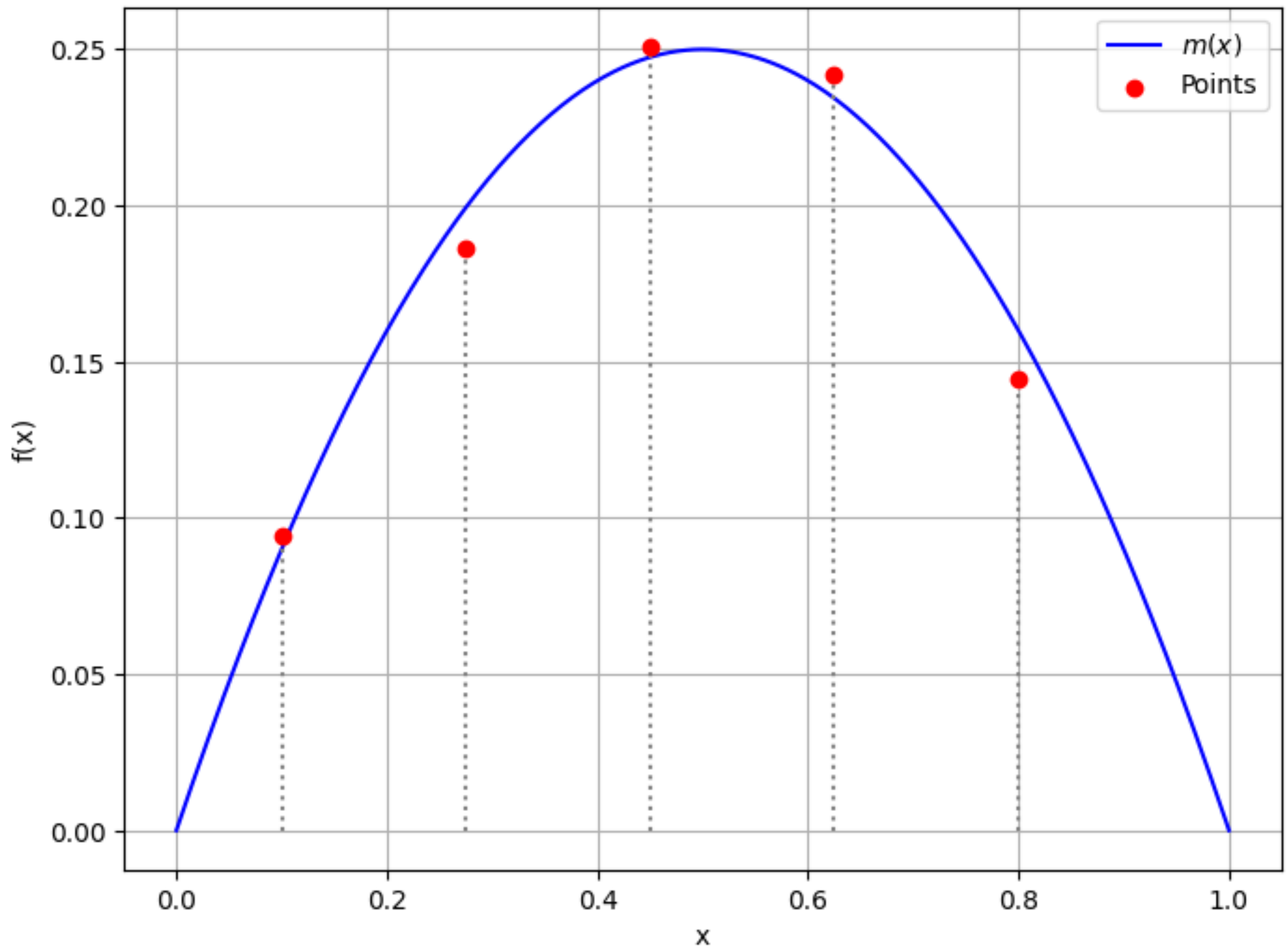
$$m : \mathcal{I} \rightarrow \mathbb{R} \, , \quad k : \mathcal{I} \times \mathcal{I} \rightarrow \mathbb{R}$$

for a GP, the vector of stochastic variables $\mathbf{f}$

$$\mathbf{x} = \{x_i ; i = 1, \ldots, N\} \, , \quad \mathbf{f} = f(\mathbf{x}) = \begin{pmatrix} f_1 \\ \vdots \\ f_N \end{pmatrix} \in \mathbb{R}^N \, , \quad f_i = f(x_i)$$

is distributed as a multidimensional Gaussian

$$\mathbf{f} \sim \mathcal{N}(\mathbf{m}, K) \, ,$$



$m(x)$
Points
0.25
0.20
0.15
0.10
0.05
0.00
f(x)
0.0
0.2
0.4
0.6
0.8
1.0
x

Gaussian Process Prior Distribution

mean & covariance

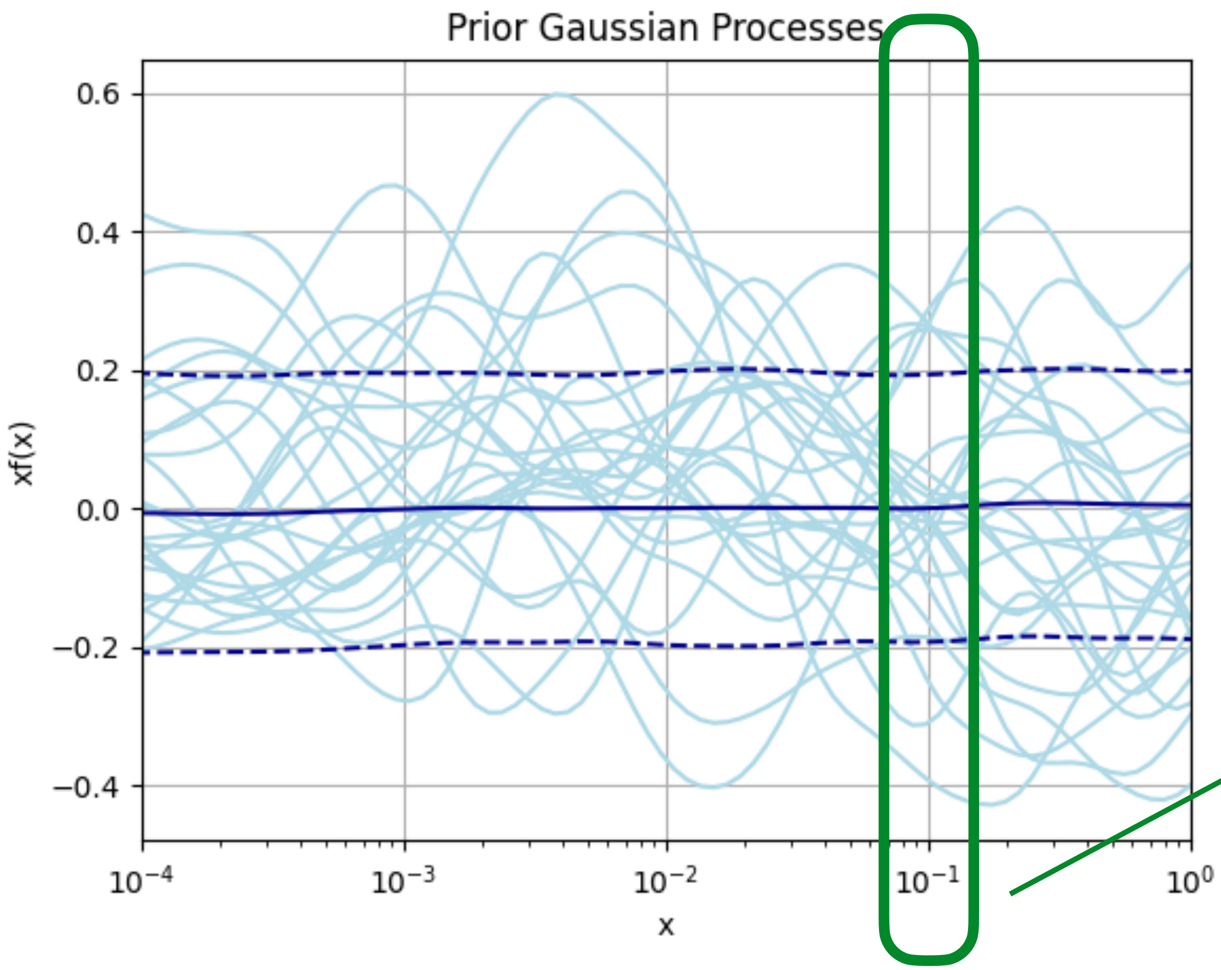
$$\mathbf{m} = m(\mathbf{x})\,, \quad K = k(\mathbf{x}, \mathbf{x}^T)\,,$$

$$E[f_i] = m_i = m(x_i)\,,$$
$$\mathrm{Cov}[f_i, f_j] = K_{ij} = k(x_i, x_j)\,.$$

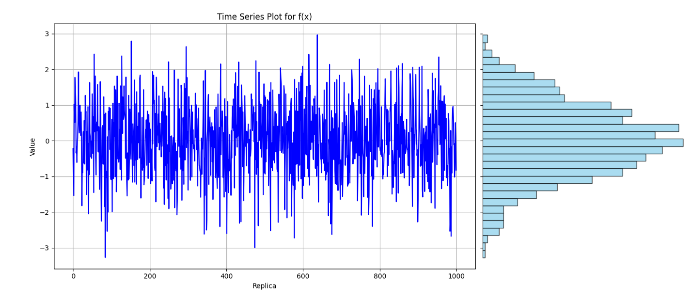
specific choices for this work: zero mean and Gibbs kernel

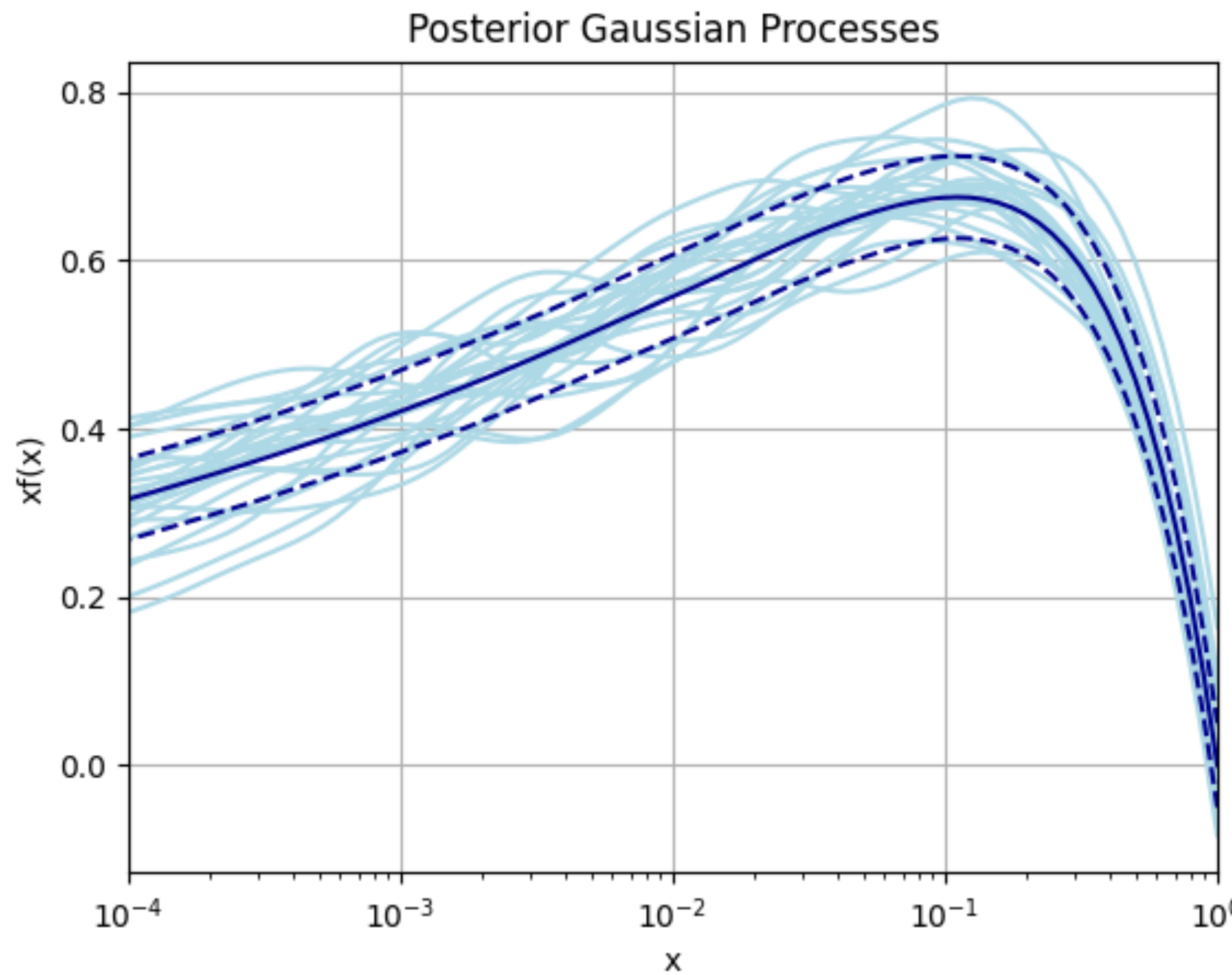
$$m(x) = 0$$
$$k(x, x') = \sigma^2 \sqrt{\frac{2l\left(x\right)l\left(x'\right)}{l\left(x\right)^2 + l\left(x'\right)^2}} \exp\left[-\frac{\left(x - x'\right)^2}{l\left(x\right)^2 + l\left(x'\right)^2}\right]$$

hyperparameters


Prior Gaussian Processes
0.6
0.4
0.2
0.0
−0.2
−0.4
xf(x)
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
x

$$l(x) = l_0(x + \delta)$$




Posterior Gaussian Processes
0.8
0.6
0.4
0.2
0.0
xf(x)
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
x

Data and Likelihood

dataset central values: $\mathbf{y} = \{y_I, I = 1, \ldots, N_{\rm dat}\}$

dataset fluctuations: $\epsilon \sim \mathcal{N}\left(0, C_Y\right)$

linear dependence on f:

$$T_I = \int_{\mathcal{I}} dx\, C_I(x) f(x) \approx \sum_{i=1}^{N} ({\rm FK})_{Ii} f_i$$

NB: applies to both quasi/pseudo-PDFs and spectral densities

$$E[T_I] = ({\rm FK})_{Ij} m_j$$

$$\mathrm{Cov}[T_I, T_J] = ({\rm FK})_{Ii} \left(K_{\mathbf{xx}}\right)_{ij} ({\rm FK})^T_{jJ}$$

Posterior Distribution

we want to determine

$$\tilde{p}(\mathbf{f}, \mathbf{f}^*) = p(\mathbf{f}, \mathbf{f}^*|y) = \int d\theta \, p\left(\mathbf{f}, \mathbf{f}^*, \theta|y\right)$$

$$p\left(\mathbf{f}, \mathbf{f}^*, \theta|y\right) = p\left(\mathbf{f}, \mathbf{f}^*|\theta, y\right) p\left(\theta|y\right)$$

compute each factor independently

$$p(\mathbf{f}, \mathbf{f}^*|\theta, y) \propto \exp\left\{-\frac{1}{2}\left((\mathbf{f}-\mathbf{m})^T, (\mathbf{f}^*-\mathbf{m}^*)^T\right) K^{-1} \begin{pmatrix} \mathbf{f}-\mathbf{m} \\ \mathbf{f}^*-\mathbf{m}^* \end{pmatrix}\right\}$$

$$\times \, \exp\left\{-\frac{1}{2}((\mathrm{FK})\mathbf{f}-y)^T C_Y^{-1}((\mathrm{FK})\mathbf{f}-y)\right\} .$$

Posterior distribution

$$\tilde{p}(\mathbf{f}) = p(\mathbf{f}|Y) = \int d\theta\, p(\mathbf{f},\theta|Y)$$

$$p(\mathbf{f},\theta|Y) = p(\mathbf{f}|\theta,Y)\,p(\theta|Y)$$

- Compute each factor independently

$$p(\mathbf{f}|\theta,Y) \propto \exp\left\{-\frac{1}{2}(\mathbf{f}-\mathbf{m})^T K(\theta)^{-1}(\mathbf{f}-\mathbf{m})\right\}$$
$$\times \exp\left\{-\frac{1}{2}((\mathrm{FK})\mathbf{f}-Y)^T C_Y^{-1}((\mathrm{FK})\mathbf{f}-Y)\right\}$$

Posterior for fixed hyper parameters

- Posterior in this case is Gaussian

$$\mathbf{f}|\theta, Y \sim \mathcal{N}(\tilde{\mathbf{m}}, \tilde{\mathbf{K}})$$

where

$$\tilde{\mathbf{m}} = \tilde{K}\left[(\mathrm{FK})^T C_Y^{-1} Y + K^{-1}\mathbf{m}\right]$$
$$\tilde{K}^{-1} = (\mathrm{FK})^T C_Y^{-1} (\mathrm{FK}) + K^{-1}$$

- Explicit dependence on the prior

Interpretation of the result: closure test

vanishing exp errors

$$y = y_0 = (\mathrm{FK})\mathbf{f}_0\,, \quad C_Y = 0$$

yields

$$\tilde{\mathbf{m}} = R^{(0)}_{\mathbf{xx}}\,\mathbf{f}_0\,, \quad \tilde{\mathbf{m}}^* = R^{(0)}_{\mathbf{x}^*\mathbf{x}}\,\mathbf{f}_0$$

where we introduced the smearing kernel

$$R^{(0)}_{\mathbf{xx}} = K_{\mathbf{xx}}(\mathrm{FK})^T\left[(\mathrm{FK})K_{\mathbf{xx}}(\mathrm{FK})^T\right]^{-1}(\mathrm{FK})$$

the result of Bayesian inference is a smeared version of the 'true' answer

$$\tilde{\mathbf{m}} - \mathbf{f}_0 = \left[R^{(0)}_{\mathbf{xx}} - \mathbb{1}\right]\mathbf{f}_0\,, \quad \tilde{K}_{\mathbf{xx}} = \left(\mathbb{1} - R^{(0)}_{\mathbf{xx}}\right)K_{\mathbf{xx}}$$

Inference for hyperparameters

using Bayes theorem

another prior

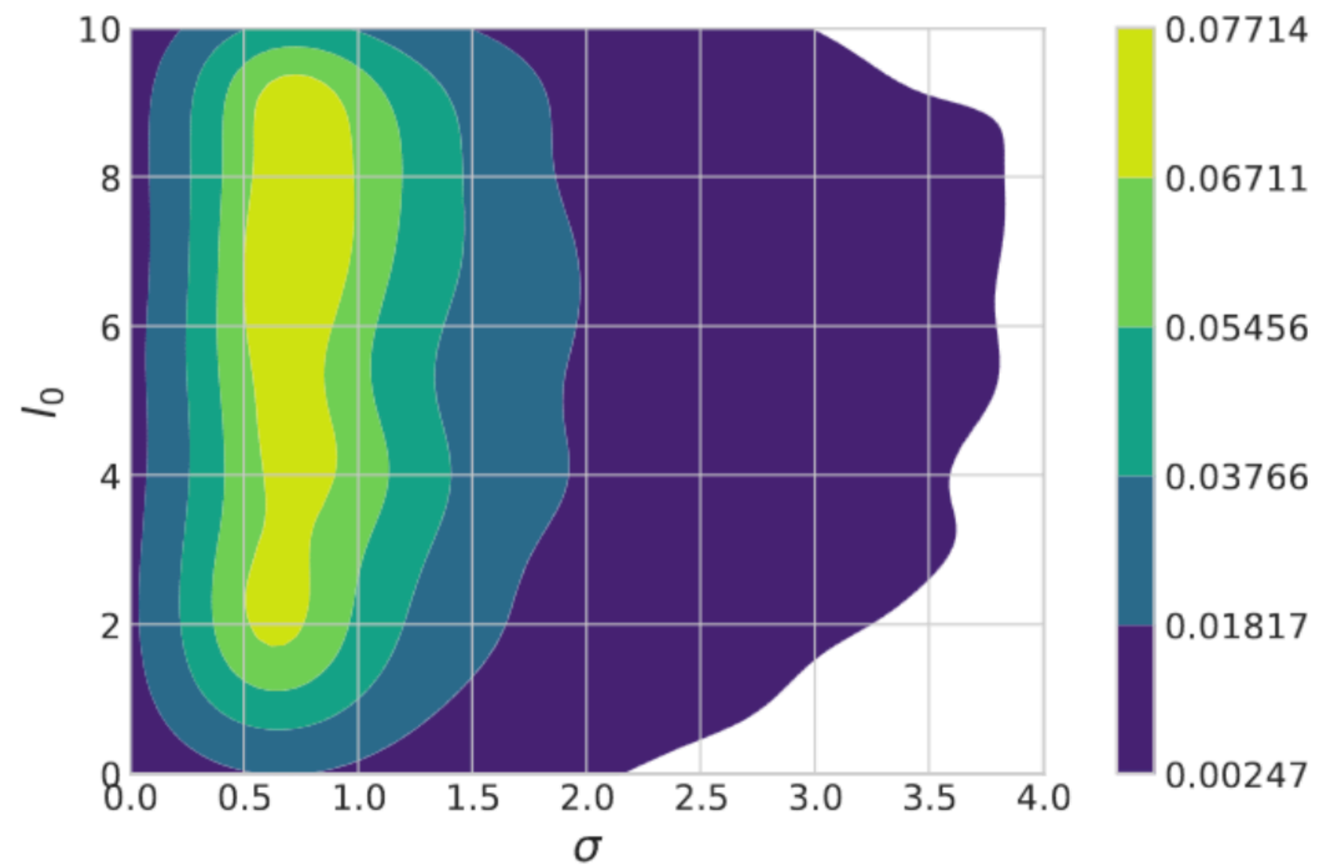
$$p(\theta|y) = \frac{p(y|\theta)\, p_\theta(\theta)}{\int d\theta\, p(y|\theta)\, p_\theta(\theta)},$$

on the RHS

$$p(y|\theta) = \frac{e^{-\frac{1}{2}(y-(\mathrm{FK})\mathbf{m})^T C_{YT}^{-1}(y-(\mathrm{FK})\mathbf{m})}}{\sqrt{\det\left[2\pi C_{YT}\right]}}.$$

$p(\theta|y)$ can be sampled by MCMC

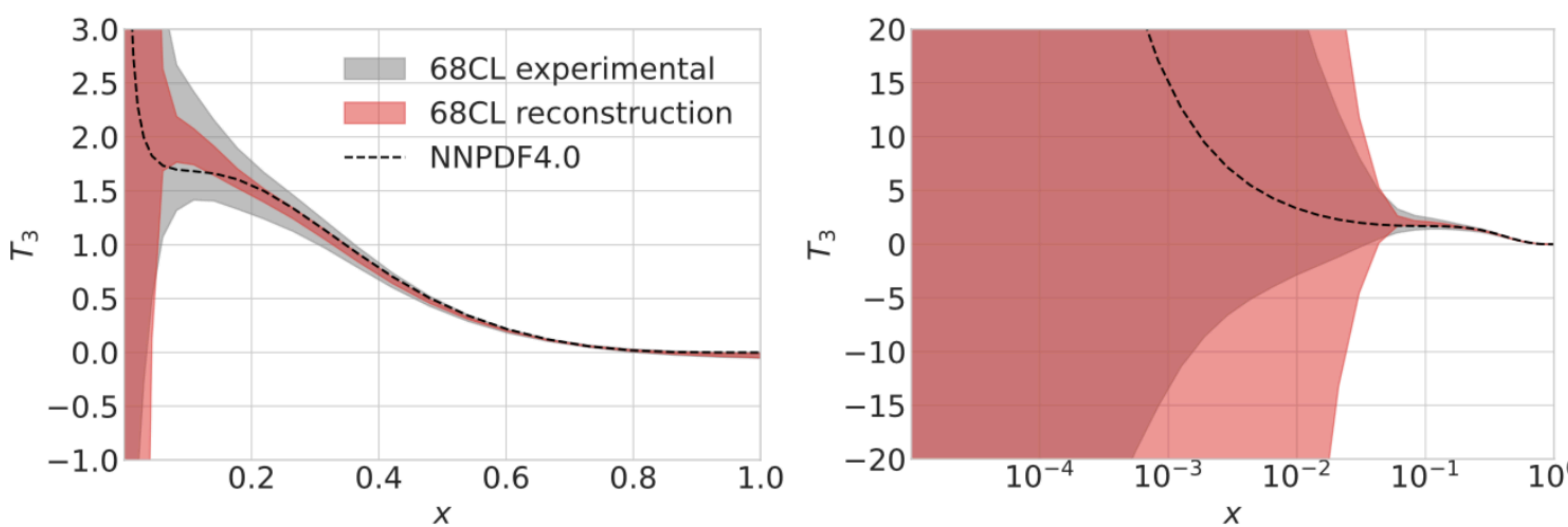
starting from **flat** priors for the hyperparameters, we get for $p(\theta|y)$



and $p(\mathbf{f}^*|\theta, y)$ is known analytically

Putting both factors together

- limited reconstruction due to smearing, functional uncertainty
- functional uncertainty is not cured by more precise data
- the term proportional to η is the propagation of the experimental error in the reconstructed function, experimental uncertainty



Comparing with fitting the data

- Parametrize the unknown function $f(x, \theta)$

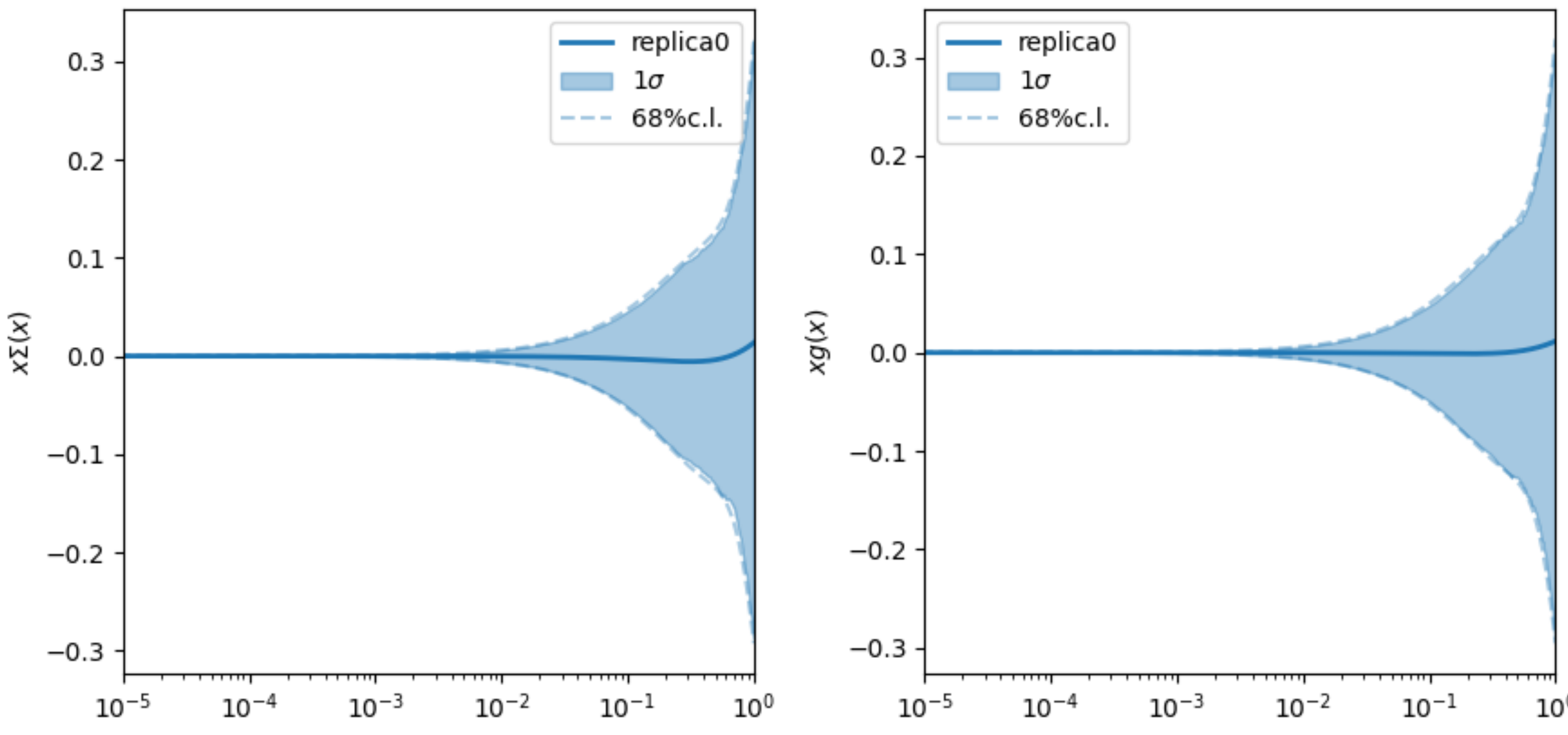
$$f(x,\theta) = \mathrm{NN}(x;\theta)$$

$$= Ax^{\delta}(1-x)^{\eta}\left(1+\sum_{i=1}^{6} a_i T_i(1-2\sqrt{x})\right)$$

- Prior for the (hyper)parameters induces a prior for f

$$p(f) = \int d\theta\, p(\theta) \prod_{x} \delta\left(f(x) - f(x,\theta)\right)$$

NNPDF @ init - distribution over replicas



(almost) Gaussian Process with K determined by the NN architecture

Training - minimising the loss

gradient descent - for all parametrizations

$$\frac{d}{dt}\theta_\mu = -\nabla_\mu \mathcal{L}$$

$$\nabla_\mu \mathcal{L} = -(\nabla_\mu f_t)^T \left(\frac{\partial T}{\partial f}\right)_t^T C_Y^{-1} \epsilon_t\,, \quad \epsilon_t = y - T[f_t]$$

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

where

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

is the Neural Tangent Kernel

for linear data & NN parametrizations

for linear data:

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

for wide neural networks

$$\Theta_t = \Theta + O(1/n)$$

hence we get a linear equation for f_t

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

- The rate at which features are learned is dictated by the eigenvalues/eigenvectors of a flow Hamiltonian

- There is a strong hierarchy in the eigenvalues (spectral bias)

- Solution of the flow equation

$$f_t = \mathcal{A} e^{-\tilde{H} t} f_0 + \mathcal{A} \left(1 - e^{-\tilde{H} t}\right) \mathcal{A}^T (\mathrm{FK})^T C_Y^{-1} Y$$

- At infinite training time, reproduces the GP result for $K^{-1} \to 0$

- WIP: understanding stopping criteria in this formalism

Outlook

- PDFs are a crucial ingredient for exploiting LHC experiments
- Bayesian approach is a convenient framework for solving the related inverse problem
- All hypotheses are explicit in the prior
- Compare different methodologies
- Robust errors defined from the comparison
- Relation between NN/Gaussian Processes/Backus-Gilbert