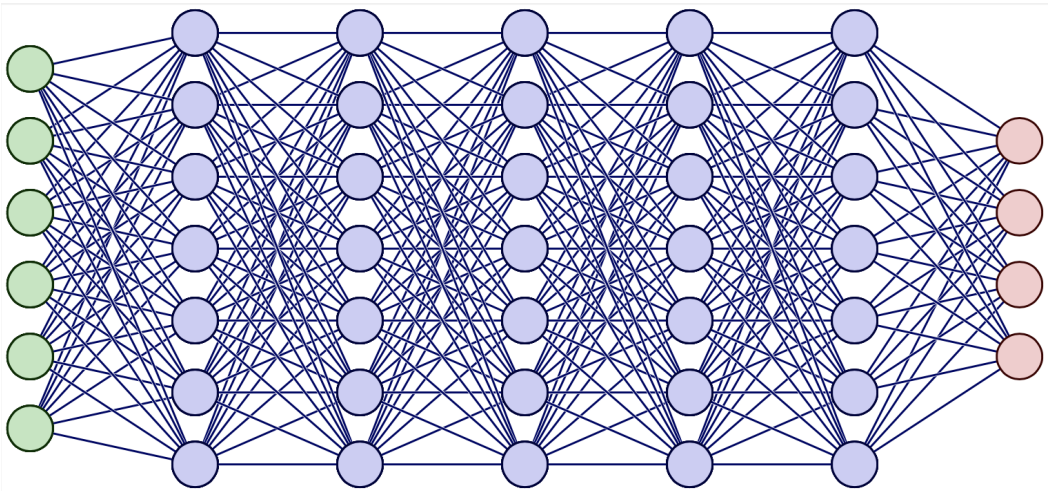
Field Theories and Machine Learning



L Del Debbio

Higgs Centre for Theoretical Physics
University of Edinburgh

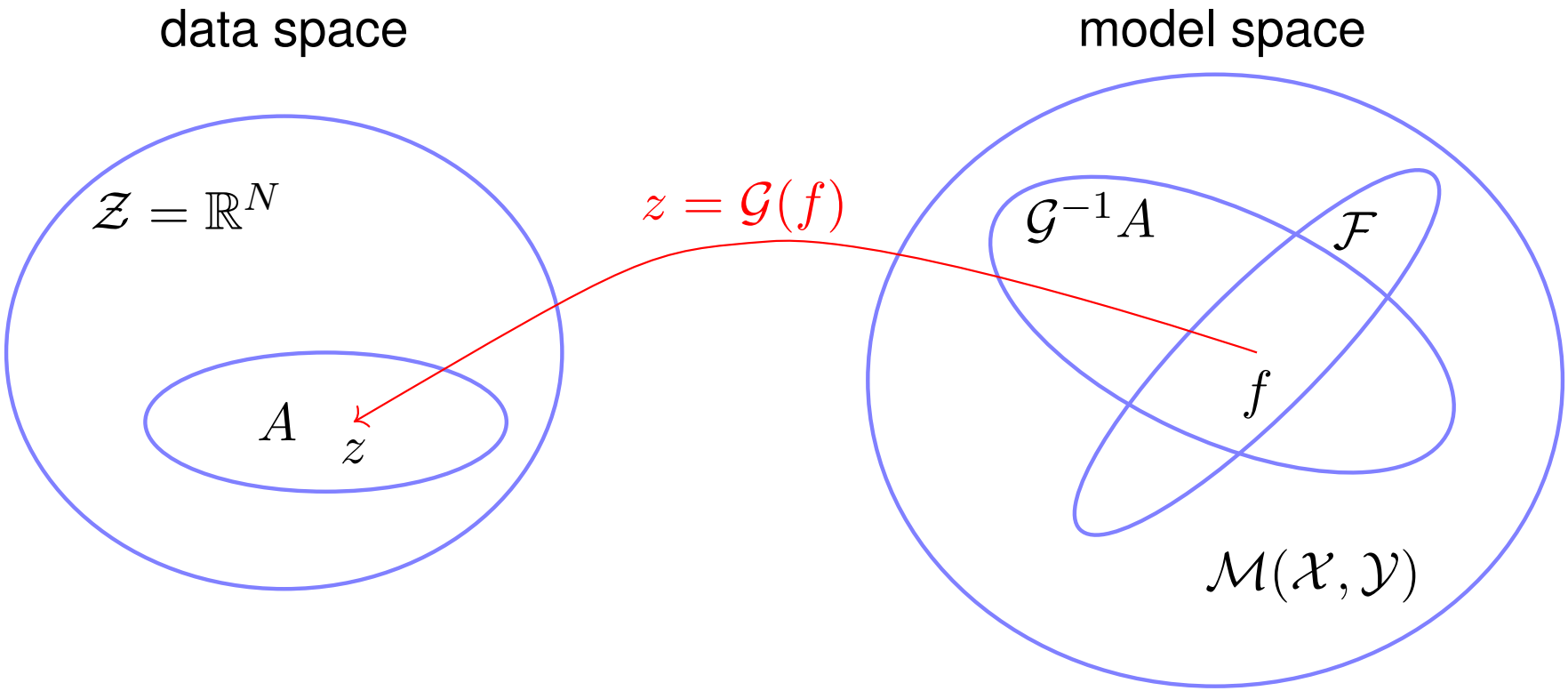
plan of the lectures

- Gaussian Processes and Inverse Problems
- FT for ML
- ML for FT

based on

- Idd et al 21
- Roberts et al 21
- Albandea et al 23

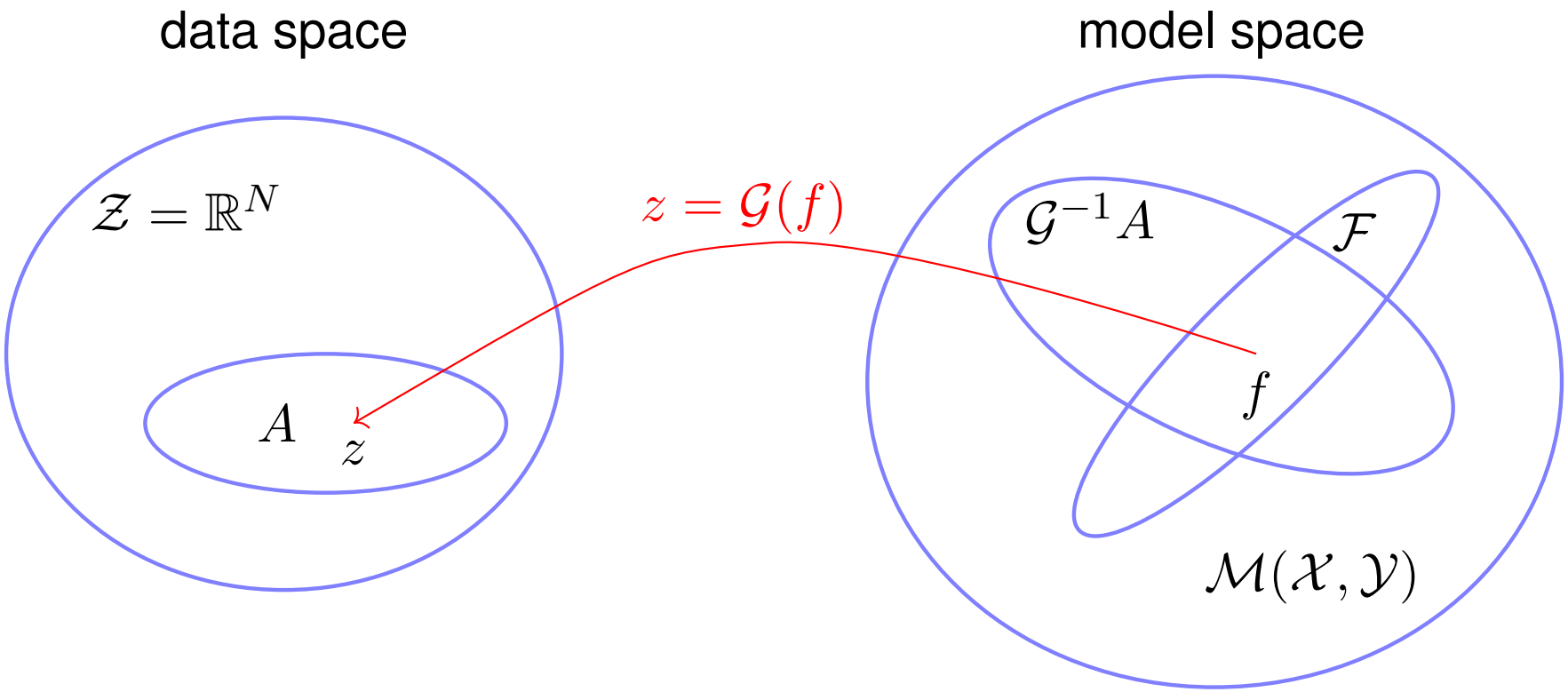
Inverse Problems



Bayes theorem

$$p(f|A) \propto p(A|f)p(f)$$

Inverse Problems



loss function $\mathcal{L} : \mathcal{M}(\mathcal{X}, \mathcal{Y}) \times \mathcal{Z} \longrightarrow \mathbb{R}$

algorithm $\mathcal{A} : \mathcal{Z} \longrightarrow \mathcal{F} \subset \mathcal{M}(\mathcal{X}, \mathcal{Y})$

a few definitions

empirical risk

$$\hat{\mathcal{R}}_s(f) = \mathcal{L}(f, s)$$

erm algorithm

$$\mathcal{A}^{\mathrm{erm}}(s) = \arg\min_{f \in \mathcal{F}} \hat{\mathcal{R}}_s(f) = \hat{f}_s$$

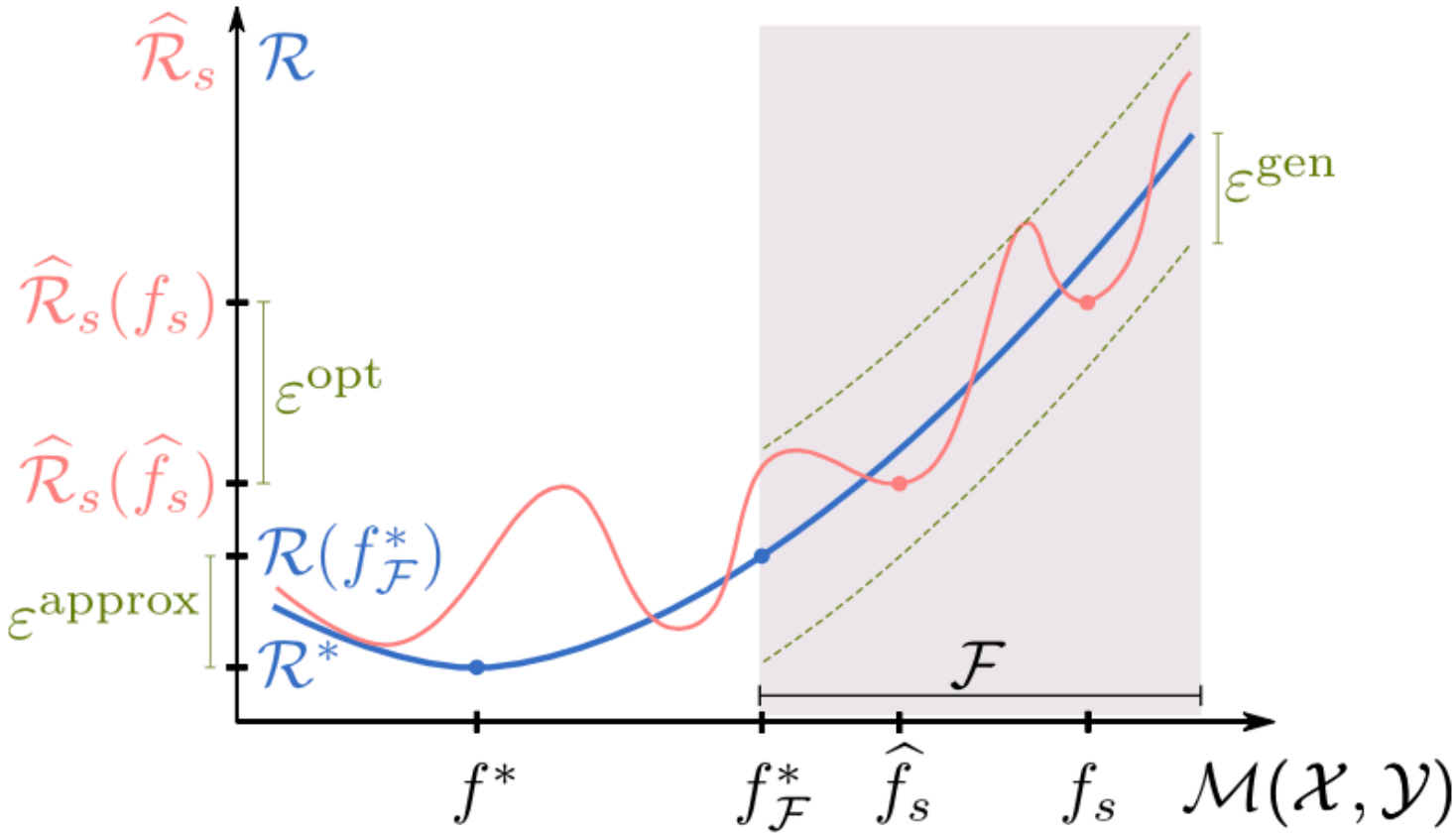
risk

$$\mathcal{R}(f) = \mathbb{E}_Z[\mathcal{L}(f, Z)] = \int dP_Z(z)\, \mathcal{L}(f, z)$$

$$= \frac{1}{N_{\mathrm{rep}}} \sum_{k=1}^{N_{\mathrm{rep}}} \mathcal{L}(f, z^{(k)})$$

Bayes optimal function (generalization)

$$\mathcal{R}^* = \mathcal{R}(f^*) = \inf_{f \in \mathcal{M}(\mathcal{X}, \mathcal{Y})} \mathcal{R}(f)$$



robust definitions are important for precision physics!

Linear Problems & Bayesian inference

$$y_I = \int dx\, C_I(x) f(x)\,, \quad I = 1, \ldots, N_{\rm dat}$$

$\hookrightarrow$ ill-defined problem (e.g. PDFs from DIS, lectures by Forte)

simple underdetermined problem

model: $u \in \mathbb{R}^{N_{\rm mod}}\,, \quad N_{\rm mod} > 1$

theory prediction: $\quad N_{\rm dat} = 1\,, \quad g \in \mathbb{R}^{N_{\rm mod}}\,, \quad y = (g^T u)$

prior: $p(u) = \dfrac{1}{\sqrt{\det(2\pi K)}} \exp\left(-\dfrac{1}{2}|u-m|^2_K\right)$

$$|u-m|^2_K = (u-m)^T K^{-1} (u-m)$$

fluctuations of the experimental data: $\eta \sim p_\eta = \mathcal{N}(0, \gamma^2)$

then

$$\begin{aligned} p\left(y|u\right) &= \text{Prob}\left(g^T u + \eta = y\right) \\ &= \int d\eta\, p_\eta(\eta) \delta\left(g^T u + \eta - y\right) \\ &= p_\eta\left(y - g^T u\right) \end{aligned}$$

Bayes theorem:

$$\begin{aligned} p\left(u|y\right) &\propto \exp\left(-\frac{1}{2\gamma^2}\left|y - g^T u\right|^2 - \frac{1}{2}\left|u - m\right|_K^2\right) \\ &\propto \exp\left(-\frac{1}{2}\left|u - \tilde{m}\right|_{\tilde{K}}^2\right) \end{aligned}$$

$$\tilde{m} = m + (Kg)\ \frac{y - g^T m}{\gamma^2 + (g^T K g)}$$

$$\tilde{K} = K - \frac{(Kg)(Kg)^T}{\gamma^2 + (g^T K g)}$$

hence:

$$\tilde{m}^* = \lim_{\gamma \to 0} \tilde{m} = \frac{(Kg)y}{(g^T K g)}$$

$$\implies g^T \tilde{m}^* = y$$

and

$$\tilde{K}^* = \lim_{\gamma \to 0} \tilde{K} = K - \frac{(Kg)(Kg)^T}{(g^T K g)}$$

$$\implies (\tilde{K}^* g) = 0$$

removing the prior

set $K = \sigma \mathbb{1}$ and take the limit $\sigma \to \infty$

yields

$$\tilde{K}^* = \sigma \left(\mathbb{1} - \frac{g g^T}{(g^T g)} \right)$$
$$= \begin{pmatrix} 0 & \\ & \sigma \end{pmatrix}$$

the error in the g direction vanishes

the error in the orthogonal subspace diverges

Bayesian Inference

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

Bayesian solution of the form:

$$f_i = f(x_i) \quad \longrightarrow \quad p(f_1 \dots f_N | y, \mathcal{H}) = \frac{p(y|f,\mathcal{H}) p(f|\mathcal{H})}{p(y|\mathcal{H})}$$

$\hookrightarrow$ dependence on the choice of the prior

Gaussian processes: $f \sim \mathcal{GP}(m, k)$

$$\mathbb{E}[f_i] = m(x_i)\,, \quad \mathrm{Cov}[f_i, f_j] = k(x_i, x_j)$$

$$\mathrm{prior} : \mathcal{GP}(m,k) \quad \Longrightarrow \quad \mathrm{posterior} : \mathcal{GP}(\tilde{m}, \tilde{k})$$

Gaussian Processes in Action

$$\mathbf{f} \in \mathbb{R}^N, \quad \mathbf{f}^* \in \mathbb{R}^M,$$

prior distribution

$$\frac{1}{\sqrt{\det(2\pi K)}} \exp\left\{-\frac{1}{2}\left((\mathbf{f}-\mathbf{m})^T, (\mathbf{f}^*-\mathbf{m}^*)^T\right) K^{-1} \begin{pmatrix} \mathbf{f}-\mathbf{m} \\ \mathbf{f}^*-\mathbf{m}^* \end{pmatrix}\right\}$$

where K is an $(N+M)\times(N+M)$ matrix

$$K = \begin{pmatrix} k(\mathbf{x},\mathbf{x}^T) & k(\mathbf{x},\mathbf{x}^{*T}) \\ k(\mathbf{x}^*,\mathbf{x}^T) & k(\mathbf{x}^*,\mathbf{x}^{*T}) \end{pmatrix} = \begin{pmatrix} K_{\mathbf{xx}} & K_{\mathbf{xx}^*} \\ K_{\mathbf{x}^*\mathbf{x}} & K_{\mathbf{x}^*\mathbf{x}^*} \end{pmatrix}.$$

$$T_I = \sum_{i=1}^{N} (\mathrm{FK})_{Ii} f_i$$

where $(\mathrm{FK})_{Ii} = b_I(x_i)$.

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

experimental error on data

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

we have three stochastic variables $\mathbf{f}, \mathbf{f}^*, \epsilon$ with

$$\mathrm{Cov} = \begin{pmatrix} K & 0 \\ 0 & C_Y \end{pmatrix} \,,$$

and we will impose that

$$y = (\mathrm{FK})\mathbf{f} + \epsilon \,,$$

Posterior Distribution

$$(\mathbf{f}|(\mathrm{FK})\mathbf{f} + \epsilon = \mathbf{y}) \sim \mathcal{N}\left(\tilde{\mathbf{m}}, \tilde{K}_{\mathbf{xx}}\right) ,$$

where

$$\begin{aligned}\tilde{\mathbf{m}} &= \mathbf{m} + K_{\mathbf{xx}}(\mathrm{FK})^T C_{YT}^{+} \left(\mathbf{y} - (\mathrm{FK})\mathbf{m}\right) , \\ \tilde{K}_{\mathbf{xx}} &= K_{\mathbf{xx}} - K_{\mathbf{xx}}(\mathrm{FK})^T C_{YT}^{+} (\mathrm{FK}) K_{\mathbf{xx}} ,\end{aligned}$$

$$C_{YT} = (\mathrm{FK}) K_{\mathbf{xx}} (\mathrm{FK})^T + C_Y .$$

exercise:

$$\tilde{K}_{\mathbf{xx}}^{-1} = K_{\mathbf{xx}}^{-1} + (\mathrm{FK})^T C_Y^{-1} (\mathrm{FK}) \quad \rightarrow \quad (\mathrm{FK})^T C_Y^{-1} (\mathrm{FK})$$

Boostrapping a fit

fluctuations of the data encoded in an ensemble of ”replicas”

$$y^{(k)} = y + \epsilon^{(k)}\,, \quad k = 1, \ldots, N_{\rm rep}$$

for each replica minimize

$$\chi^{2,(k)} = \frac{1}{2}\left(y^{(k)} - (\mathrm{FK})\mathbf{f}\right)^T C_Y^{-1}\left(y^{(k)} - (\mathrm{FK})\mathbf{f}\right)$$
$$f^{(k)} = \left((\mathrm{FK})^T C_Y^{-1}(\mathrm{FK})\right)^{-1}(\mathrm{FK})^T C_Y^{-1} y^{(k)}$$

$f^{(k)}$ are normally distributed with mean $\tilde{\mathbf{m}}$ and covariance $\tilde{K}_{\mathbf{xx}}$

for $\mathbf{m} = 0$, $K^{-1}_{\mathbf{xx}} = 0$

More on the Posterior

for the *unseen* points

$$(\mathbf{f}^*|(\mathrm{FK})\mathbf{f} + \epsilon = y) \sim \mathcal{N}\left(\tilde{\mathbf{m}}^*, \tilde{K}_{\mathbf{x}^*\mathbf{x}^*}\right) ,$$

where in this case

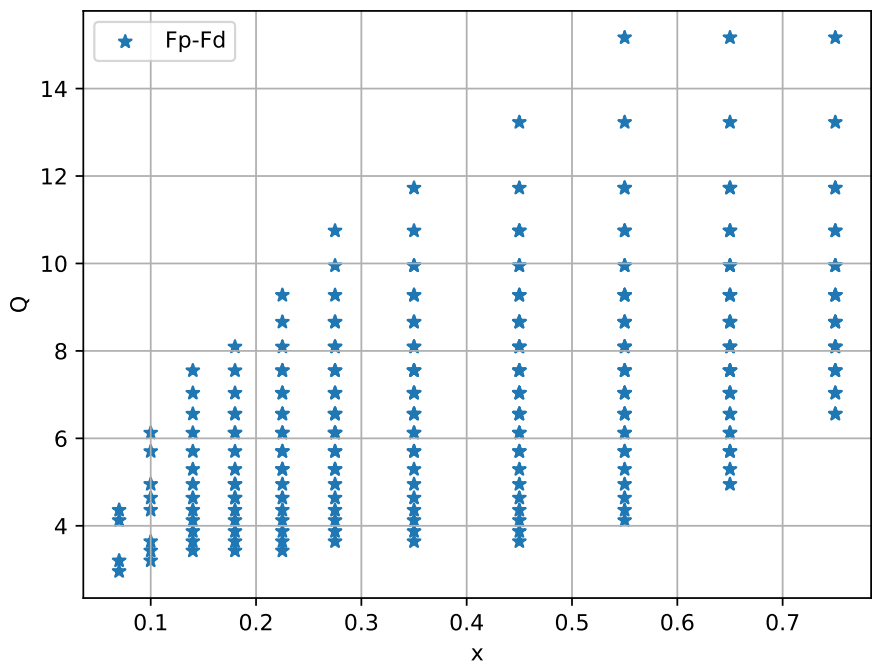
$$\begin{aligned}
\tilde{\mathbf{m}}^* &= \mathbf{m}^* + K_{\mathbf{x}^*\mathbf{x}}(\mathrm{FK})^T C_{YT}^+ \left(\mathbf{y} - (\mathrm{FK})\mathbf{m}\right) , \\
\tilde{K}_{\mathbf{x}^*\mathbf{x}^*} &= K_{\mathbf{x}^*\mathbf{x}^*} - K_{\mathbf{x}^*\mathbf{x}}(\mathrm{FK})^T C_{YT}^+ (\mathrm{FK}) K_{\mathbf{x}\mathbf{x}^*} .
\end{aligned}$$

Example

$$F_2^p - F_2^d = C_{T_3} \otimes T_3$$

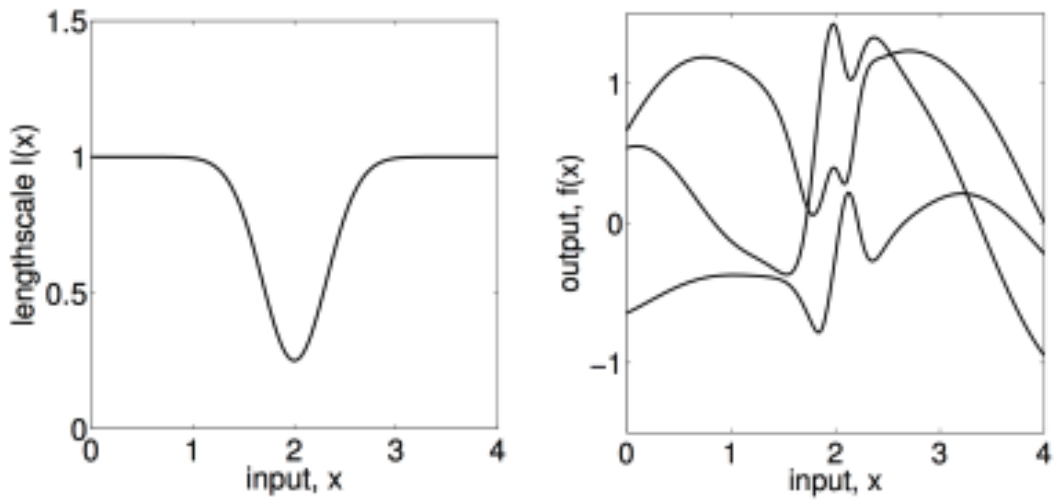
data covariance

$$C_Y = \mathrm{Cov}\left[F_2^p, F_2^p\right] + \mathrm{Cov}\left[F_2^d, F_2^d\right] - 2\mathrm{Cov}\left[F_2^p, F_2^d\right]$$



Gibbs Kernel

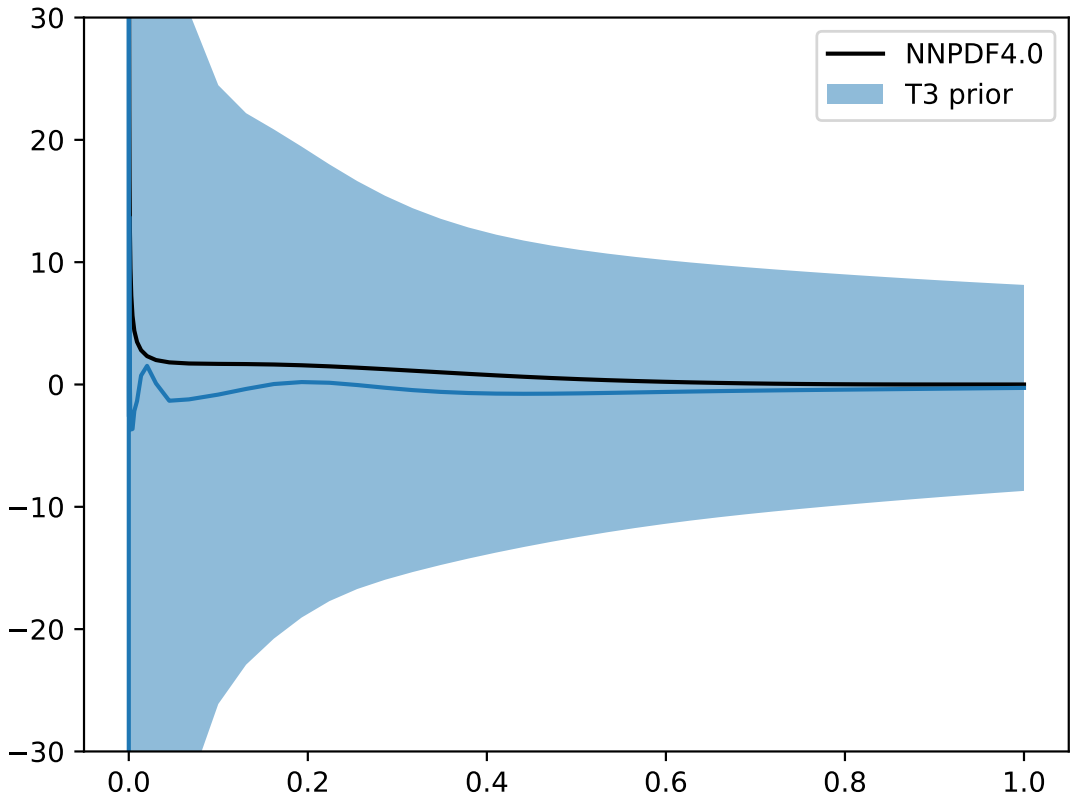
$$k\left(x,y\right)=\sigma^{2}\sqrt{\frac{2l\left(x\right)l\left(y\right)}{l^{2}\left(x\right)+l^{2}\left(y\right)}}\exp\left[-\frac{\left(x-y\right)^{2}}{l^{2}\left(x\right)+l^{2}\left(y\right)}\right]$$



here we use

$$l\left(x\right)=l_{0}\times\left(x+\epsilon\right)\ .$$

Prior Distribution



Closure Test

generate artificial data: $y = (\mathrm{FK})\mathbf{f}_0$

introduce a *filter* function

$$R_{\mathbf{xx}} = K_{\mathbf{xx}}(\mathrm{FK})^T C_{YT}^{-1}(\mathrm{FK})$$
$$R_{\mathbf{x}^*\mathbf{x}} = K_{\mathbf{x}^*\mathbf{x}}(\mathrm{FK})^T C_{YT}^{-1}(\mathrm{FK})$$

then

$$\tilde{\mathbf{m}} - \mathbf{m} = R_{\mathbf{xx}}\,(\mathbf{f}_0 - \mathbf{m})$$
$$\tilde{\mathbf{m}}^* - \mathbf{m}^* = R_{\mathbf{x}^*\mathbf{x}}\,(\mathbf{f}_0 - \mathbf{m}) \quad \Longrightarrow \quad \tilde{\mathbf{m}}^* = R_{\mathbf{x}^*\mathbf{x}}\,\mathbf{f}_0$$

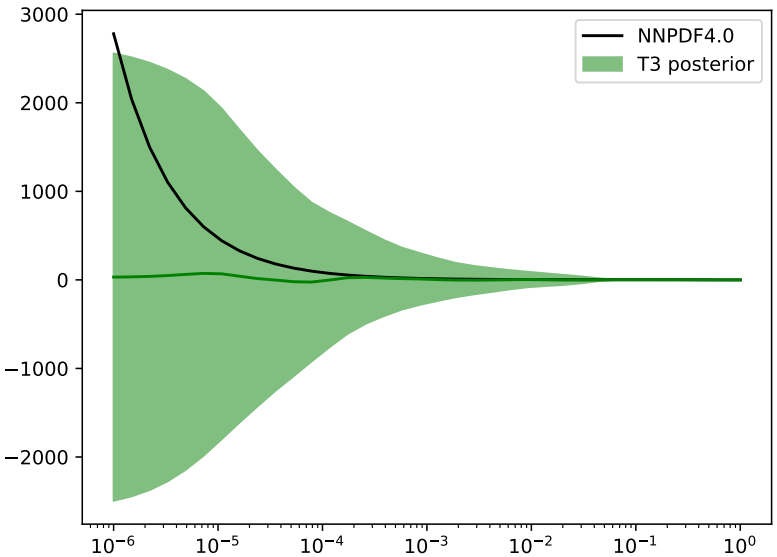
considering the case $\mathbf{x}^* = \mathbf{x}$ we define the bias

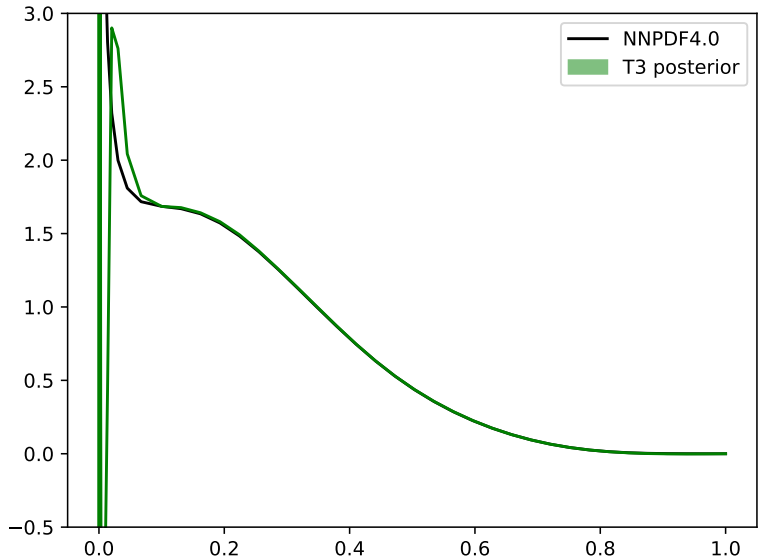
$$\tilde{\mathbf{m}} - \mathbf{f}_0 = \left[R_{\mathbf{xx}} - \mathbb{1}\right]\mathbf{f}_0\,,$$

and the covariance of the posterior can be written as

$$\tilde{K} = (\mathbb{1} - R_{\mathbf{xx}})\,K_{\mathbf{xx}}\,(\mathbb{1} - R_{\mathbf{xx}})^T + a_{\mathbf{xx}}^T C_Y a_{\mathbf{xx}}\,.$$

Posterior Distribution



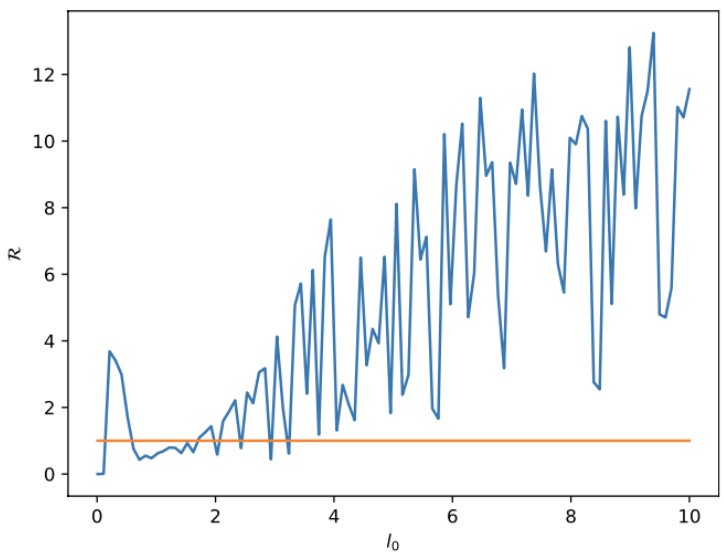


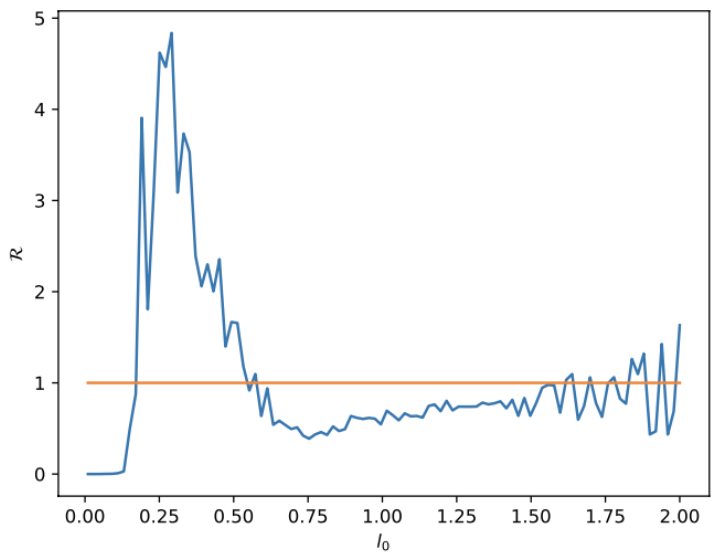
Bias & Variance

define

$$\mathcal{B} = (\mathrm{FK})(\tilde{\mathbf{m}} - \mathbf{f}_0) = T - y$$
$$\mathcal{V} = (\mathrm{FK})\tilde{K}(\mathrm{FK})^T$$

$$\mathcal{R} = \sum_{I,J} \mathcal{B}_I (\mathcal{V}^{-1})_{IJ} \mathcal{B}_J$$





Outlook

- GP are an effective way to explore the space of functions
- prior is explicitly specified AND determines the posterior
- careful study of the bias/variance ratio in order to choose the prior