Machine Learning for Applied Physics and High Energy Physics

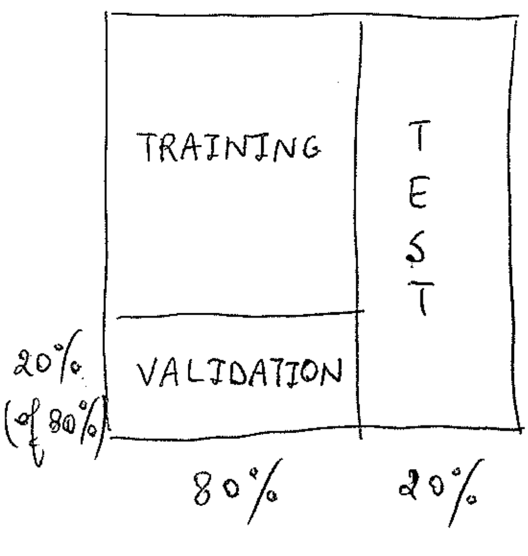
Lecture 6: K-folding. Closure tests

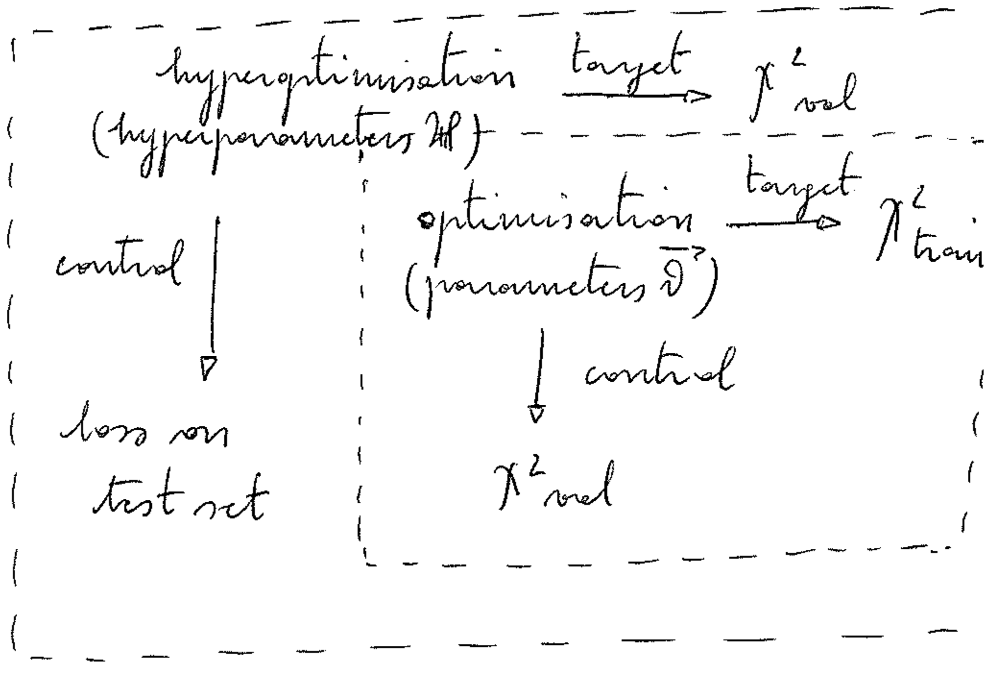
6.1 K-folding

We have formulated the inverse problem of determining the posterior conditional probability $p(f(\vec{x},\vec{\vartheta})/\vec{y})$ $= p(f/\mathcal{D})$ in Bayesian terms, that is

$$p(\vec{\vartheta}/\mathcal{D}) = \frac{p(\vec{\vartheta})\, p(\mathcal{D}/\vec{\vartheta})}{\int d\vec{\vartheta}\, p(\mathcal{D}/\vec{\vartheta})}$$

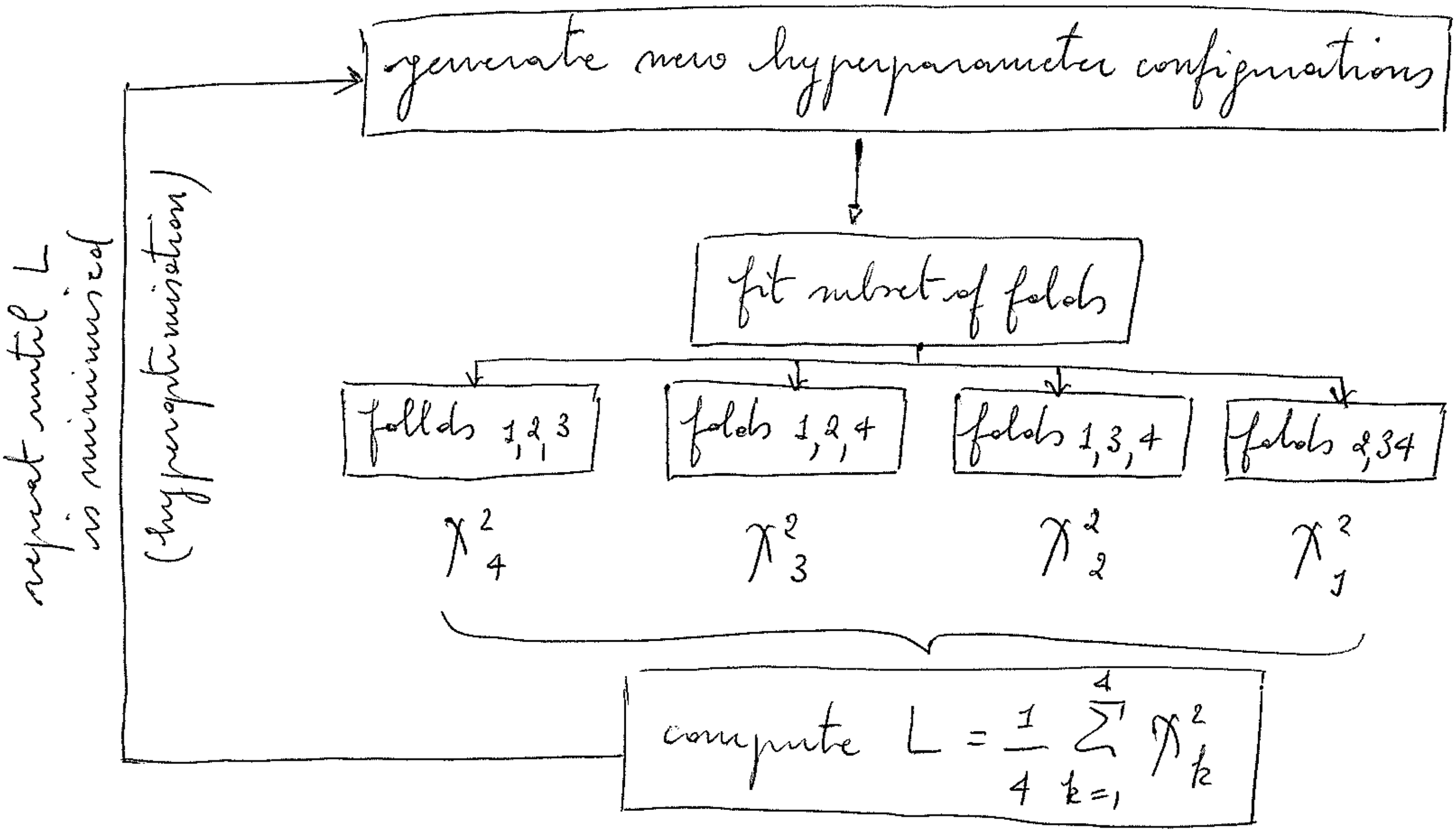
This formulation is however incomplete, because the prior depends not only on the model parameters $\vec{\vartheta}$, but also on hyperparameters $\mathcal{H}$. These hyperparameters include, e.g., the architecture of the neural network, the details of the optimisation algorithm, etc. A way of "fitting the methodology" is hyperparametrisation. We can imagine to split the dataset in three subsets





Hyperparameters are optimised on the unseen test set.
The way to choose the test set is provided by K-folding

1. Divide the data set in N folds. Folds MUST be homogeneous.
2. Train the model on $N-1$ folds. Use the excluded fold as test set.
3. Repeat 2 for all folds and compute the average loss.
4. Repeat 2 and 3 for a scan over models (aka new hyperparameter configurations) to minimise the loss.



5. The optimal hyperparameter configuration is the one that minimises the average loss.

Remarks:

1 - there is not a unique way to define the loss; there

may be other, equally effective losses than the average loss 2 - it is convenient to perform hyperoptimisation in stages, separating hyperparameters that enter the definition of the model and hyperparameters that enter the optimisation of the model.

6.2 Closure tests – stochastic uncertainty representation.

As mentioned multiple times, we have formulate the inverse problem of determining the posterior conditional probability $p(f(x, \vec{\vartheta})/\vec{y}) = p(f/\mathcal{D})$ in Bayesian terms

$$p(\vec{\vartheta}/\mathcal{D}) = \frac{p(\vec{\vartheta})\, p(\mathcal{D}/\vec{\vartheta})}{\int d\vec{\vartheta}\, p(\mathcal{D}/\vec{\vartheta})}$$

Remember that f is a forward mapping

$$f : \mathbb{R}^{D} \longrightarrow \mathbb{R}^{D'} \quad \text{where we parametrise } f = f(\vec{x}, \vec{\vartheta})$$
$$(X) \longrightarrow (Y)$$

The underlying true mapping is $f(\vec{x}) = \mathcal{G}(\vec{x})$ and the data is subjected to noise : $\vec{y} = \mathcal{G}(\vec{x}) + \vec{\mathcal{E}}$. We assume that this noise is sampled from a multi-Gaussian distribution centered on zero : $\vec{\mathcal{E}} \sim \mathcal{N}(0, C_Y)$ where $Y = \mathbb{R}^{D'}$ and C_Y is the covariance matrix in the space of the labels (the experimental covariance matrix). The problem is solved by determining

$p(\vec{\vartheta}/\mathcal{D})$ with Bayes' theorem as MAP $\hat{\vec{\vartheta}} = \underset{\vec{\vartheta}}{\arg\max}\, p(\vec{\vartheta}/\mathcal{D})$

by maximum a posteriori, which corresponds to the maximum log-likelihood estimation with a credibility interval

There are two ways of addressing the problem.

1 MAXIMUM LIKELIHOOD

$$\mathcal{P}(\vec{\vartheta}/\mathcal{D}) \longrightarrow f(\vec{x}, \hat{\vec{\vartheta}})$$

2 MONTE CARLO

$$\mathcal{P}(\vec{\vartheta}/\mathcal{D}) \longrightarrow \{ f^{(k)}(\vec{x}, \hat{\vec{\vartheta}}^{(k)}) \}$$

In these expressions $\mathcal{P}(\vec{\vartheta}/\mathcal{D})$ is such that

$$\mathbb{E}[Y] = \int_f \mathcal{D} f \, \mathcal{P}(\vec{\vartheta}/\mathcal{D}) \, f \qquad \text{expectation value}$$

$$\mathbb{V}[Y] = \int_f \mathcal{D} f \, \mathcal{P}(\vec{\vartheta}/\mathcal{D}) \left[f - \mathbb{E}[Y] \right]^2 \qquad \text{variance}$$

$$\mathbb{E}[Y] = \mathbb{E}[f(\vec{x}, \hat{\vec{\vartheta}})] \qquad \text{for HESSIAN}$$

$$\mathbb{E}[Y] = \frac{1}{N_{rep}} \sum_{k=1}^{N_{rep}} f^{(k)}(\vec{x}, \hat{\vec{\vartheta}}^{(k)}) \qquad \text{MONTE CARLO}$$

$$\mathbb{V}[Y] \longrightarrow \text{Hessian contour } \Delta\chi^2 = 1 \qquad \text{HESSIAN}$$

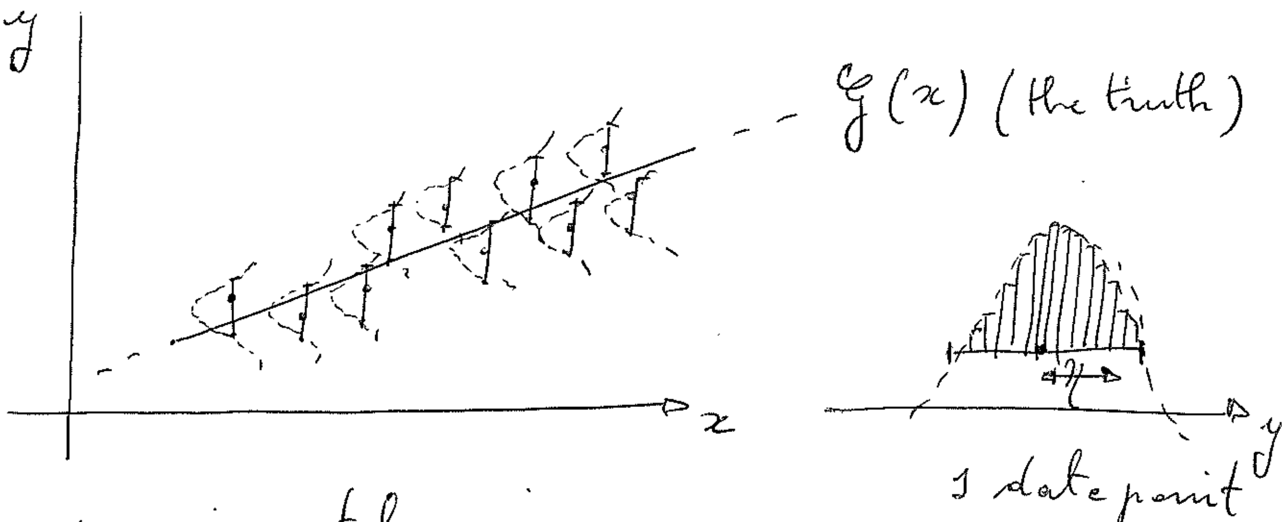
$$\mathbb{V}[Y] = \frac{1}{N_{rep}} \sum_{k=1}^{N_{rep}} \left(f^{(k)}(\vec{x}, \hat{\vec{\vartheta}}^{(k)}) - \mathbb{E}[Y] \right)^2 \qquad \text{MONTE CARLO}$$

In the MONTE CARLO method, I sample the experimental data distribution by generating an ensemble of replicas

$$\vec{y} \to \vec{y}^{(k)} = \vec{y} + \vec{\eta}^{(k)} = \mathcal{G}(x) + \varepsilon + \eta^{(k)}$$

where $\eta^{(k)} \sim \mathcal{N}(0, C_Y)$

In other words (in a 1D model)



ε is the experimental noise

$\eta^{(k)}$ is the sampling fluctuation

How many replicas should I generate? Require that the difference between $\mathcal{N}(0, C_Y)$ and the same distribution sampled from replicas is smaller than a given threshold.

HESSIAN	MONTE CARLO
rely on quadratic expansion around $\hat{\vec{\vartheta}}$	insensitive to details of expansion
perform only one fit	perform Nrep fits

$$\chi^2 = \frac{1}{N}\sum_{n,n'=1}^{N} \left(\vec{y}_n - f(\vec{x}_n, \vec{\vartheta})\right) C_Y^{-1} \left(\vec{y}_{n'} - f(\vec{x}_{n'}, \vec{\vartheta})\right) \quad \text{HESSIAN}$$

$$\chi^{2(k)} = \frac{1}{N}\sum_{n,n'=1}^{N} \left(\vec{y}_n^{\,(k)} - f^{(k)}(\vec{x}_n, \vec{\vartheta})\right) C_Y^{-1} \left(\vec{y}_{n'}^{\,(k)} - f^{(k)}(\vec{x}_{n'}, \vec{\vartheta})\right) \quad \text{MONTE CARLO}$$

6.3 Closure tests - levels.

Suppose to use the MONTE CARLO approach to represent the noise of the data into the noise of $f(\vec{x}_n, \vec{\vartheta})$.

We can distinguish three ways of performing a closure test. These three ways are called "levels" of a closure tests.

1 LEVEL 0

Pseudodata are generated without statistical noise

$$\vec{y} = \mathcal{G}(\vec{x}) = \vec{y}^{(0)}$$

The fitting proceeds as usual, with minimisation of the loss function

$$\chi_0^{2(k)} = \frac{1}{N}\sum_{n,n'=1}^{N}\left(\vec{y}_n^{(0)} - f(\vec{x}_n,\vec{\vartheta})^{(k)}\right) C_y^{-1}\left(\vec{y}_{n'}^{(0)} - f(\vec{x}_{n'},\vec{\vartheta})^{(k)}\right)$$

Note that, for each replica k, $f(\vec{x}_n,\vec{\vartheta})^{(k)}$ differ only because minimisation is performed starting from a different point in parameter space. For an unbiased methodology, we expect $\chi_0^{2(k)} \xrightarrow[\text{training length}]{\text{for large}} 0$. Note that C_y is the experimental covariance matrix.

2 LEVEL 1

Pseudodata are generated with statistical noise

$$\vec{y} = \mathcal{G}(x) + \vec{\varepsilon} = \vec{y}^{(0,\varepsilon)}$$

The fitting proceeds as usual, with minimisation of the loss function

$$\chi_1^{2(k)} = \frac{1}{N}\sum_{n,n'=1}^{N}\left(\vec{y}_n^{(0,\varepsilon)} - f(\vec{x}_n,\vec{\vartheta})^{(k)}\right) C_y^{-1}\left(\vec{y}_{n'}^{(0,\varepsilon)} - f(\vec{x}_{n'},\vec{\vartheta})^{(k)}\right)$$

Again, for each replica k, $f(\vec{x}_n,\vec{\vartheta})^{(k)}$ differ only because minimisation is performed starting from a different point in parameter space. For an unbiased methodology we expect $\chi_1^{2(k)} \xrightarrow[\text{training length}]{\text{for large}} 1$.

3 LEVEL 2

Pseudodata are generated with stochastic noise; replicas are also generated

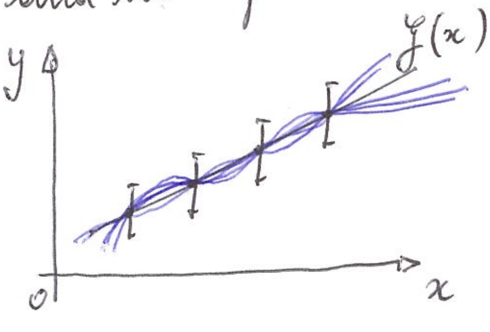
$$\vec{y} = \mathcal{G}(\vec{x}) + \vec{\varepsilon} + \vec{\eta}^{(k)} \equiv \vec{y}^{(k)}$$

The fitting proceeds with minimisation of the loss function

$$\chi_2^{2(k)} = \frac{1}{N} \sum_{n,n'=1}^{N} \left(\vec{y}_n^{(k)} - f(\vec{x}, \vec{\vartheta})^{(k)} \right) C_Y^{-1} \left(\vec{y}_{n'}^{(k)} - f(\vec{x}_{n'}, \vec{\vartheta})^{(k)} \right)$$

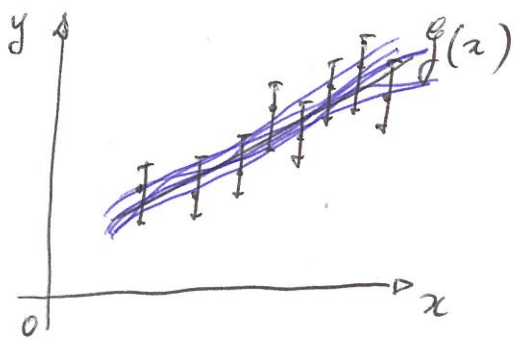
For each replica k, $f(\vec{x}, \vec{\vartheta})^{(k)}$ differ for the starting point in parameter space AND for the pseudodata set to which the parameters are trained. For an unbiased methodology, we expect $\chi_2^{2(k)} \xrightarrow[\text{training length}]{\text{for large}} 2$.

level 0
no fluctuations
(no stochastic noise)
and no replicas



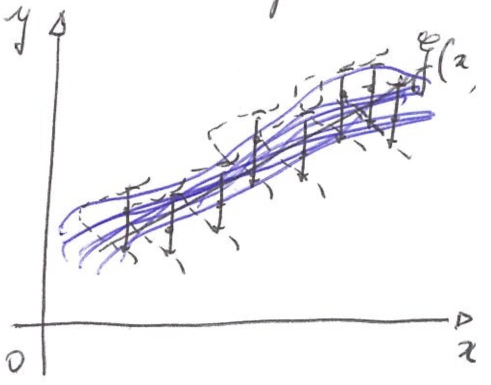
INTERPOLATION
and EXTRAPOLATION
uncertainty

level 1
stochastic noise
no replicas



FUNTIONAL
uncertainty
(epistemic)

level 2
stochastic noise
and replicas



STOCHASTIC
uncertainty

6.4 Closure tests – estimators

Let us define an error function as the expectation value across replicas, denoted as $\mathbb{E}_\eta[\cdot]$ of the loss function between predictions from the k-th replica, $f(\vec{x_n},\vec{\vartheta})^{(k)}$, and the unfluctuated data (the generalisation error)

$$\mathbb{E}_\eta\left[\chi^{2(k)}\right] = \frac{1}{N}\,\mathbb{E}_\eta\left[\sum_{n,n'=1}^{N}\left(\vec{y}_n - f(\vec{x_n},\vec{\vartheta})^{(k)}\right)C_Y^{-1}\left(\vec{y}_{n'} - f(\vec{x}_{n'},\vec{\vartheta})^{(k)}\right)\right]$$

As already shown, this expression can be decomposed as

$$\mathbb{E}_\eta\left[\chi^{2(k)}\right] = \text{noise} + \text{bias}^2 + \text{variance}$$

where

$$\text{noise} = \frac{1}{N}\sum_{n,n'=1}^{N}\left(\vec{y}_n - \mathcal{G}_n(\vec{x})\right)C_Y^{-1}\left(\vec{y}_{n'} - \mathcal{G}_{n'}(\vec{x})\right)$$

$$\text{bias}^2 = \frac{1}{N}\sum_{n,n'=1}^{N}\left(\mathcal{G}_n(\vec{x}) - \mathbb{E}_\eta[f(\vec{x_n},\vec{\vartheta})]\right)C_Y^{-1}\left(\mathcal{G}_{n'}(\vec{x}) - \mathbb{E}_\eta[f(\vec{x}_{n'},\vec{\vartheta})]\right)$$

$$\text{variance} = \frac{1}{N}\sum_{n,n'=1}^{N}\mathbb{E}_\eta\left[\sum_{n,n'=1}^{N}\left(f(\vec{x_n},\vec{\vartheta})^{(k)} - \mathbb{E}_\eta[f(\vec{x_n},\vec{\vartheta})]\right)C_Y^{-1}\right.$$
$$\left.\times\left(f(\vec{x}_{n'},\vec{\vartheta})^{(k)} - \mathbb{E}_\eta[f(\vec{x}_{n'},\vec{\vartheta})]\right)\right]$$

We further define $\Delta\chi^2$ as the difference between the χ^2 evaluated from comparing $\mathbb{E}_\eta[f(\vec{x_n},\vec{\vartheta})]$ and the level-one data, i.e. $\vec{y} = \mathcal{G}(\vec{x}) + \vec{\varepsilon}$, and the χ^2 evaluated from comparing the truth $\mathcal{G}(\vec{x})$ and the same level-one data

$$\Delta\chi^2 = \chi^2\left[\mathbb{E}_\eta[f(\vec{x_n},\vec{\vartheta})],\vec{y}\right] - \chi^2\left[\mathcal{G}(\vec{x}),\vec{y}\right].$$

It is clear that, for $\Delta\chi^2 = 0$, we have optimal learning; for $\Delta\chi^2 < 0$ we have we have overfitting; for $\Delta\chi^2 > 0$ we have underfitting.

Due to its dependence on the shift vector $\vec{\varepsilon}$, $\Delta\chi^2$ is a stochastic variable. We can sample it by "running" the universe, i.e. by simulating many measurements that differ for stochastic noise, and then averaging over noise.

One can then compute

$$\mathbb{E}_{\varepsilon}[\text{bias}^2] = \frac{1}{N}\,\mathbb{E}_{\varepsilon}\Big[\sum_{n,n'=1}^{N}\big(\mathbb{E}_{\eta}[f(\vec{x}_n,\vec{\vartheta})] - g(\vec{x})\big)\,C_Y^{-1} \times \big(\mathbb{E}_{\eta}[f(\vec{x}_{n'},\vec{\vartheta})] - g(\vec{x})\big)$$

$$\mathbb{E}_{\varepsilon}[\text{variance}] = \frac{1}{N}\,\mathbb{E}_{\varepsilon}\Big[\mathbb{E}_{\eta}\Big[\sum_{n,n'=1}^{N}\big(\mathbb{E}_{\eta}[f(\vec{x}_n,\vec{\vartheta})] - \vec{y}_n^{\,(k)}\big)\,C_Y^{-1}\,\big(\mathbb{E}_{\eta}[f(\vec{x}_{n'},\vec{\vartheta})] - \vec{y}_{n'}^{\,(k)}\big)\Big]\Big]$$

For an unbiased, optimally learned model

$$\sqrt{\frac{\mathbb{E}_{\varepsilon}[\text{bias}^2]}{\mathbb{E}_{\varepsilon}[\text{variance}]}} = R_{BV} \longrightarrow 1$$

Finally, we can define the quantile estimator:

$$\xi_{n\sigma} = \frac{1}{n_x}\frac{1}{n_{fit}}\sum_{j}^{n_x}\sum_{l}^{n_{fit}} \underbrace{I_{[-n\sigma^{(l)}(x_j),\, n\sigma^{(l)}(x_j)]}}_{1 \text{ if true; } 0 \text{ if false}}\big(\mathbb{E}_{\eta}[f^{(l)}(x_j,\hat{\vec{\vartheta}})] - g(x_j)\big)$$

For an unbiased, optimally learned model

$$\xi_{1\sigma} = 0{,}68.$$