Deep learning

1. Predictive modelling

Clément Gorin
clement.gorin@univ-paris1.fr

Harbin Institute of Technology, Shenzhen
August 2024

Predictive modelling

This lecture introduces (supervised) machine learning models and their connection with traditional econometrics

- We provide a conceptual framework to understand the differences and similarities between the two approaches
- There is a substantial overlap between the two disciplines as both are grounded in statistics
- The views of the two disciplines are voluntarily contrasted and does not reflect the diversity of methods

Function approximation

Function approximation

Consider an unknown function of interest F that maps some input array x_i to a response y_i given some error ϵ_i

$$y_i = F(x_i) + \epsilon_i \tag{1}$$
$$i = 1, \ldots, N$$

where N is the number of observable population (x_i, y_i) pairs generated by the function

- Many tasks can be formulated this way e.g. estimating one aspect or the entire data-generating process
- The input may be pixel intensities in an image or characters in a text mapped to a probability

Function approximation

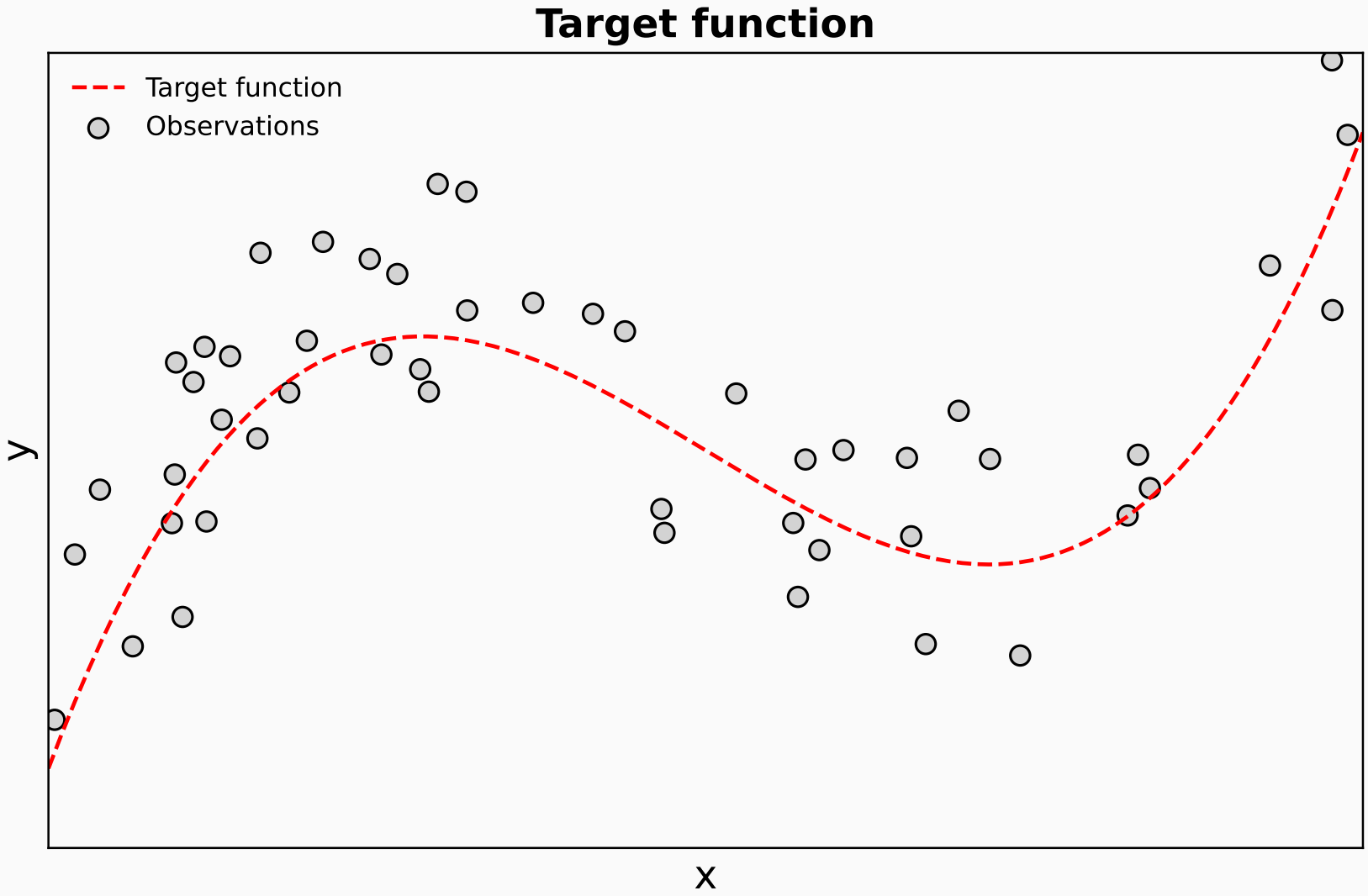
To illustrate the different approaches, let's simulate a data-generating process or target function

Target function

```
1  import numpy as np
2  from numpy import random
3
4  # Function
5  def F(x): return x**3 - x * 5
6
7  # Data
8  random.seed(0)
9  n = 50
10 s = np.random.choice([True, False], n, p=[.5, .5])
11 x = random.uniform(-3, 3, n)
12 e = random.normal(0, 3, n)
13 y = f(x) + e
```

In practice we do not know the target function, otherwise statistical modelling would not be required

Function approximation



Function approximation

In practice, there are two reasons why we may want to compute an estimate $\hat{F}$ for this function

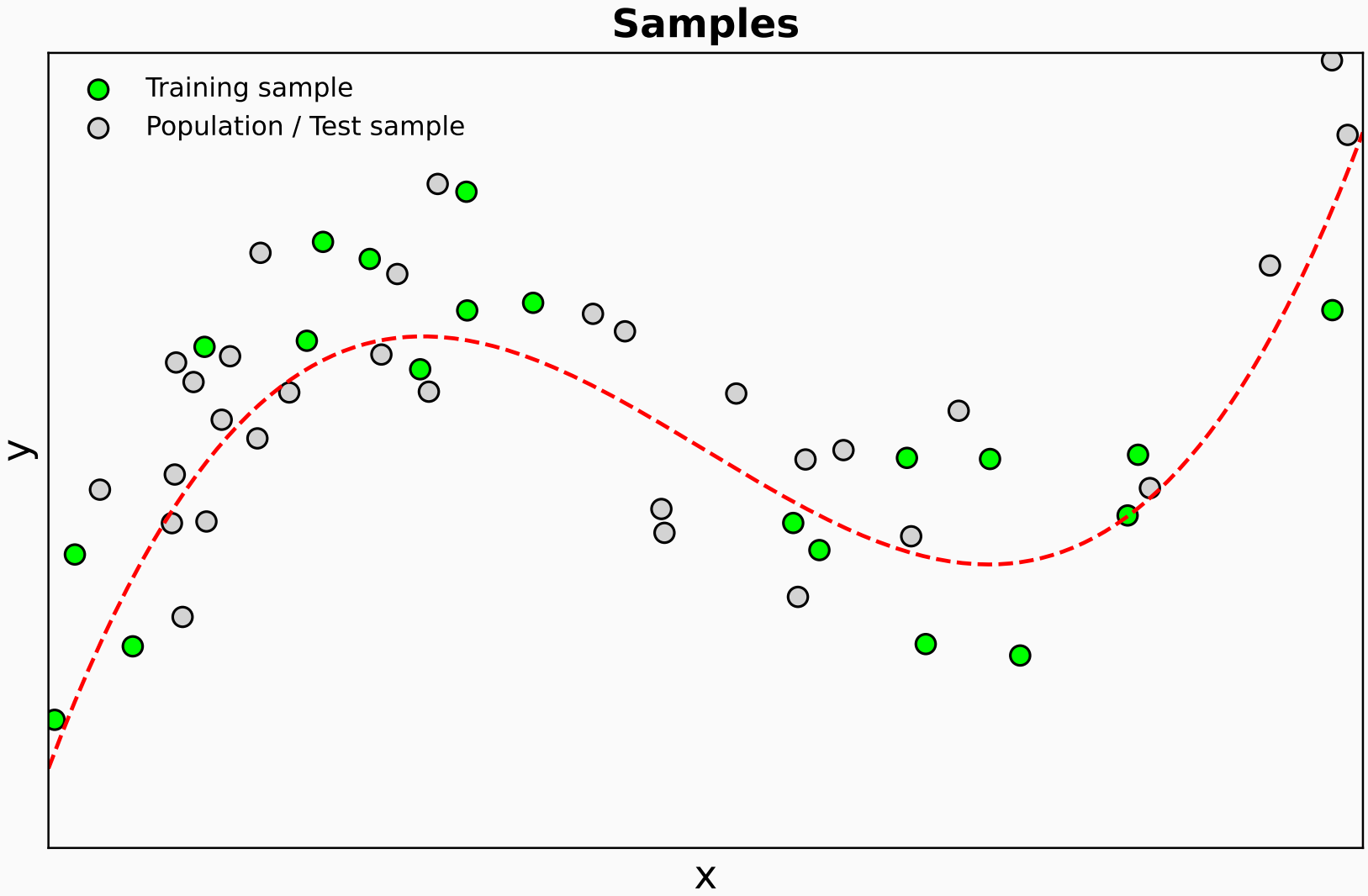
- To interpret the relationship between x_i and $y_i \rightarrow$ focus on estimated parameters $\hat{\beta}$ and explanatory power
- To predict accurately y_i using $x_i \rightarrow$ focus on estimated response $\hat{y}_i$ and predictive power (out-of-sample)
- Parametric econometrics is used for the former while the latter is best solved using supervised learning

Function approximation

Regardless of the application, we only observe a random sample representative of the population

- A random sample means that every observation has an equal probability of being sampled
- This implies that the sample's characteristics are distributed similarly as those of the population
- Unrepresentative or insufficient samples causes models to produce biased and/or inefficient estimates

Function approximation



Parameter estimation

Parameter estimation

There is an infinity of functions passing through the observed data points and function approximation is unsolvable

$$y_i = f(x_i, \beta) + \epsilon_i \tag{2}$$
$$i = 1, \ldots, n$$

where f is an empirical model, β a set of parameters and n the number of observed sample (x_i, y_i) pairs

- We impose constraints on the problem by restricting the search space to a parametric family of functions
- The model structure encodes those constraints and depend on the application and prior knowledge

Empirical model

```
import pasty
from statsmodels import api

class Spline:

 def __init__(self, d:int, df:int):
     self.d     = d
     self.df    = df
     self.model = None

 def transform(self, x:np.ndarray, d:int, df:int) -> np.ndarray:
     X = patsy.dmatrix(f'bs(x, degree={d}, df={df})', {'x': x})
     return X

 def fit(self, x:np.ndarray, y:np.ndarray) -> None:
     X = self.transform(x, d=self.d, df=self.df)
     self.model = api.GLM(y, X).fit()

 def predict(self, x:np.ndarray) -> np.ndarray:
     X  = self.transform(x, d=self.d, df=self.df)
     yh = self.model.predict(X)
     return yh
```

Parameter estimation

Within this parameter-space the optimisation searches for the values $\hat{\beta}$ that minimise the loss over the sample

$$\hat{\beta} = \underset{\beta}{argmin} \sum_{i=1}^{n} \mathcal{L}(y_i, f(x_i, \beta)) \tag{3}$$

$$i = 1, \ldots, n$$

where $\mathcal{L}$ is a loss function, a measure of distance between the observed and the estimated response

- The form of the loss function depends on the distribution of the response (e.g. continuous, categorical)
- The optimisation may have a closed-form solution or require iterative routines (e.g. gradient-based)

Inference

Inference

To compute interpretable parameters estimates, the model is often additive and linear in the parameters

- The additivity allows parameter to be interpreted separately (i.e. partialling out, ceteris paribus)
- The linearity ensures that parameters have a simple interpretation (e.g. average unit-increase)
- Transformations are restricted to those producing interpretable parameters (e.g. logs, interactions)

Inference

Importantly, the model is derived from either theory or intuition and requires strong prior assumptions

- The researcher defines what inputs enter the model and their functional relationship with the response
- Non-linearities and interactions are defined by transforming the inputs before estimating the model
- This approach is sensible for low-dimensional problems with a well established theoretical background

Inference

Under restrictive conditions on the form of F and the conditional distribution of the population error

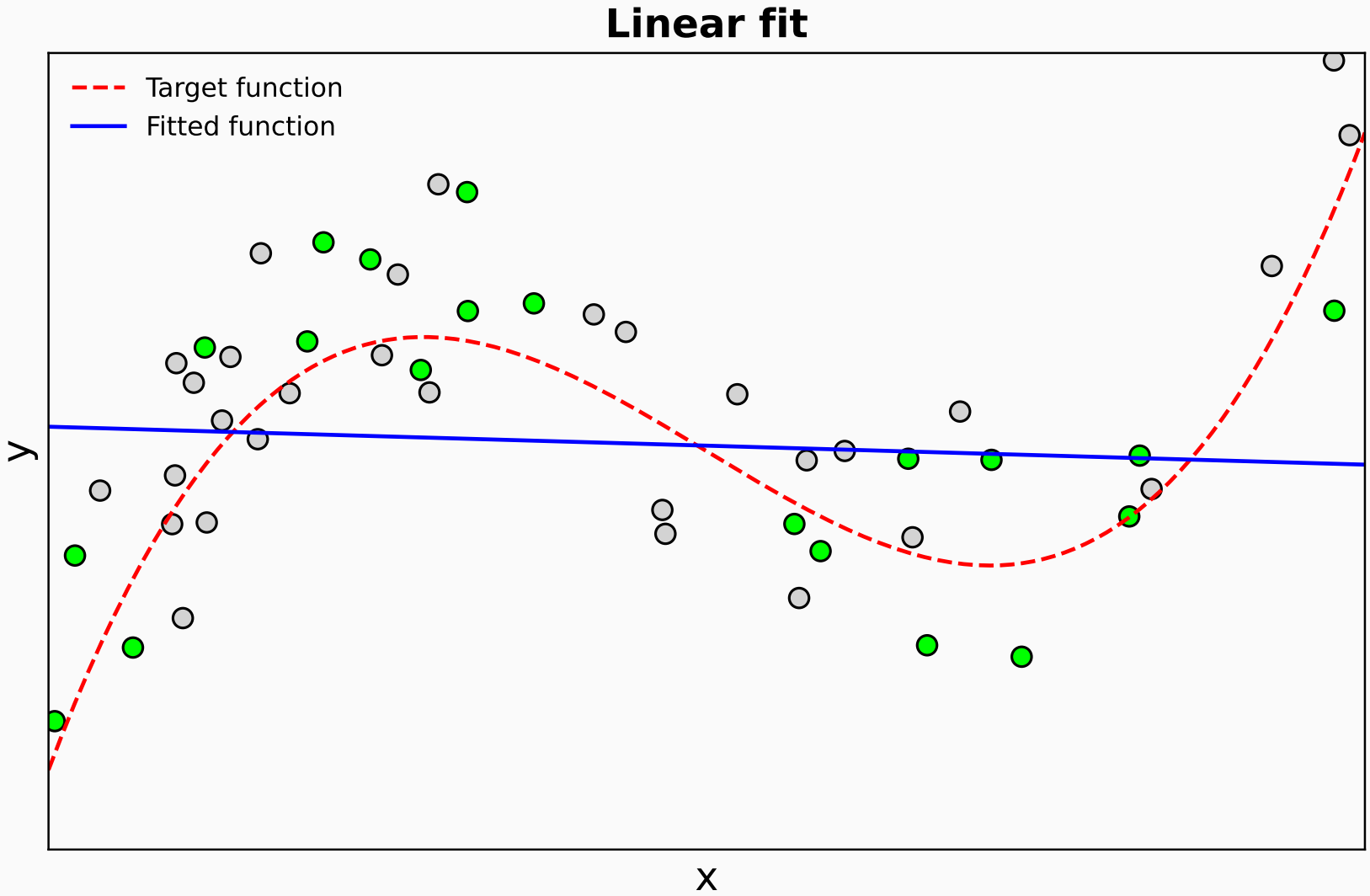
- Parametric econometric models deliver interpretable and meaningful average parameter estimates
- Statistical inference provides confidence intervals to establish generality beyond the sample used
- The optimisation has a closed-form solution or can be solved efficiently (e.g Newton-Raphson)

Inference

Linear models are usually not suited for predictions except when the target function is linear

- Focus on unbiasedness (under linearity) and does not trade-off bias and variance (more on this)
- When the target function is non-linear, linear model provide stable but biased predictions
- Cannot approximate complex functions, high-dimensional with unknown non-linearities and interactions

Inference



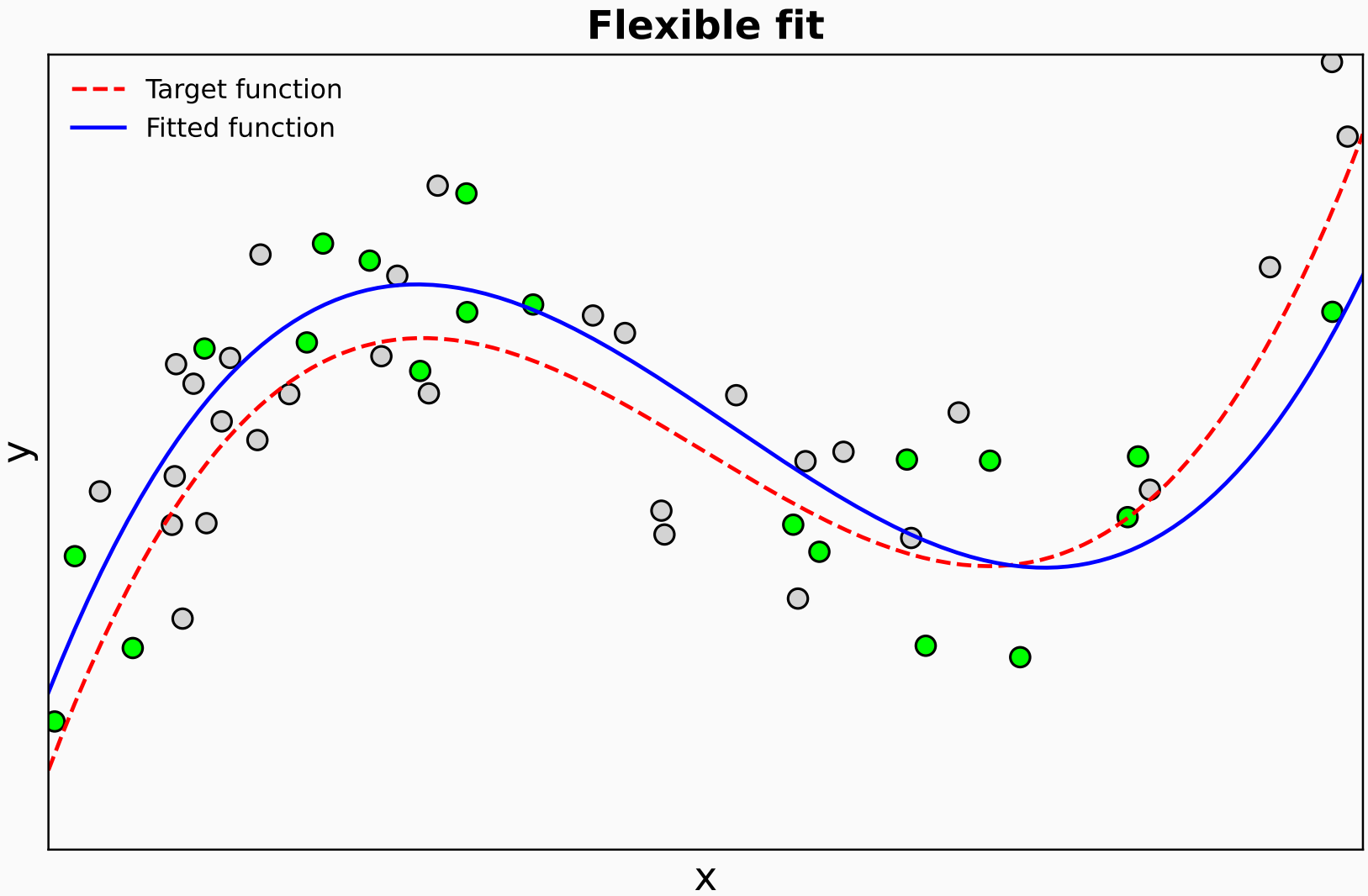
Prediction

Prediction

For predictive tasks, parameter interpretability is not required and more flexible methods become available

- Supervised learning models make much fewer assumptions about the target function
- The models use the sample data and numerous parameters to learn about its shape
- They uncover key non-linearities and interactions that were not defined in advance (i.e. functional form)

Prediction



Motivating example

- We predict the probability that aerial images contain a house (e.g. logistic)
- An image is represented as a multi-dimensional array of pixel intensities
- This observation contains $1024 \times 1024 \times 3 = 3.14$ million "variables"

Prediction

The target function is high-dimensional and contains numerous unknown interactions and non-linearities

- Interactions: objects are depicted by particular spatial patterns of pixels with specific intensities
- Non-linearities: e.g. no simple mapping between the colour intensities and the probability
- We have no idea about this function and cannot possibly devise a sensible parametric model

Prediction

A more sensible approach is to build a model directly from the sample data using supervised learning

- Other models naturally handle the high dimensions and uncover non-linearities and interactions
- The estimated function be general enough to predict accurately out-of-sample observations
- When the estimated model approximates the target function with accuracy, it “learns” the function

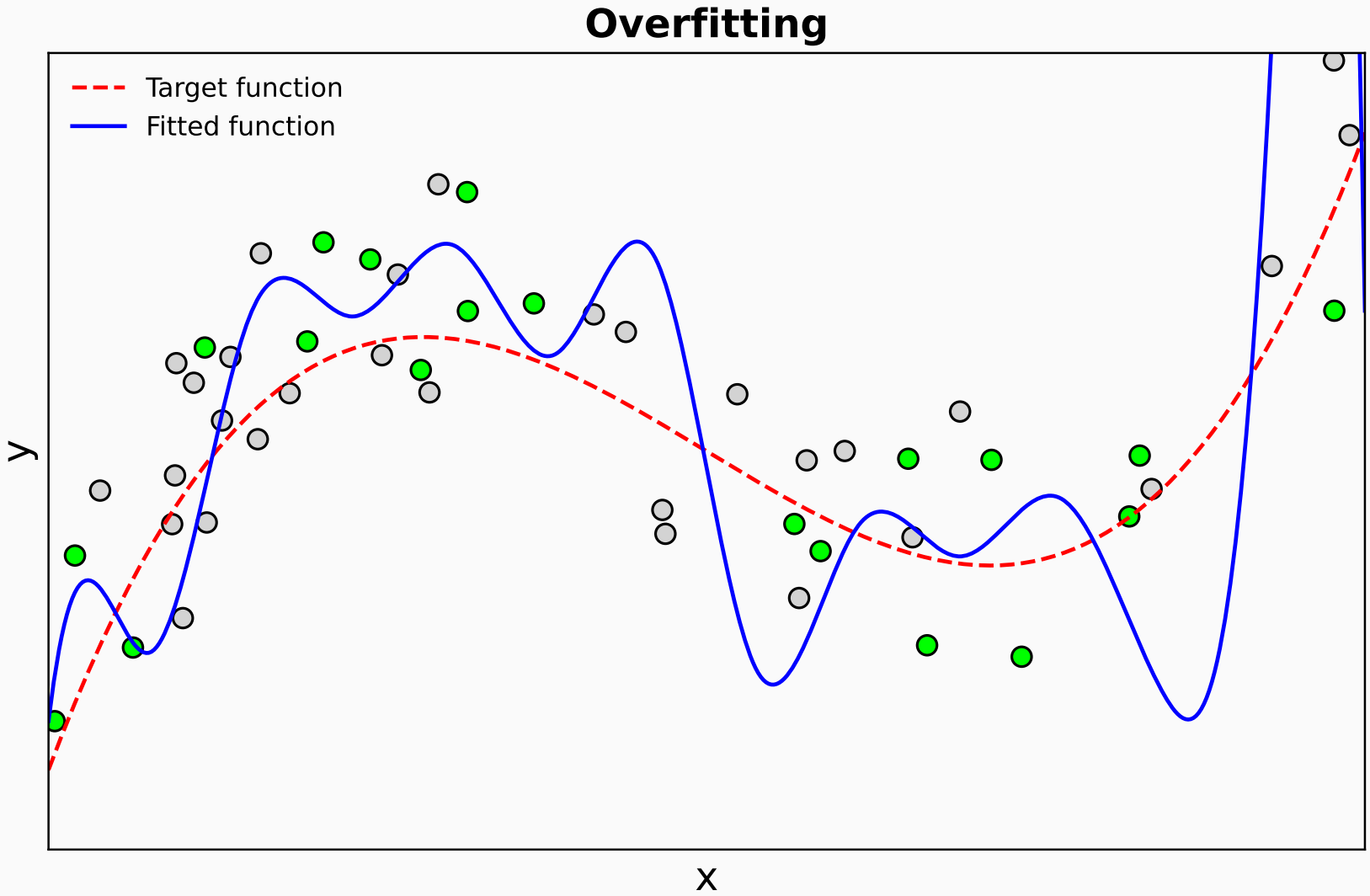
Regularisation

Regularisation

For predictive applications, the model must be assessed using observations outside the (training) sample

- Flexible methods approximate not only the target function but also the observational error (i.e. overfitting)
- The training sample loss provides a biased estimate of predictive performance (i.e. tends toward 0)
- The model is assessed on another random sample that has not been used for training i.e. test sample

Regularisation



Regularisation

During optimisation, we minimise the training sample error but we really care about the test sample error

- The empirical model must be flexible enough to approximate complex functions (i.e. low bias)
- However the estimated function must also generalise to out-of sample observations (i.e. low variance)
- This fundamental trade-off is addressed by adding a regularisation term to the loss function

Regularisation

A regularisation term is added to the loss to encourage more general models. Equation (3) becomes

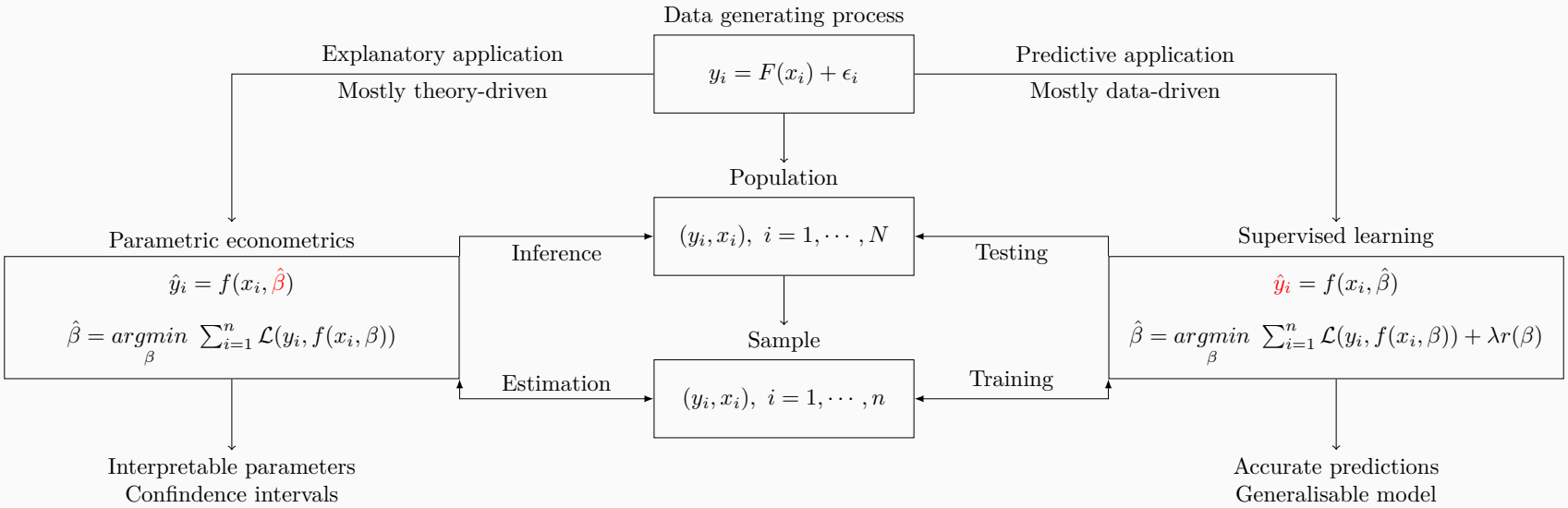
$$\hat{\beta} = \underset{\beta}{argmin} \sum_{i=1}^{n} \mathcal{L}\left(y_i, f(x_i, \beta)\right) + \lambda r(\beta) \tag{4}$$

$$i = 1, \ldots, n$$

where r is a regularisation function that increases with large parameter values and λ is the regularisation parameter

- Large parameter values allow the estimated function to change its behaviour over a small space
- The relative strength of the two counteracting effects is governed by the tuning parameter λ

Inference and predictive methods



▸ Methods comparison

Bias and Variance

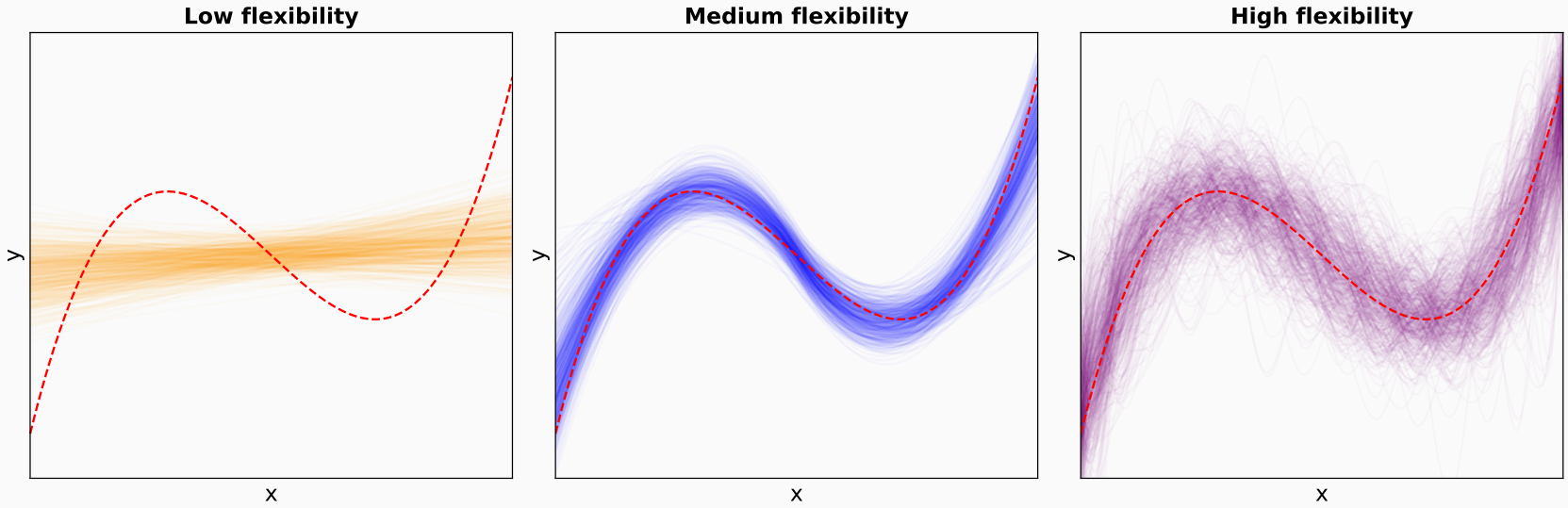
Bias and Variance

To produce accurate preditions our model must minimise simultaneously prediction bias and variance

- These statistics refer to the distribution of $\hat{y}$ when the same model is estimated on different random samples
- Estimators can be considered as random variables with the estimated response being its realisation
- For predictions, we consider the integrated effect of bias and variance over the entire function space

▸ Random variable

Bootstrapped predictions for different estimators

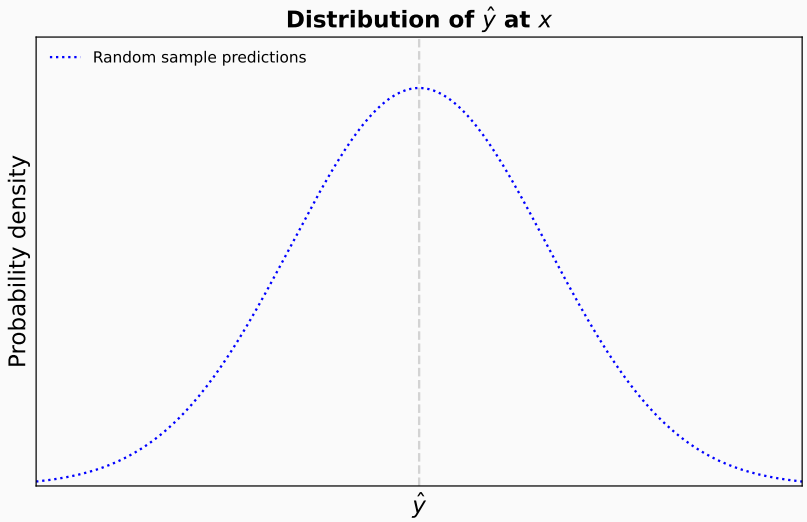


- Bias is the tendency of an estimator to systematically produce an over-estimate or under-estimate
- Variance is the amount by which the estimates change when fitting the model on another random sample

Bias and Variance

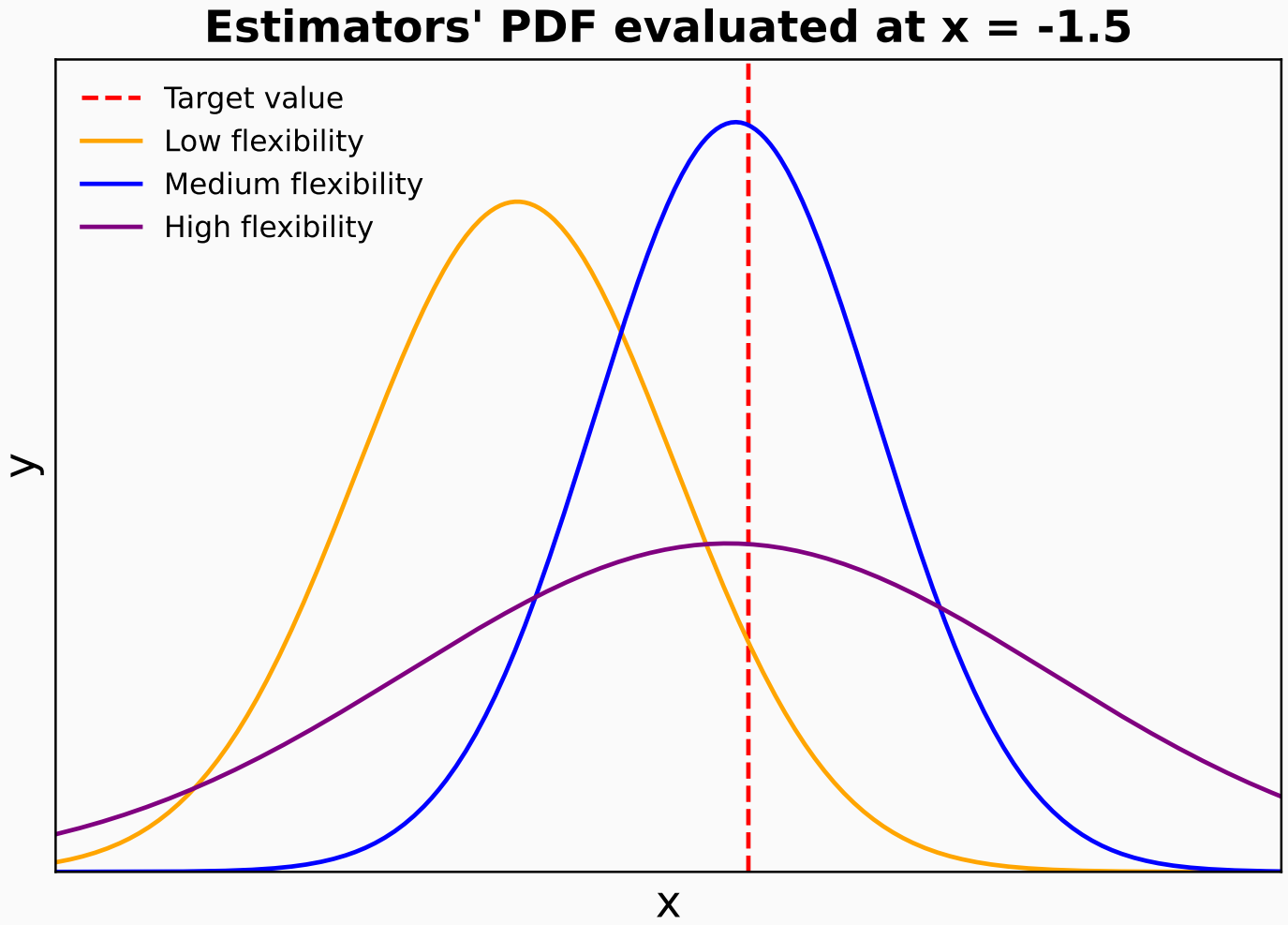
The shape of the probability density function gives important information about the estimator's properties

- The first moment measures the central tendency (bias)
- The second moment measures the dispersion (variance)



Integrating over a range of the PDF provides the probability that the estimate falls within that range

Bias and Variance



Bias and Variance

To derive formally bias-variance trade-off, consider the mean squared error loss function

$$\mathcal{L}_{mse}(y_i, \hat{y}_i) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i, \hat{\beta}))^2 \tag{5}$$

in predictive applications, the error is typically expressed at the observation level, rathen than the sample

- The squared term ensures that positive and negative estimated errors do not cancel out
- Observations whose predicted response lie far from the true response are given more weight

Bias and Variance

The expected training mean squared error can be expressed as the sum of reducible and irreducible error

$$E\left[(y-\hat{f})^2\right] = \underbrace{(f-\hat{f})^2}_{\text{reducible error}} + \underbrace{Var[\epsilon]}_{\text{irreducible error}} \tag{6}$$

where $f = F(x_i)$, $\hat{f} = f(x_i, \hat{\beta})$, and the subscript i is dropped to simplify notation

- We can only minimise reducible error, the distance between the target and the estimated function
- Even with a perfect estimate for F, the training error is non-zero because ϵ_i cannot be predicted using x_i

Bias and Variance

Note that f is constant and $\hat{f}$ is assumed to be fixed. If the error is truly random $E(\epsilon) = 0$

$$
\begin{aligned}
E\left[(y - \hat{f})^2\right] &= E\left[(f + \epsilon - \hat{f})^2\right] \\
&= E\left[(f - \hat{f} + \epsilon)^2\right] \\
&= E\left[(f - \hat{f})^2 + 2\epsilon(f - \hat{f}) + \epsilon^2\right] \\
&= E\left[(f - \hat{f})^2\right] + E\left[2\epsilon(f - \hat{f})\right] + E[\epsilon^2] \\
&= E\left[(f - \hat{f})^2\right] + 2\underline{E[\epsilon]}E[f - \hat{f}] + E[\epsilon^2]
\end{aligned}
$$

$$
E\left[(y - \hat{f})^2\right] = \underbrace{(f - \hat{f})^2}_{\text{reducible error}} + \underbrace{Var[\epsilon]}_{\text{irreducible error}}
$$

Bias and Variance

The reducible error in equation (6) can be further decomposed as the sum of bias (squared) and variance

$$E\left[(y-\hat{f})^2\right] = \underbrace{\left(E[\hat{f}]-f\right)^2}_{\text{bias}^2} + \underbrace{E\left[\left(\hat{f}-E[\hat{f}]\right)^2\right]}_{\text{variance}} + Var[\epsilon] \quad (7)$$

- Minimising the test sample error implies minimising both the bias and variance on the training sample
- Increased regularisation can be seen as reducing the estimator's variance at the cost of introducing bias

Bias and Variance

$$
\begin{aligned}
E\left[\left(y-\hat{f}\right)^2\right] &= E\left[\left(f+\epsilon-\hat{f}\right)^2\right] \\
&= E\left[\left(f+\epsilon-\hat{f}+E[\hat{f}]-E[\hat{f}]\right)^2\right] \\
&= E\left[\left(f-E[\hat{f}]\right)+\epsilon+\left(E[\hat{f}]-\hat{f}\right)^2\right] \\
&= E\left[\left(f-E[\hat{f}]\right)^2\right]+E(\epsilon^2)+E\left[\left(E[\hat{f}]-\hat{f}\right)^2\right] \\
&+ 2E\left[\left(f-E[\hat{f}]\right)\epsilon\right]+2E\left[\epsilon\left(E[\hat{f}]-\hat{f}\right)\right] \\
&+ 2E\left[\left(f-E[\hat{f}]\right)\left(E[\hat{f}]-\hat{f}\right)\right]
\end{aligned}
$$

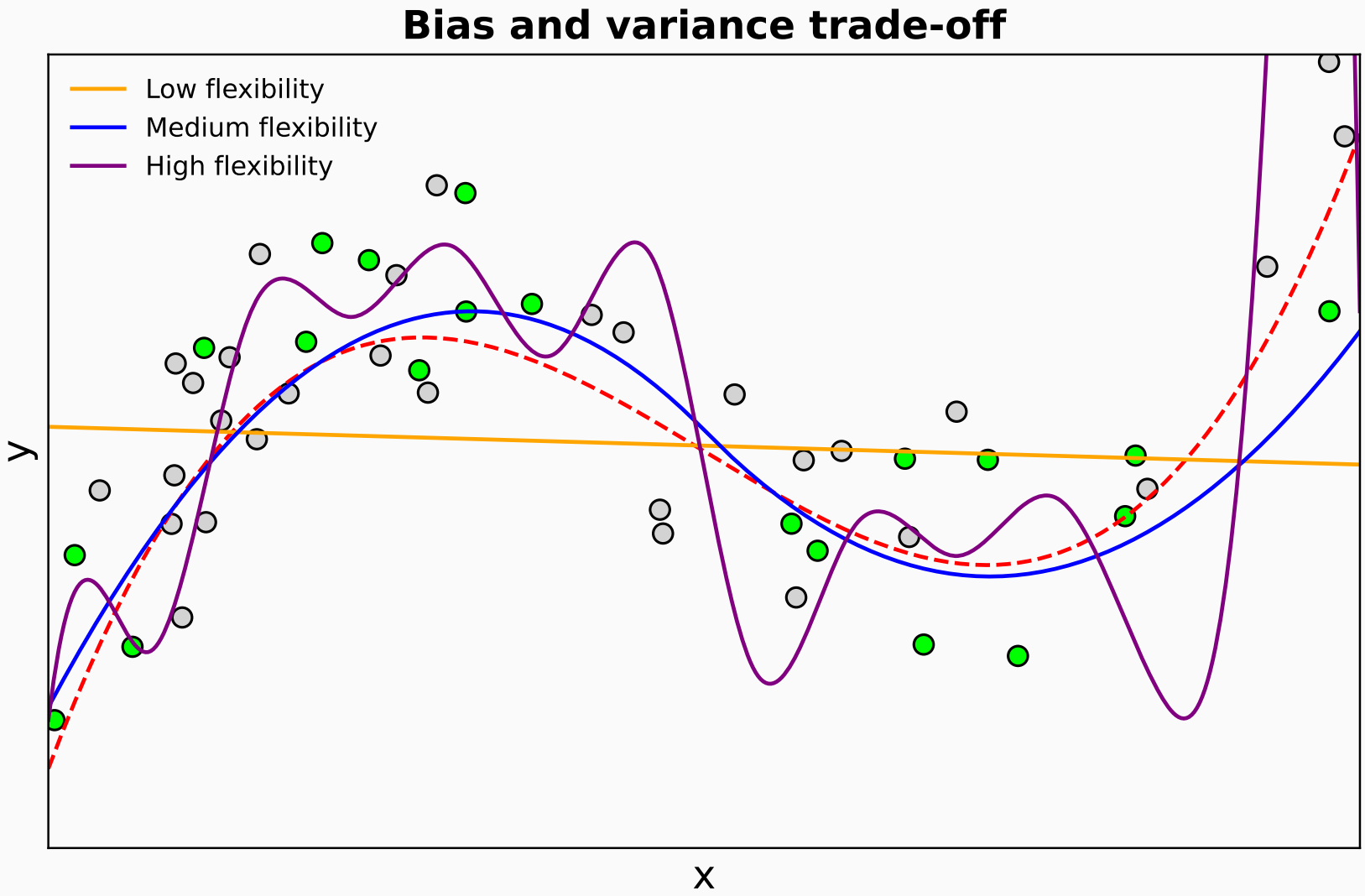
Bias and Variance

$$
\begin{aligned}
&= \left(f - E[\hat{f}]\right)^2 + E(\epsilon^2) + E\left[\left(E[\hat{f}] - \hat{f}\right)^2\right] \\
&+ 2\left(f - E[\hat{f}]\right)\underline{E[\epsilon]} \\
&+ 2\underline{E[\epsilon]}E\left[E[\hat{f}] - \hat{f}\right] \\
&+ 2E\left[\left(f - E[\hat{f}]\right)\underline{\left(E[\hat{f}] - \hat{f}\right)}\right] \\
E\left[\left(y - \hat{f}\right)^2\right] &= Bias^2[\hat{f}] + Var[\hat{f}] + Var[\epsilon]
\end{aligned}
$$

Note that f is constant and $\hat{f}$ is assumed to be fixed so $E\left(\hat{f} - E(\hat{f})\right) = 0$. If the error is truly random $E(\epsilon) = 0$

Bias-variance trade-off

Bias-variance trade-off



Bias-variance trade-off

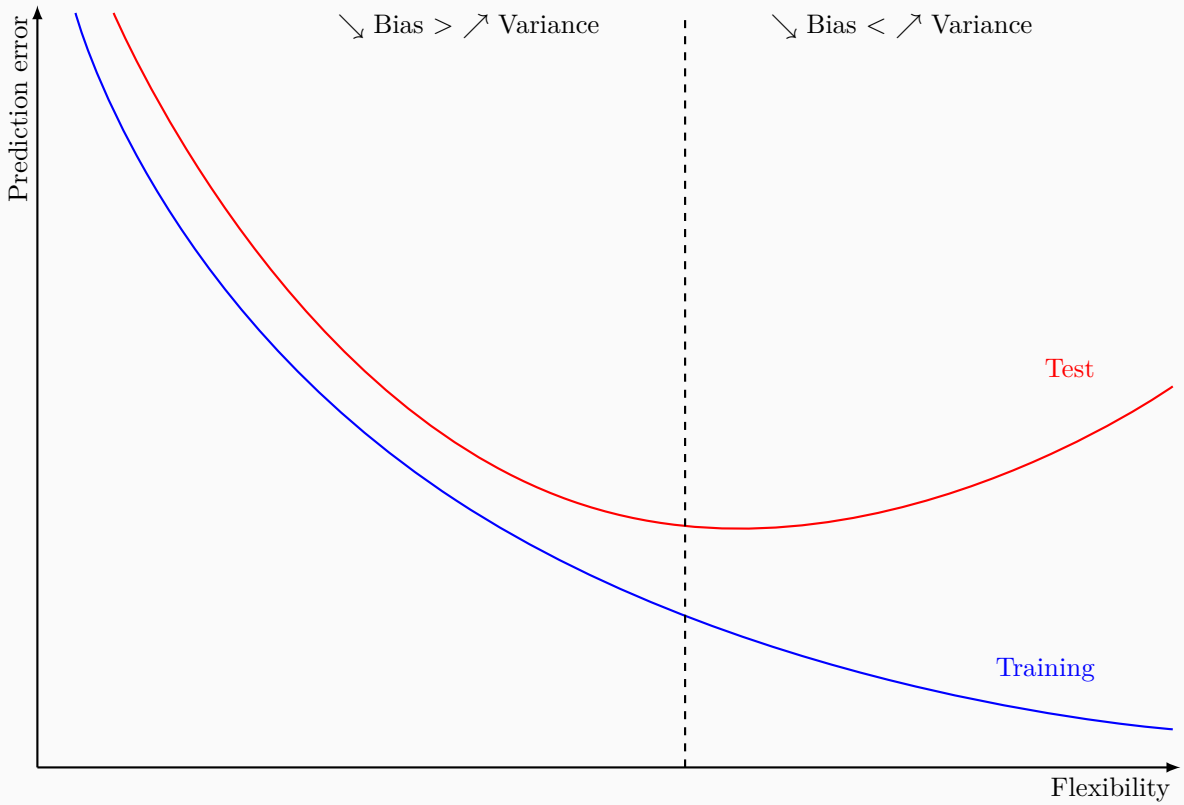
As we add flexibility to the model (i.e. more parameters) the bias decreases and the variance increases

Bootstrapped bias and variance

Flexibility	MSE	Bias^2	Var.	Δ MSE	Δ Bias^2	Δ Var.
Low	18.10	16.87	1.23			
Medium	2.53	0.90	1.63	-15.57	-15.97	0.40
High	6.87	0.52	6.36	4.34	-0.38	4.73

The challenge of regularisation is to find the model that strikes the right balance between bias and variance

Training and test sample error

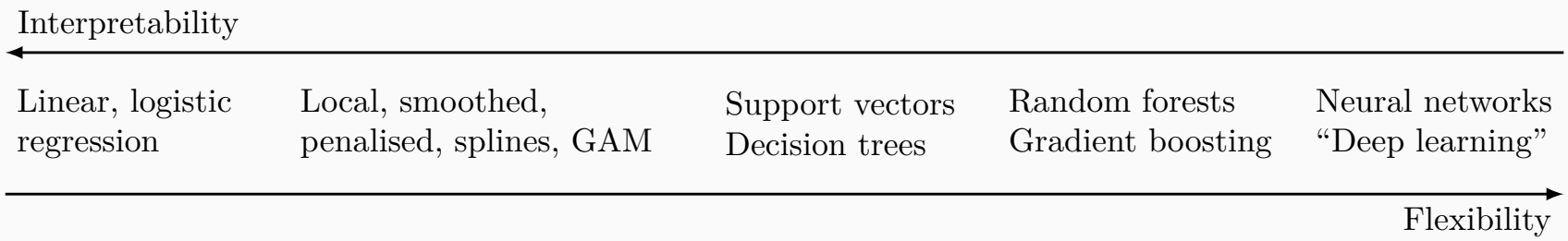


Before the dashed line, the bias decrease faster than the variance increase and the test error decreases. After that point, additional flexibility has little impact on the bias but increases variance and the error increases.

Bias-variance trade-off

Predictive models can be understood as striking a different balance between interpretability and flexibility

- Linear models are interpretable but lack the flexibility to approximate complex non-linear functions
- Flexibility means that a model with different sets of parameters can produce similar predictions



Re-sampling

Re-sampling

The statistical properties (e.g. finite sample, asymptotic) of flexible models can often not be derived analytically

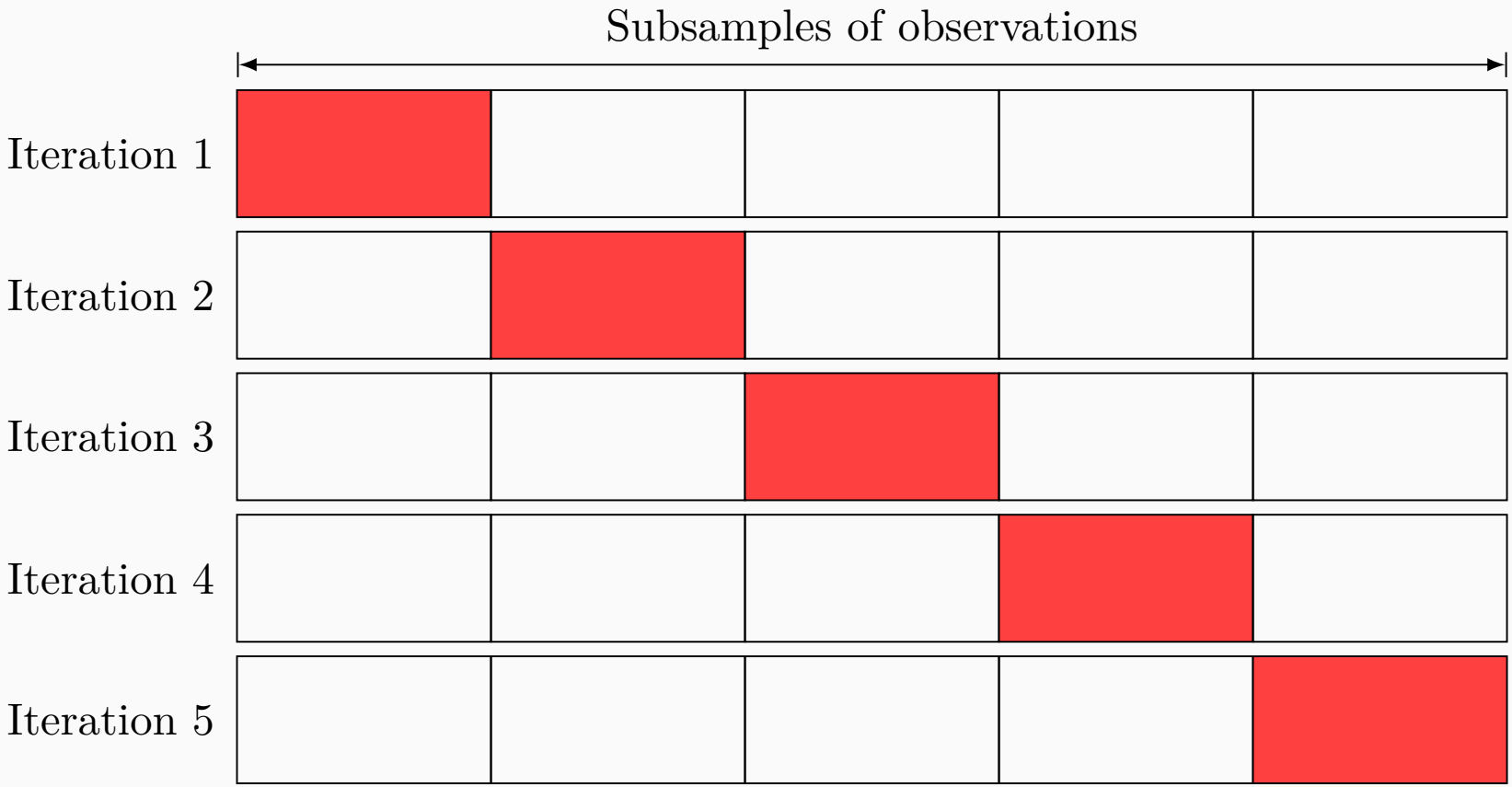
- We can repeatedly draw subsamples from a larger sample (proxy for the population) and compute statistics
- The distribution of these statistics (e.g. predictions) gives information about the model's statistical properties
- Useful techniques include cross-validation to compare models and bootstrap to discover their properties

A predictive model is estimated on a training sample and assessed on a test sample

- There is a trade-off with the partition of the observed data into the training and the test sample
- The best model is estimated using all the observed data, but the performance estimate would be biased
- An unbiased estimate requires a sufficiently large test sample, which could be used to estimate a better model

Cross-validation (Stone 1974) is a resampling procedure that addresses both these issues simultaneously

- The observations are randomly partitioned into k distinct groups without replacement
- The model is repeatedly estimated on $k-1$ partitions and evaluated on the remaining partition
- There are k iterations, so that each observation is used $k-1$ times for training and once for testing


Subsamples of observations
Iteration 1
Iteration 2
Iteration 3
Iteration 4
Iteration 5

The cross-validated mean squared error is calculated by averaging the MSE on the k test partitions

$$\mathcal{L}_{mse}^{cv}(k) = \frac{1}{k}\sum_{1}^{k}\left[\frac{1}{n_t}\sum_{i=1}^{n_t}\left(y_i - \hat{f}(x_i)\right)^2\right]$$

where $i = 1, \ldots, n_t$ are the observations of the test fold

- The value of k is a trade-off between the computational cost and the bias of the estimator
- "Leave-one-out" cross-validation involves n folds. $k = \{5, 10\}$ provide reasonably good estimates

K-fold cross validation

```
model = Spline(d=3, df=3)
k     = 10

# Splitting folds
folds = np.split(random.permutation(n), k)
folds = [~np.isin(np.arange(n), fold) for fold in folds]
eh_cv = np.zeros(k)

# K-fold CV
for i, s in enumerate(folds) :
    model.fit(x[s], y[s])
    eh_cv[i] = np.mean((y[~s] - model.predict(x[~s]))**2)
np.mean(eh_cv)
```

Bootstrap (Efron and Tibshirani 1994) is used to estimate standard errors for the estimated parameters and response

- Non-parametric methods do not provide naturally standard errors for the estimates
- We want to draw additional samples from the population, yet we have only have a single one
- An estimate can computed by sampling repeatedly with replacement (to ensure independence across samples) observations from that sample

For instance, standard error of an estimated parameter $\hat{\beta}$ can be estimated via bootstrap using

$$se(\hat{\beta}) = \sqrt{\frac{\sum_{b=1}^{B} \left(\hat{\beta}_b - \bar{\hat{\beta}}\right)^2}{B-1}} \tag{8}$$

$$\bar{\hat{\beta}} = \sum_{b=1}^{B} \frac{\hat{\beta}_b}{B}$$

where B is the number of bootstraps

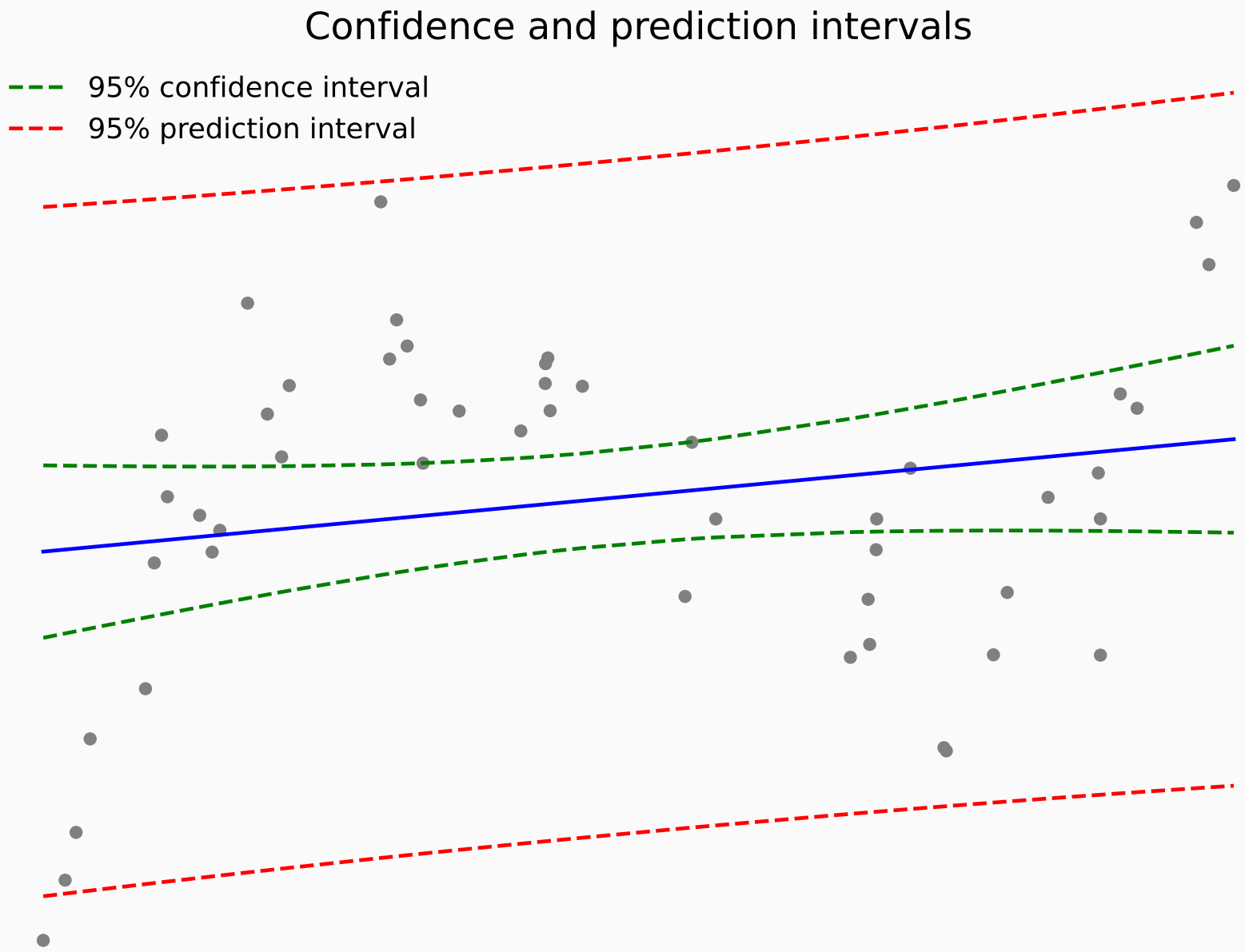
- There is no agreed convention concerning the size of the subsample and the number of bootstraps
- The bias vs. computation trade-off also applies to bootstrap methods (more is better)

Bootstrap

```
nb = 100
ns = 40
samples = [random.choice(np.arange(ns), ns) for _ in range(nb)]

bh_bs = np.zeros((nb, 2))
for i, sample in enumerate(samples):
    bh_bs[i] = model.fit(x[sample], y[sample])

se_bs = np.sum((bh_bs - np.mean(bh_bs, axis=0))**2, axis=0)
se_bs = np.sqrt(se_bs / (nb-1))
se_bs.round(4)
# array([0.7249, 0.4965])
```


Confidence and prediction intervals
95% confidence interval
95% prediction interval

Summary

Summary

Supervised learning and econometrics can be seen as function approximation methods. You may use the former when

- Prediction accuracy is more important than the interpretability of estimated parameters
- No functional form is suggested by theory and no sensible parametric model can be written
- The target function is non-linear, potentially high-dimensional and there are many observations

Summary

Every statistical model can be seen as striking a different balance between prediction bias and variance

- This trade-off can also be interpreted as a compromise between flexibility and interpretability
- The model choice depends on the research problem and the nature of the target function
- Re-sampling methods provide additional information about the statistical properties of a model

Thank you for your attention!

References i

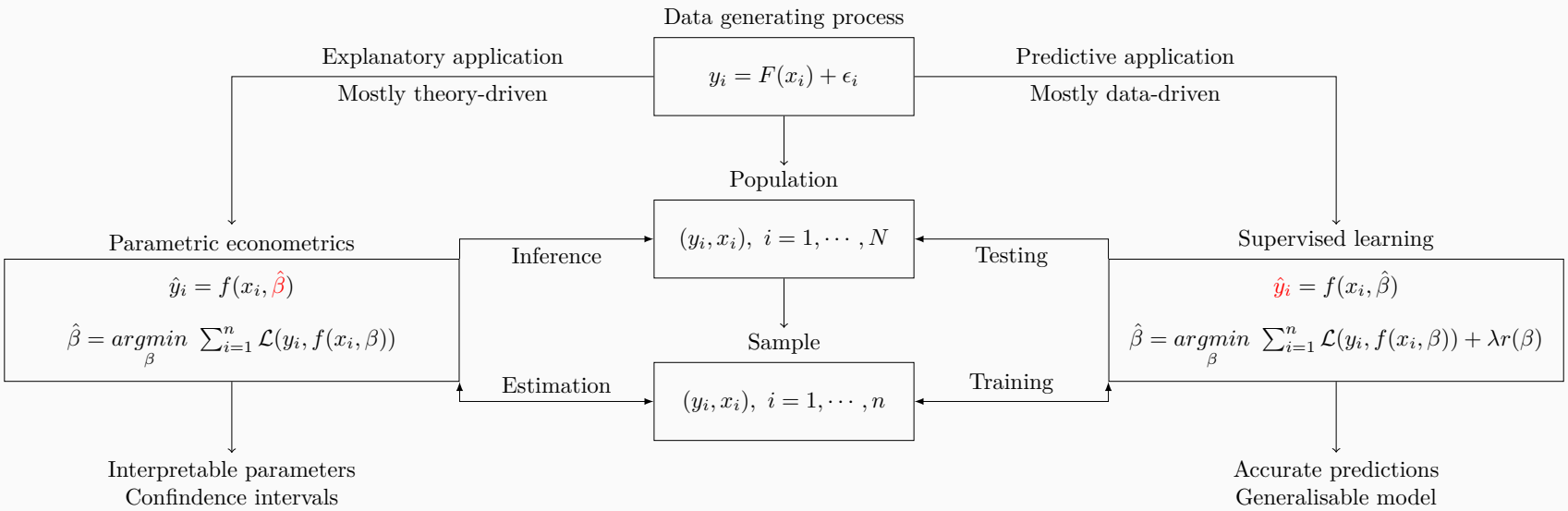
- Stone, Mervyn (1974). **"Cross-validatory choice and assessment of statistical predictions".** In: *Journal of the Royal Statistical Society* 36.2, pp. 111–147 (cit. on p. 51).
- Harrison, David J. and Daniel L. Rubinfeld (1978). **"Hedonic housing prices and the demand for clean air".** In: *Journal of Environmental Economics and Management* 5.1, pp. 81–102 (cit. on p. 68).
- Efron, Bradley and Robert Tibshirani (1994). ***An introduction to the bootstrap*.** CRC Press (cit. on p. 55).
- Breiman, Leo (2001). **"Statistical modeling: The two cultures".** In: *Statistical Science* 16.3, pp. 199–231.

References ii

- Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). ***The elements of statistical learning.*** Springer.
- Mullainathan, Sendhil and Jann Spiess (2017). **"Machine learning: An applied econometric approach".** In: *Journal of Economic Perspective* 31.2, pp. 87–106.

Appendix

Inference and prediction problems are best solved using specific methods



◂ Presentation

Estimators can be considered as random variables with the estimated response being its realisation

- A random variable is a series of realisations (i.e. estimates) that take real values with a probability
- With a sample we observe one realisation, one value out of the set of possible values (i.e. random samples)
- These values and the associated probabilities are respesented as a probability density function

Various predictive models for a hedonic price model on the Boston dataset (Harrison and Rubinfeld 1978)

Model	R^2 train	R^2 test
Least-squares	74.14	72.20
GAM	87.63	83.41
Random forest	97.76	88.26
Neural networks	98.86	86.36

▸ Application

	Count	Mean	Std.	Min.	25%	50%	75%	Max.
medv	506	22.53	9.20	5.00	17.02	21.20	25.00	50.00
crim	506	3.61	8.60	0.01	0.08	0.26	3.68	88.98
zn	506	11.36	23.32	0.00	0.00	0.00	12.50	100.00
indus	506	11.14	6.86	0.46	5.19	9.69	18.10	27.74
chas	506	0.07	0.25	0.00	0.00	0.00	0.00	1.00
nox	506	0.55	0.12	0.38	0.45	0.54	0.62	0.87
rm	506	6.28	0.70	3.56	5.89	6.21	6.62	8.78
age	506	68.57	28.15	2.90	45.02	77.50	94.07	100.00
dis	506	3.80	2.11	1.13	2.10	3.21	5.19	12.13
rad	506	9.55	8.71	1.00	4.00	5.00	24.00	24.00
tax	506	408.24	168.54	187.00	279.00	330.00	666.00	711.00
ptratio	506	18.46	2.16	12.60	17.40	19.05	20.20	22.00
black	506	356.67	91.29	0.32	375.38	391.44	396.22	396.90
lstat	506	12.65	7.14	1.73	6.95	11.36	16.96	37.97