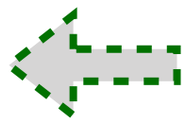
Machine Learning for Physical Scientists

Lecture 1

Introduction to Statistical Learning Theory

This Course

Inference/Learning

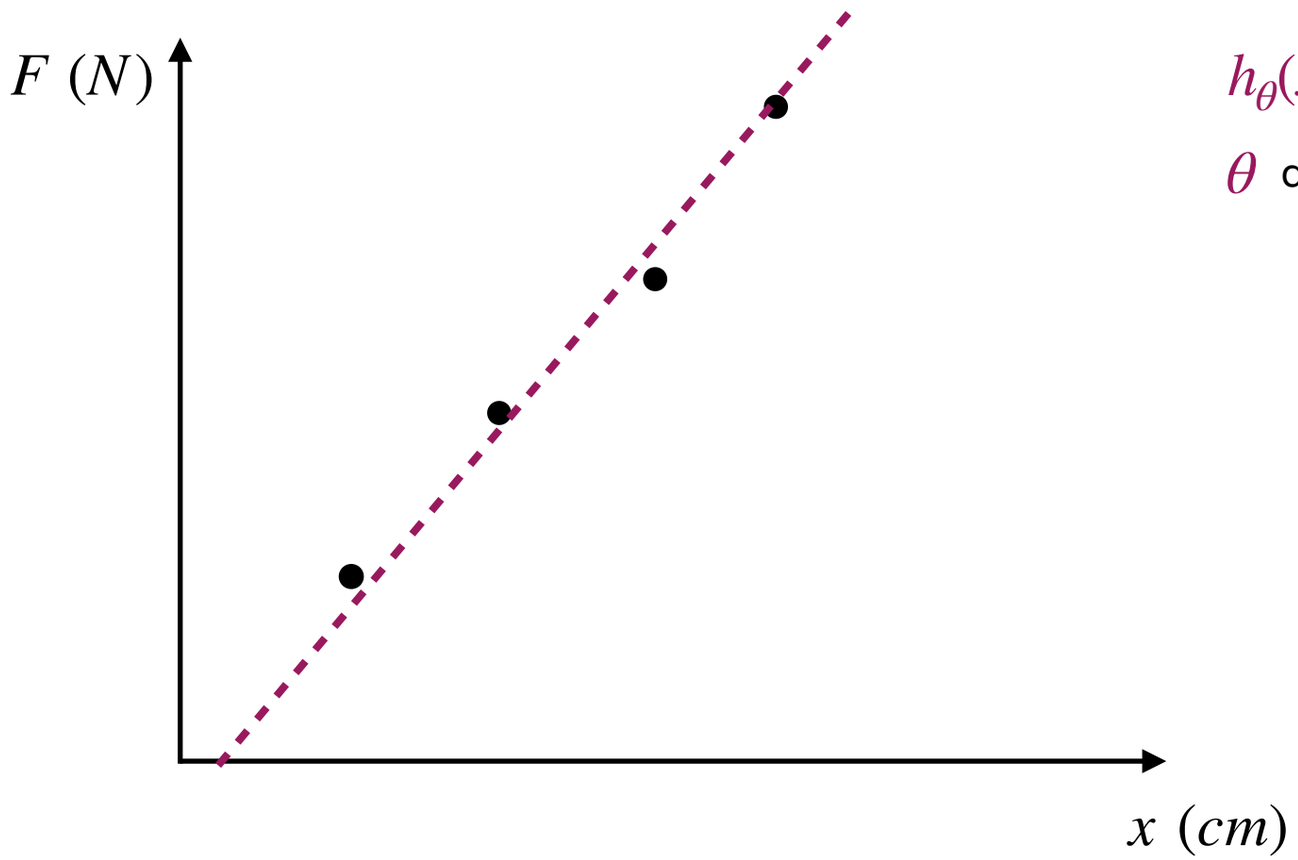


Rules or Patterns

Data

Physics 101 example

Finding the relationship between force vs displacement of a spring in equilibrium

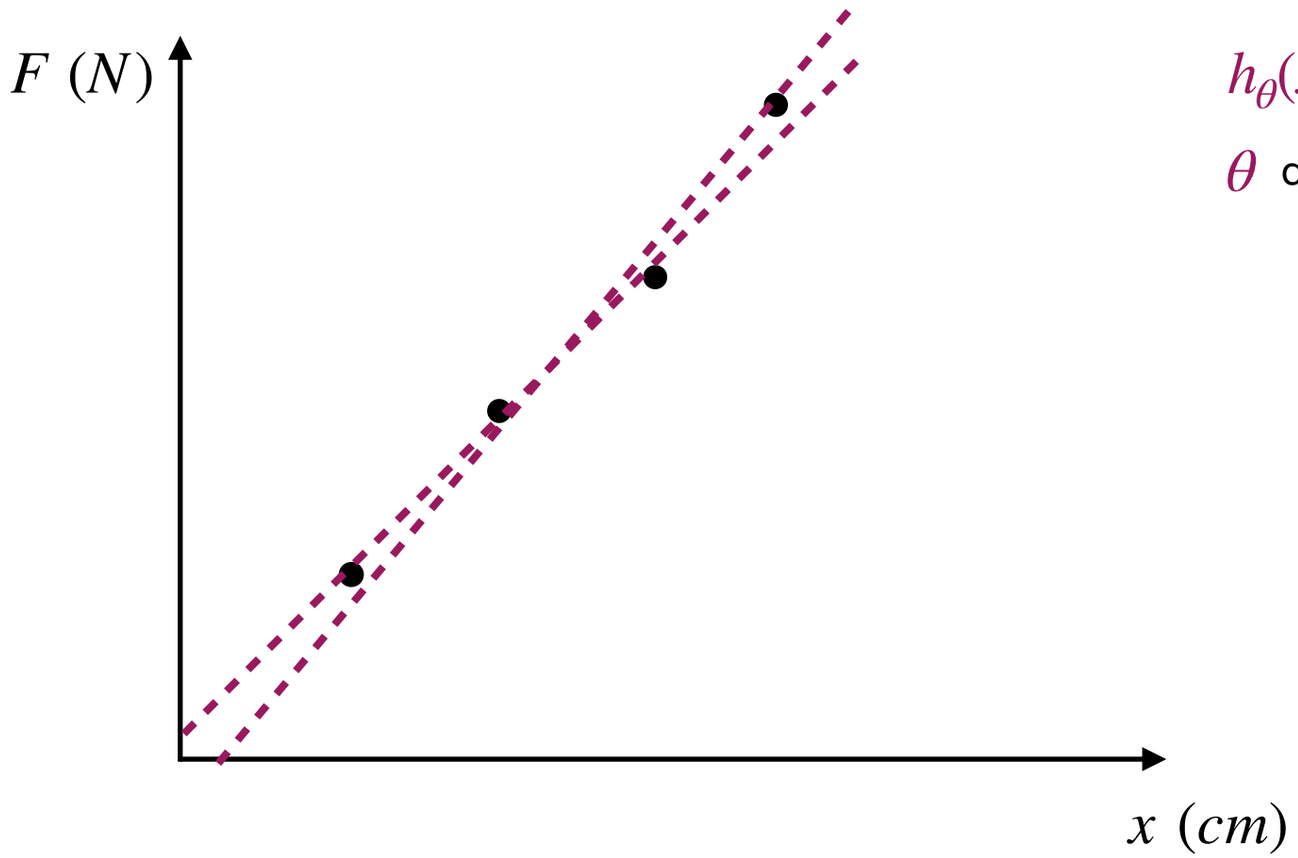


$$h_\theta(x) = kx + b$$

θ denotes a set of parameters, here *k* and *b*

Physics 101 example

Finding the relationship between force vs displacement of a spring in equilibrium

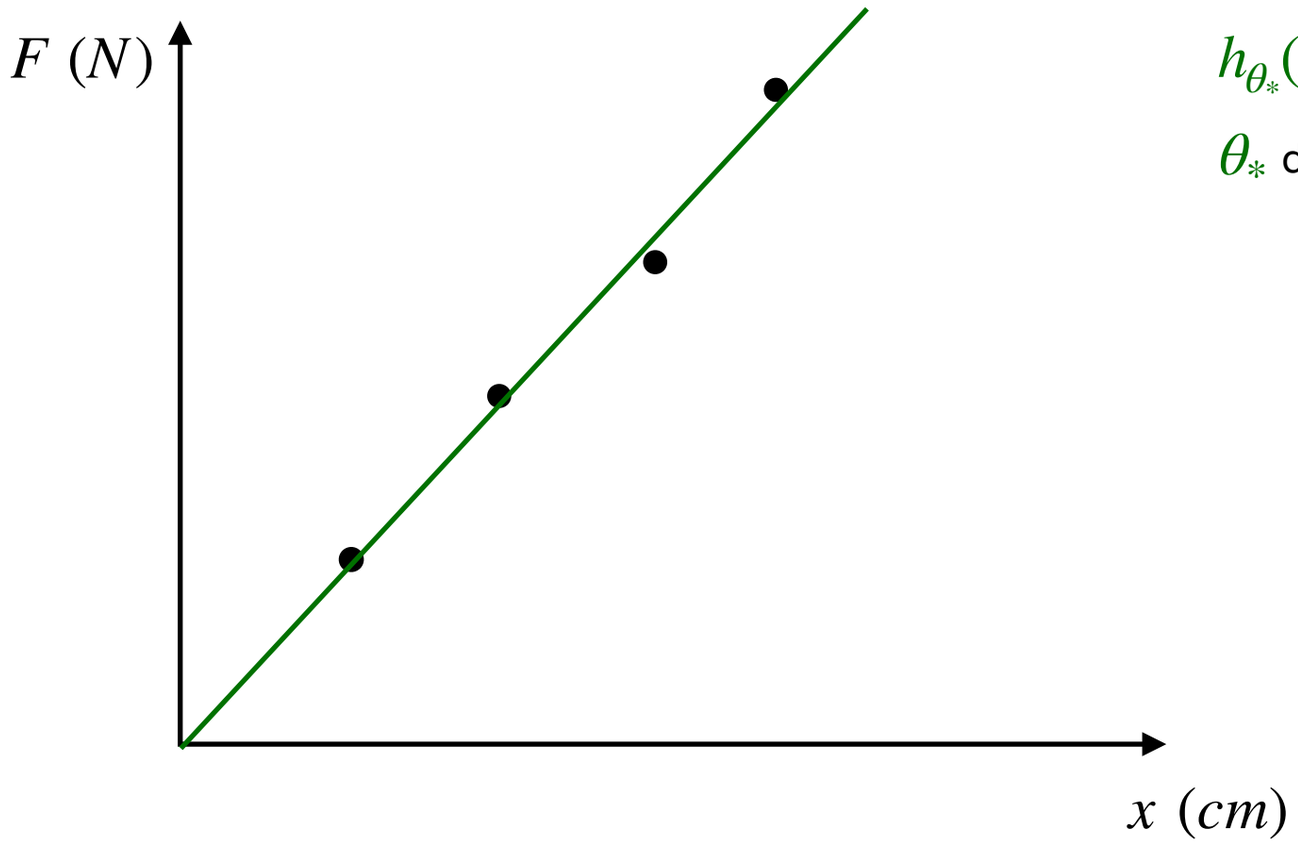


$$h_\theta(x) = kx + b$$

θ denotes a set of parameters, here *k* and *b*

Physics 101 example

Finding the relationship between force vs displacement of a spring in equilibrium



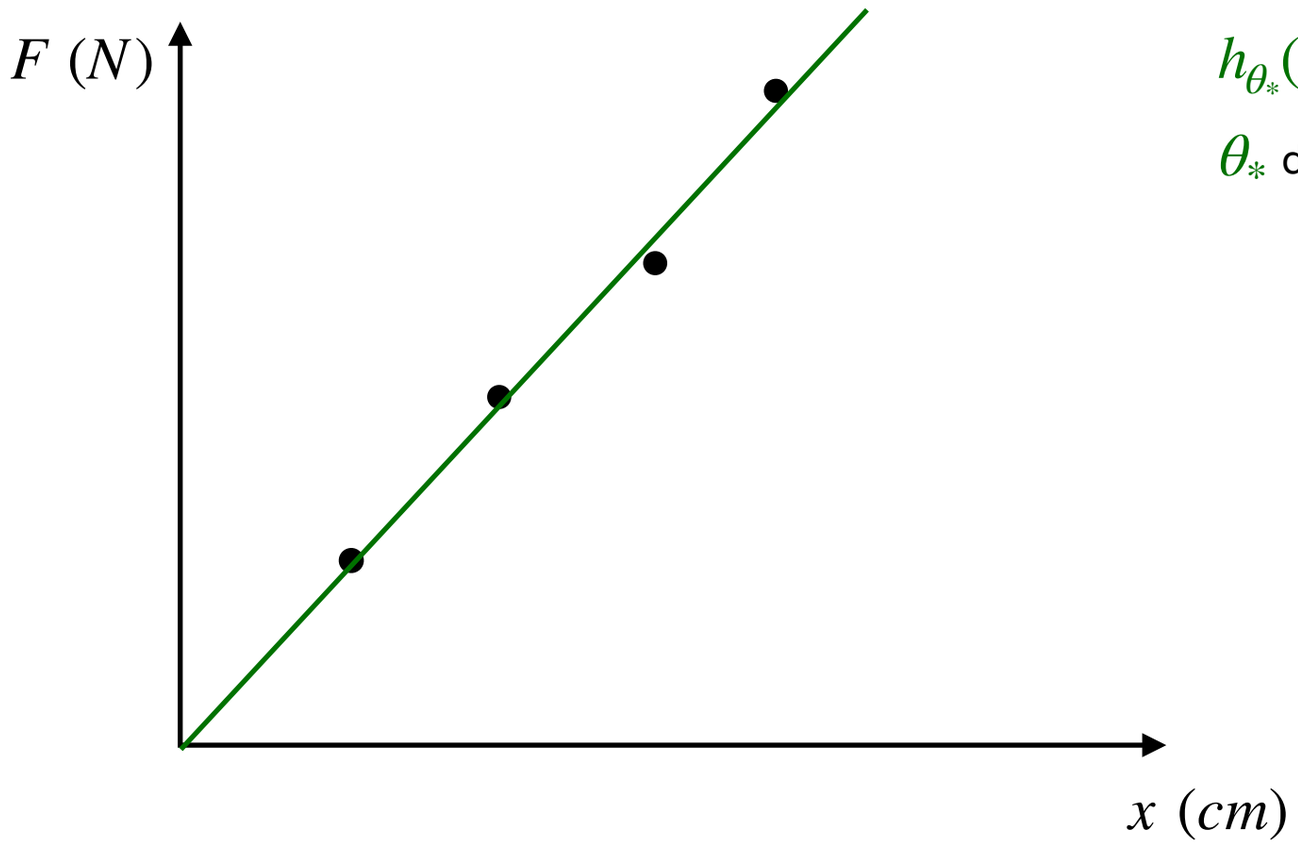
$h_{\theta_*}(x) = k_* x$

θ_* denotes the "optimal" parameters, here k_* and $b_* = 0$

Physics 101 example

Finding the relationship between force vs displacement of a spring in equilibrium

Is this hypothesis convincing?



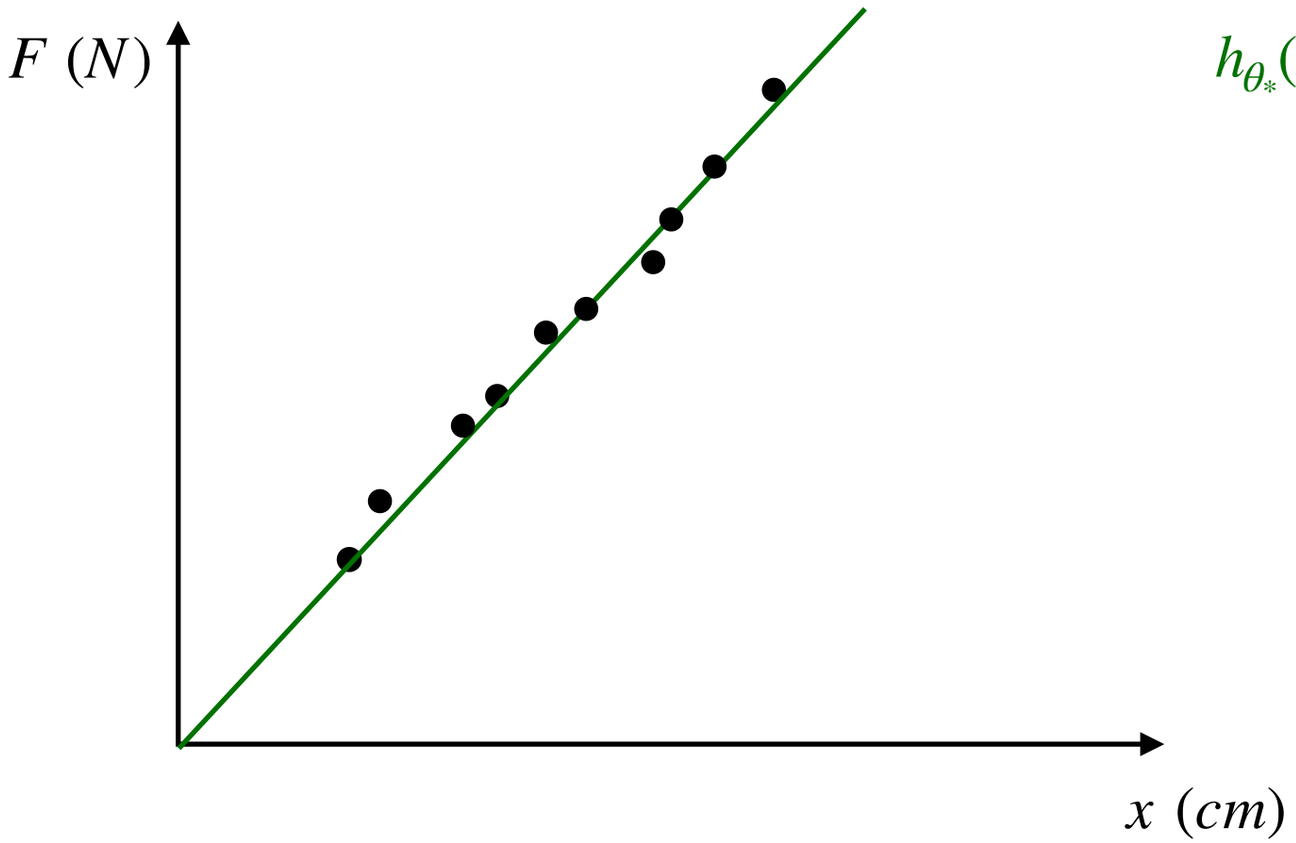
$h_{\theta_*}(x) = k_* x$

θ_* denotes the "optimal" parameters, here k_* and $b_* = 0$

Physics 101 example

Finding the relationship between force vs displacement of a spring in equilibrium

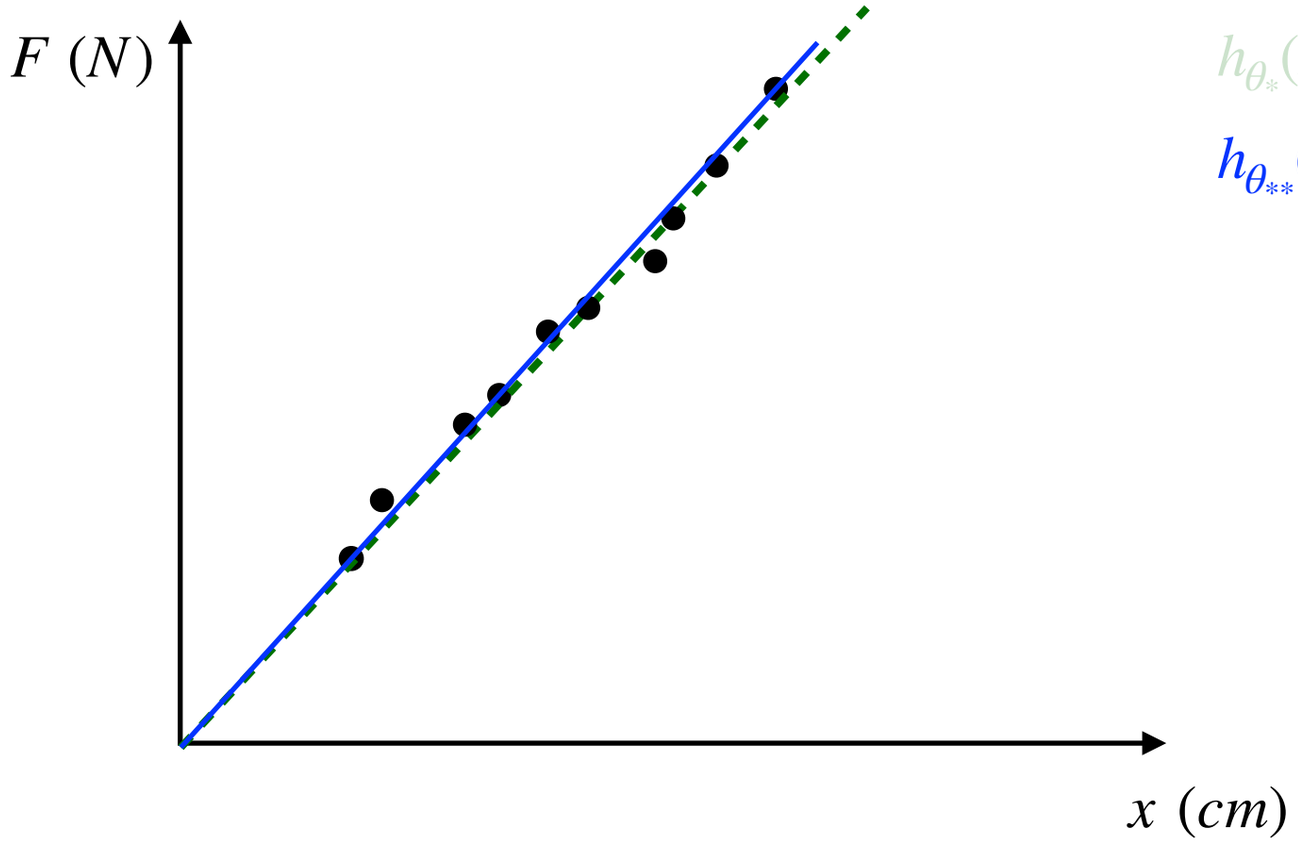
sample more data!



$$h_{\theta_*}(x) = k_* x$$

Physics 101 example

Finding the relationship between force vs displacement of a spring in equilibrium



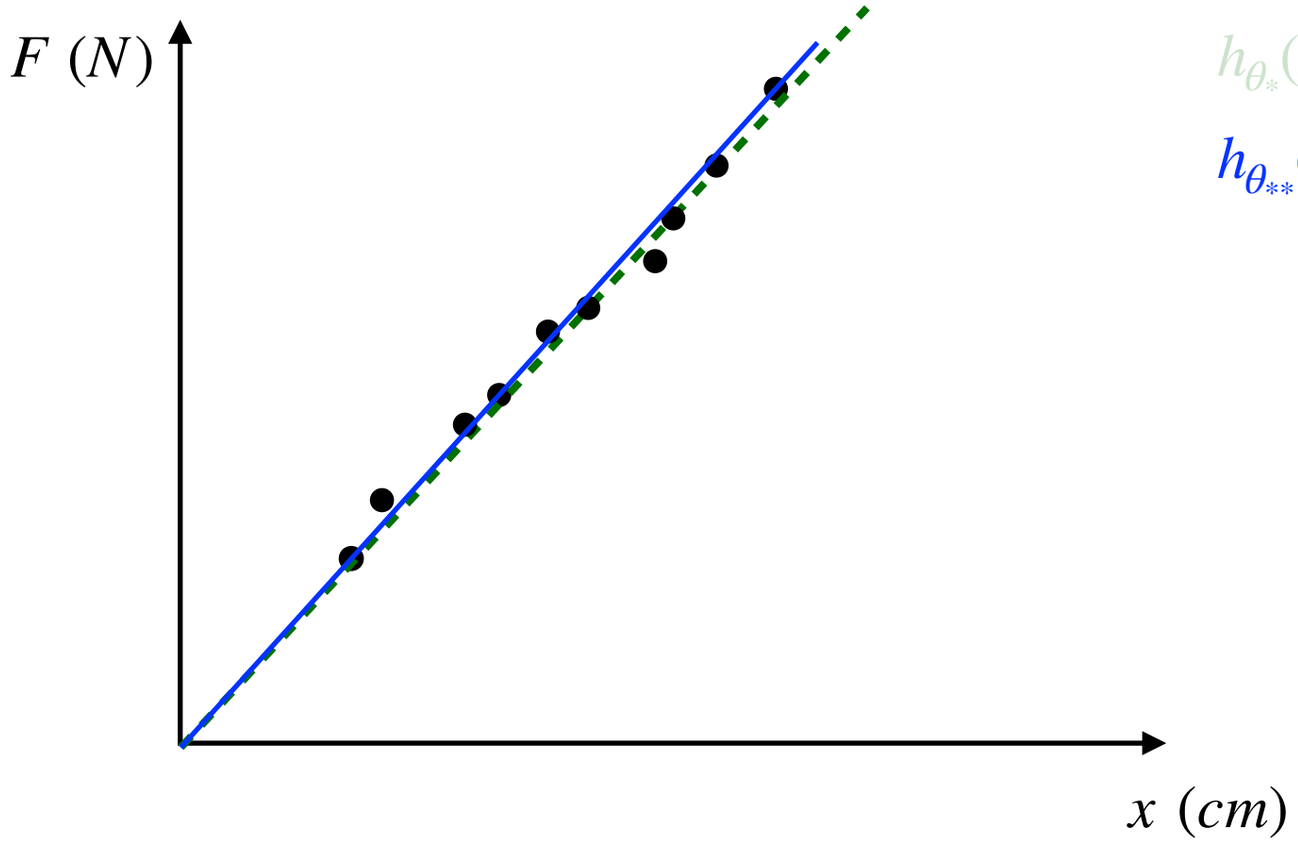
$h_{\theta_*}(x) = k_* x$

$h_{\theta_{**}}(x) = k_{**} x$

Physics 101 example

Finding the relationship between force vs displacement of a spring in equilibrium

Is this hypothesis convincing?



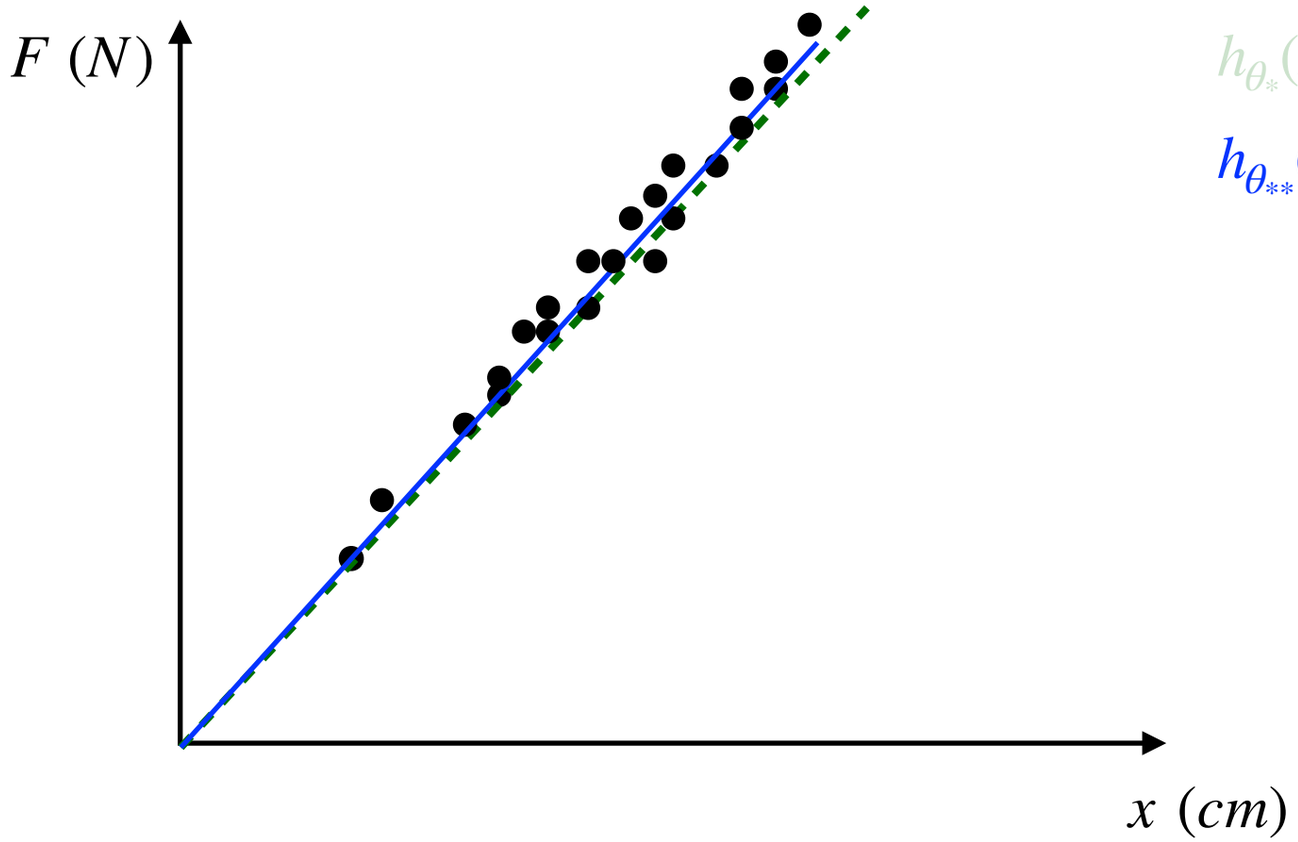
$h_{\theta_*}(x) = k_* x$

$h_{\theta_{**}}(x) = k_{**} x$

Physics 101 example

Finding the relationship between force vs displacement of a spring in equilibrium

sample more data!

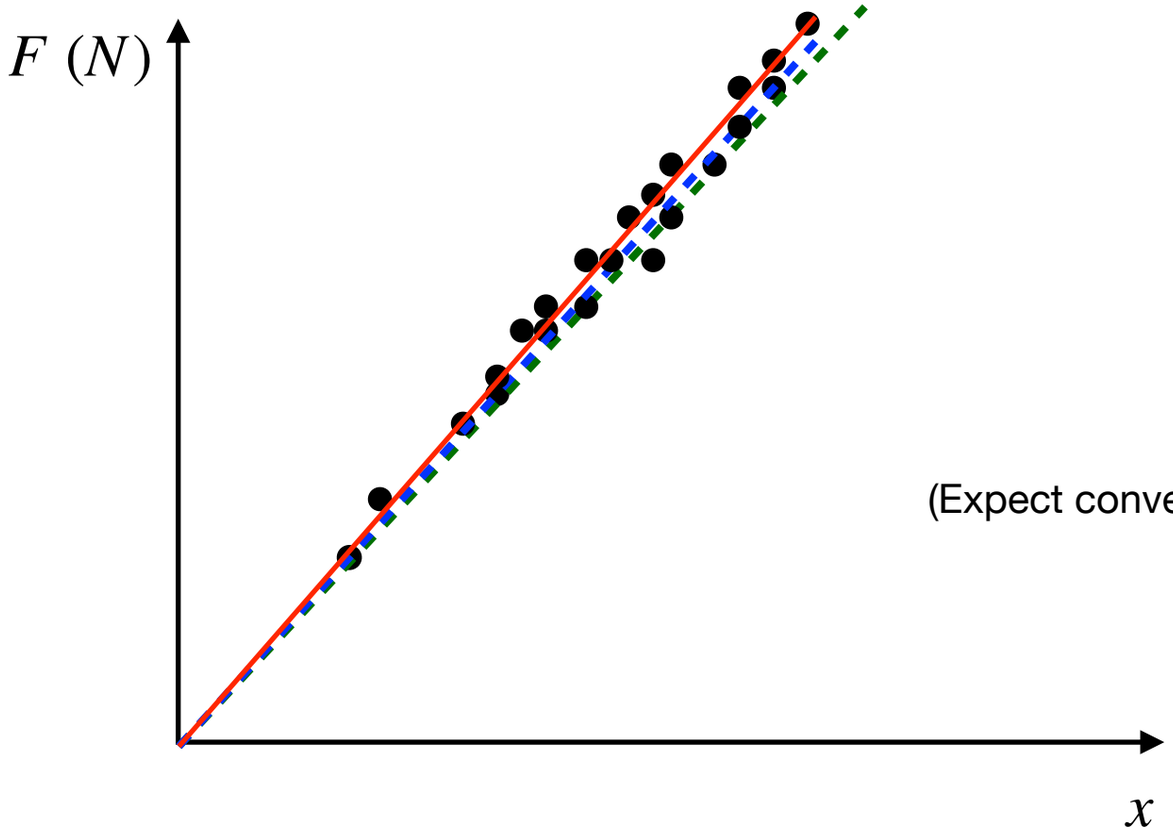


$h_{\theta_*}(x) = k_* x$

$h_{\theta_{**}}(x) = k_{**} x$

Physics 101 example

Finding the relationship between force vs displacement of a spring in equilibrium



$$h_{\theta_*}(x) = k_* x$$

$$h_{\theta_{**}}(x) = k_{**} x$$

...

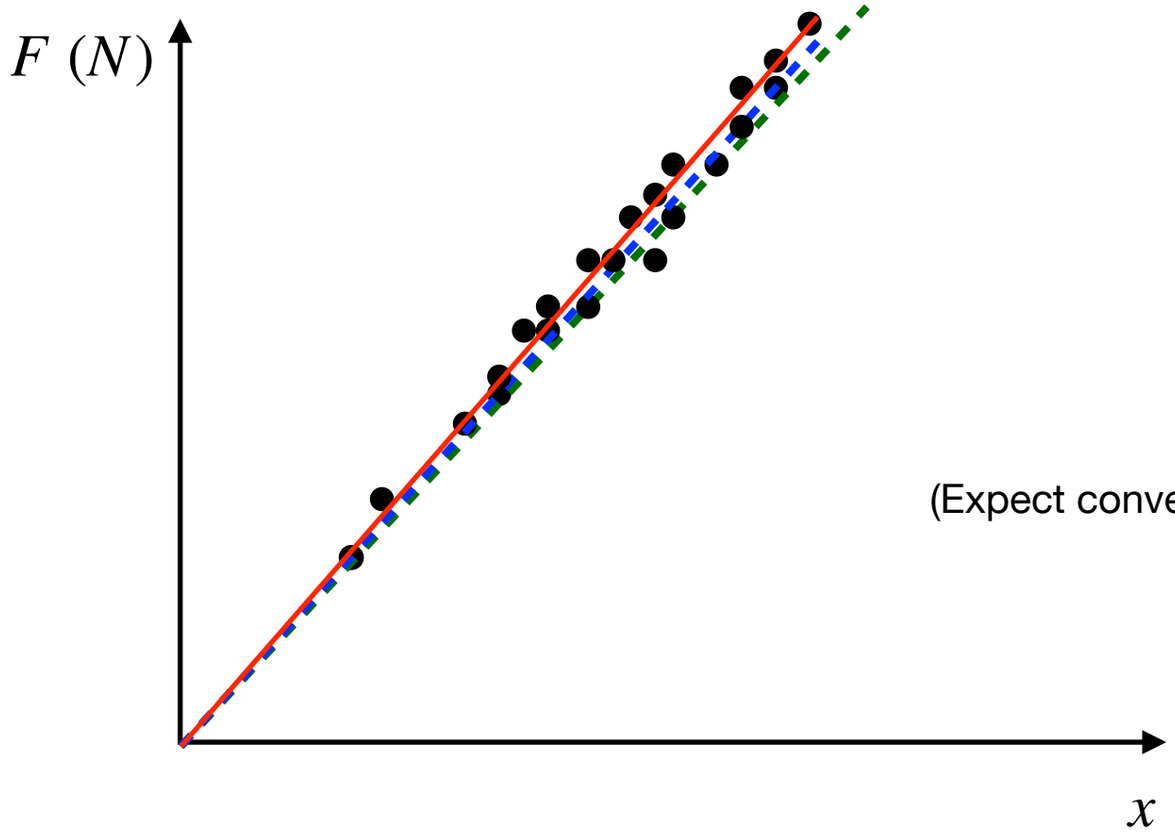
$$\tilde{h}_{\theta_*}(x) = \tilde{k}_* x$$

(Expect convergence in the limit of large sample size)

Physics 101 example

Finding the relationship between force vs displacement of a spring in equilibrium

Is this hypothesis convincing?



$h_{\theta_*}(x) = k_* x$

$h_{\theta_{**}}(x) = k_{**} x$

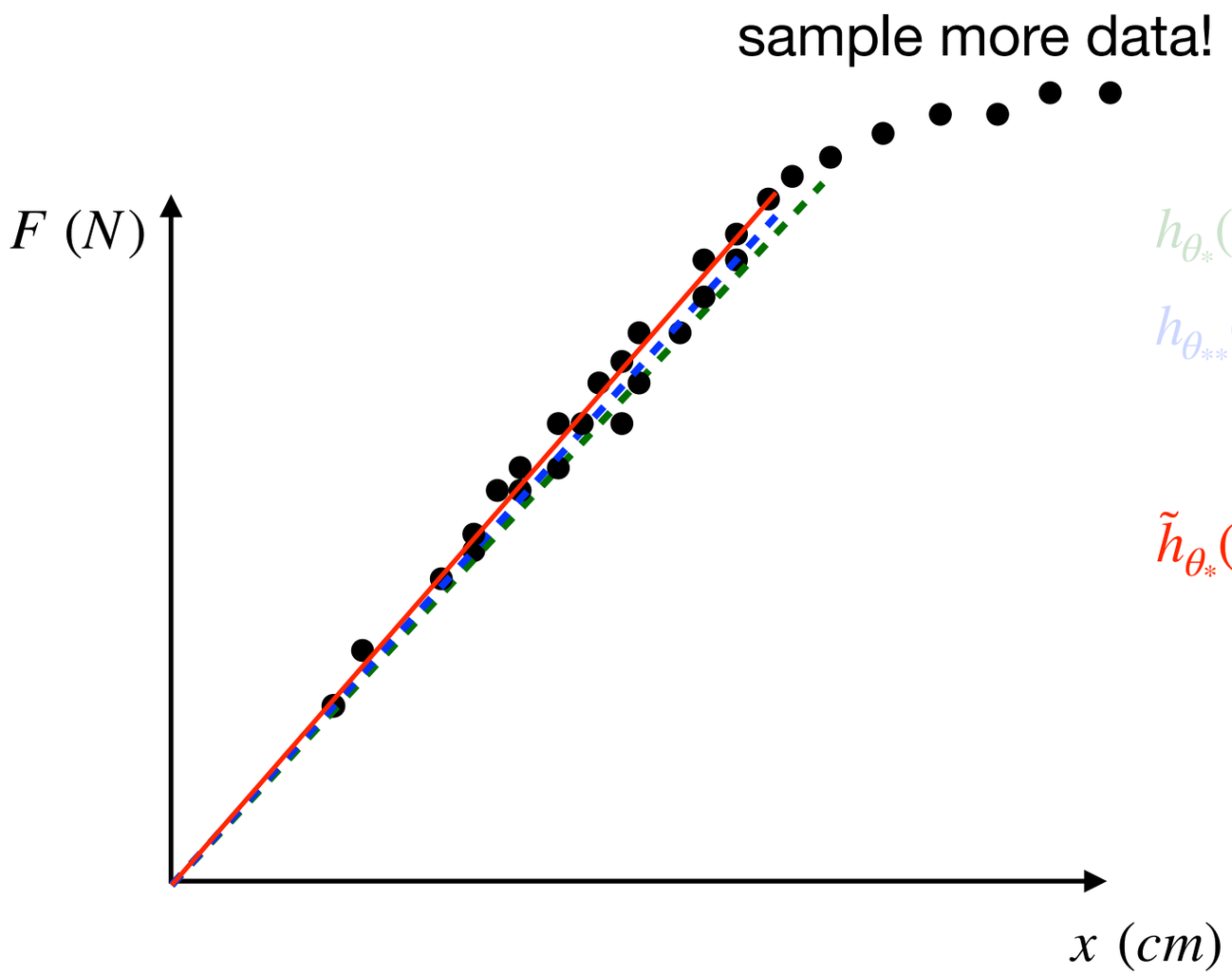
…

$\tilde{h}_{\theta_*}(x) = \tilde{k}_* x$

(Expect convergence in the limit of large sample size)

Physics 101 example

Finding the relationship between force vs displacement of a spring in equilibrium

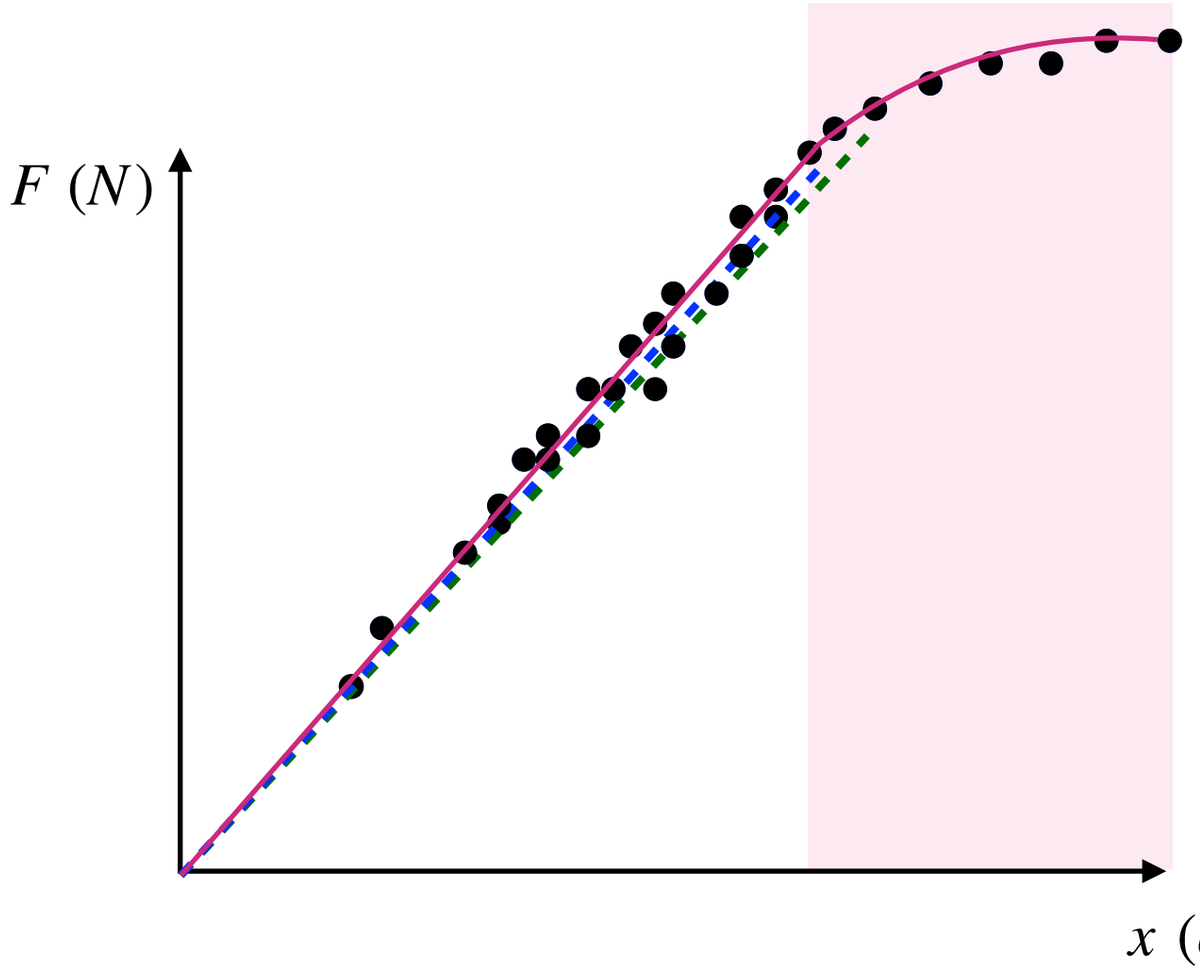


The previously optimal hypothesis *doesn't generalise well*!

(Typically is the case for physical laws in different length and time scales!)

Physics 101 example

Finding the relationship between force vs displacement of a spring in equilibrium



$h_{\theta_*}(x) = k_* x$

$h_{\theta_{**}}(x) = k_{**} x$

...

$\tilde{h}_{\theta_*}(x) = \tilde{k}_* x$

$\hat{h}_{\theta_*}(x) = \hat{k}_* x + \hat{n}_* x^2$

"Solid mechanics beyond linear elasticity theory"

The previously optimal hypothesis *doesn't generalise well*!

(Typically is the case for physical laws in different length and time scales!)

What can be done?

Assume a *more complex class* of model/hypothesis to *express* a non-linear relationship.

(e.g. quadratic polynomial, but not more or you'll fit the noise (overfitting)…)

It’d be useful to have an algorithm to automatically discover the relationship between observables, given that measurements of observables are subject to noise. In other words, we want an algorithm that can “learn” the functional relationship between observables subject to noisy measurements, so we can make a good prediction when we don’t have measurement data.

It'd be useful to have an algorithm to automatically discover the relationship between observables, given that measurements of observables are subject to noise. In other words, we want an algorithm that can "learn" the functional relationship between observables subject to noisy measurements, so we can make a good prediction when we don't have measurement data.

A simple instance of problem would be the following (hw 0):
Suppose we sample the pairs $\{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$ (training set) from a god-given stochastic process $y_i = f(x_i) + \eta_i$ where η_i is an i.i.d. Gaussian random variable with zero mean and variance σ^2, denoted by $\eta_i \in \mathcal{N}(0,\sigma^2)$.

It'd be useful to have an algorithm to automatically discover the relationship between observables, given that measurements of observables are subject to noise. In other words, we want an algorithm that can "learn" the functional relationship between observables subject to noisy measurements, so we can make a good prediction when we don't have measurement data.

A simple instance of problem would be the following (hw 0):
Suppose we sample the pairs $\{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$ (training set) from a god-given stochastic process $y_i = f(x_i) + \eta_i$ where η_i is an i.i.d. Gaussian random variable with zero mean and variance σ^2, denoted by $\eta_i \in \mathcal{N}(0,\sigma^2)$.

- Can one train an algorithm to express the correct relationship $y = f(x)$ in the limit $m \to \infty$?
- What if the noise σ is large, will the algorithm confuse noise with signal $f(x)$?
- What if we have a limited measurement so m is small, can we learn a meaningful relationship?
- How shall one constrain the class of hypothesis function so it generalises to unseen data well beyond its finite training set?

These questions can be answered and studied more rigorously in the framework of Statistical Learning Theory of Supervised Learning.

Applications of high-dimensional supervised learning algorithm

Facial Recognition

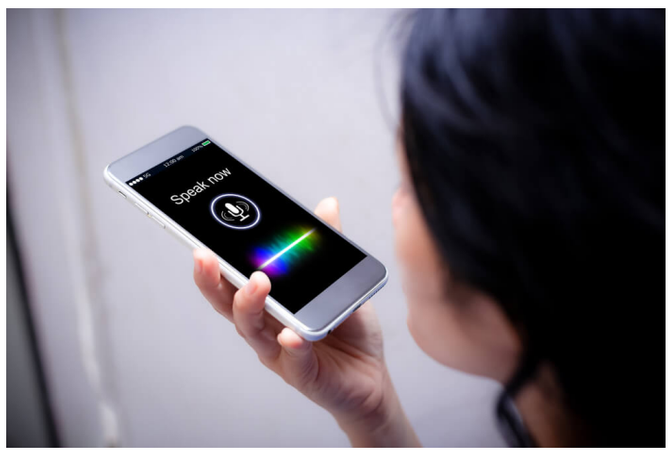
$$h_\theta(image) = sex$$

Applications of high-dimensional supervised learning algorithm

Facial Recognition

$$h_\theta(image) = sex$$

Speech Synthesis



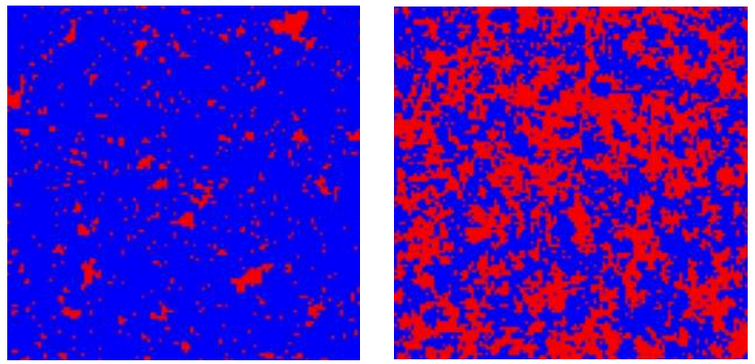
$$h_\theta(speech) = text$$

Applications of high-dimensional supervised learning algorithm

Facial Recognition

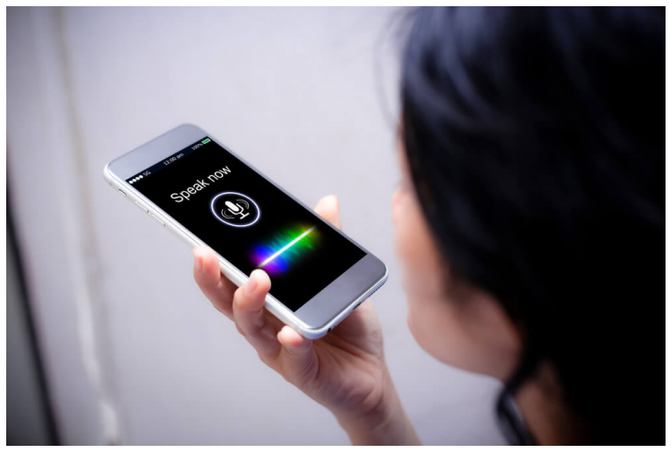
$h_\theta(image) = sex$

Phase of Matter Recognition



$h_\theta(spin\ config) = phase\ of\ matter$

Speech Synthesis

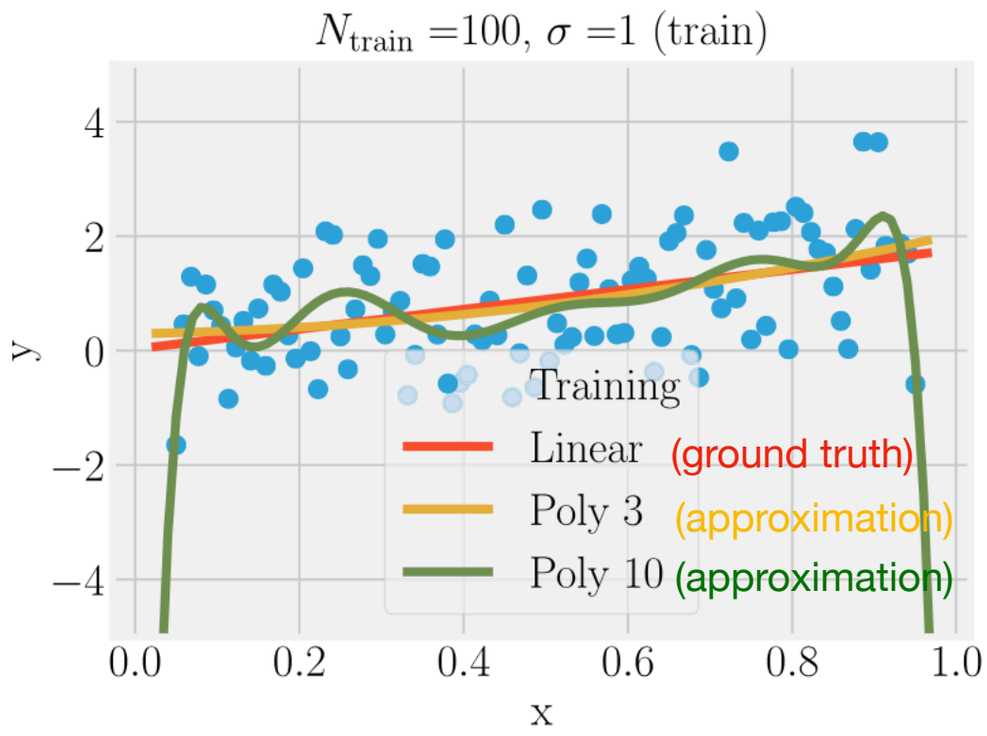


$h_\theta(speech) = text$

The main objective of Statistical Learning Theory

To make prediction effectively, not memorisation or tabulating the description.

In other words, learn from available data to effectively predict the value of function where there is no data!



$$h_\theta(image) = sex$$

Framework of Statistical Learning Theory

(supervised learning)

X: Instance Space (e.g. $\mathbb{R}^{16\times16}$ for 16x16 greyscale images)

Y: Label Space (e.g. $\mathbb{R}$ for regression or $\{1,\ldots,k\}$ for multi-class classification)

$\mathscr{D}$: Probability Distribution over $X \times Y$ (*unknown, but can sample from*)

$\ell : Y \times Y \to \mathbb{R}_{\geq 0}$ Loss or Cost Function (e.g. $\ell(y,\hat{y}) = (y-\hat{y})^2$ for $Y = \mathbb{R}$)

Framework of Statistical Learning Theory

(supervised learning)

X: Instance Space (e.g. $\mathbb{R}^{16\times16}$ for 16x16 greyscale images)

Y: Label Space (e.g. $\mathbb{R}$ for regression or $\{1,\ldots,k\}$ for multi-class classification)

$\mathscr{D}$: Probability Distribution over $X \times Y$ (*unknown, but can sample from*)

$\ell : Y \times Y \to \mathbb{R}_{\geq 0}$ Loss or Cost Function (e.g. $\ell(y,\hat{y}) = (y-\hat{y})^2$ for $Y = \mathbb{R}$)

Objective

Given a training set $S = \left\{ \left(x_i, y_i\right) \right\}_{i=1}^{m}$ drawn i.i.d. from $\mathscr{D}$, return hypothesis (predictor) $h : X \to Y$ that minimizes the population loss or expected risk:

$$L_{\mathscr{D}}(h) := \mathbb{E}_{(x,y)\sim\mathscr{D}}[\ell(y, h(x))]$$

Framework of Statistical Learning Theory

(supervised learning)

X: Instance Space (e.g. $\mathbb{R}^{16\times16}$ for 16x16 greyscale images)

Y: Label Space (e.g. $\mathbb{R}$ for regression or $\{1,\ldots,k\}$ for multi-class classification)

$\mathscr{D}$: Probability Distribution over $X \times Y$ (*unknown, but can sample from*)

$\ell : Y \times Y \to \mathbb{R}_{\geq 0}$ Loss or Cost Function (e.g. $\ell(y,\hat{y}) = (y-\hat{y})^2$ for $Y = \mathbb{R}$)

Objective

Given a training set $S = \left\{ \left(x_i, y_i\right) \right\}_{i=1}^{m}$ drawn i.i.d. from $\mathscr{D}$, return hypothesis $h : X \to Y$ that minimizes the population loss or expected risk:

$$L_{\mathscr{D}}(h) := \mathbb{E}_{(x,y)\sim\mathscr{D}}[\ell(y, h(x))]$$

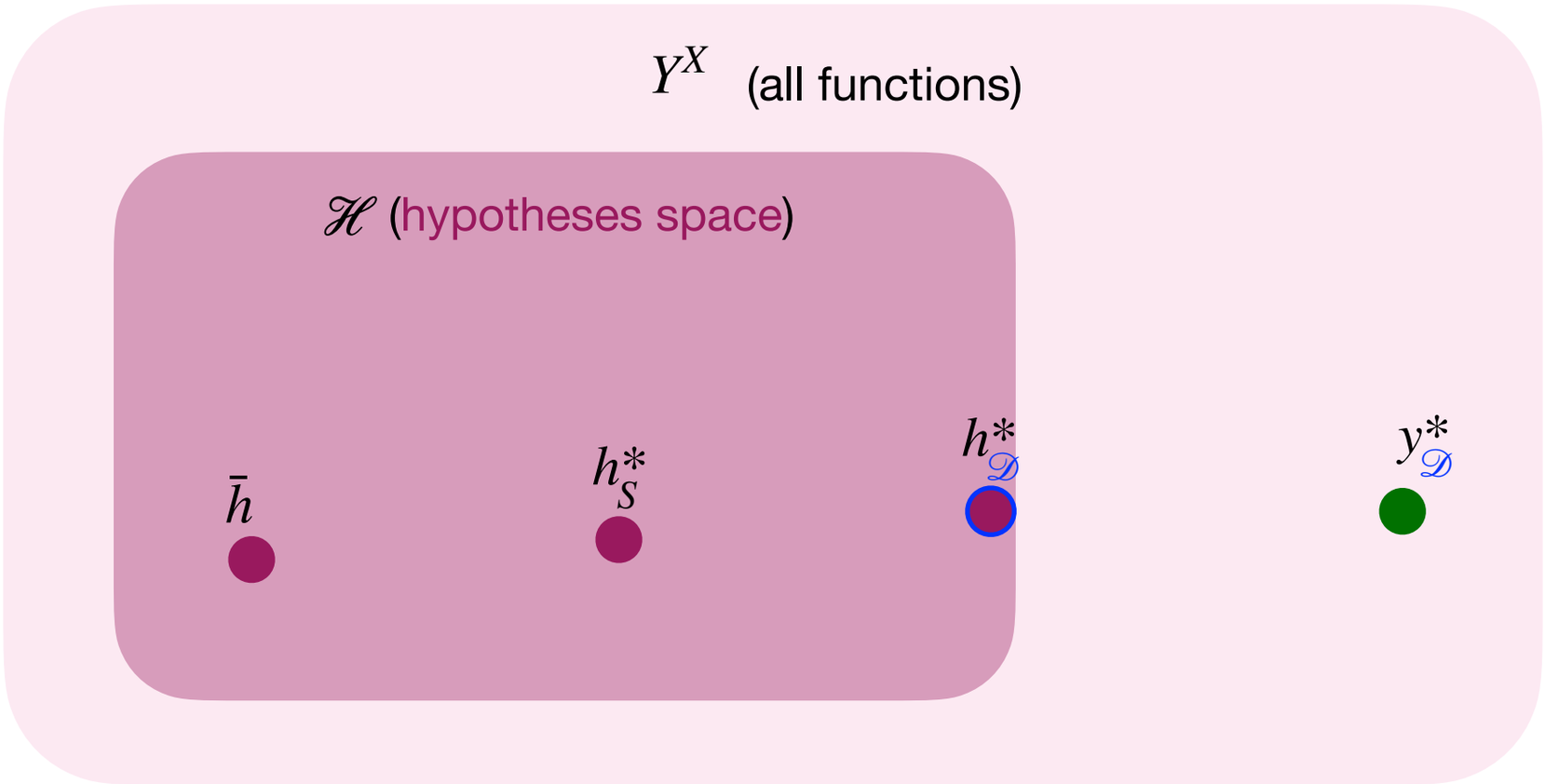
Approximate Approach

Predetermine or assume a hypotheses space $\mathscr{H} \subset Y^X$, and return hypothesis $h \in \mathscr{H}$ that minimizes sample loss or empirical loss or empirical risk:

$$L_S(h) := \frac{1}{m}\sum_{i=1}^{m} \ell\left(y_i, h\left(x_i\right)\right)$$

Jargons in Statistical Learning Theory (SLT)

(Expressiveness, Generalization, Optimization)



$y^*_{\mathscr{D}}$: ground truth (minimizer of population loss over Y^X)

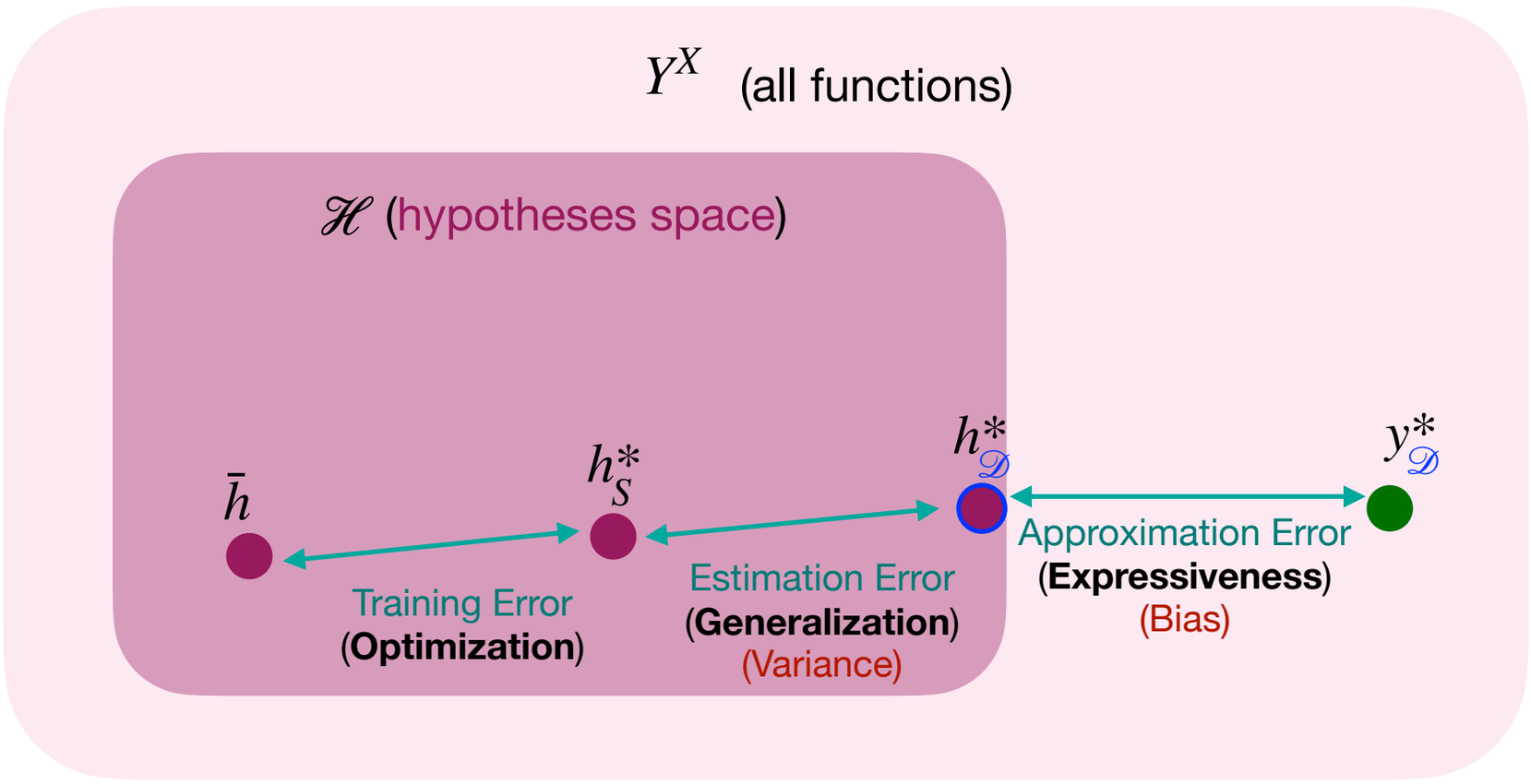
$h^*_{\mathscr{D}}$: optimal hypothesis (minimizer of population loss over $\mathscr{H}$ -infinite data sample)

h^*_S: empirically optimal hypothesis (minimizer of sample loss over $\mathscr{H}$)

$\bar{h}$: returned hypothesis

Jargons in Statistical Learning Theory (SLT)

(Expressiveness, Generalization, Optimization)



$y^*_{\mathscr{D}}$: ground truth (minimizer of population loss over Y^X)

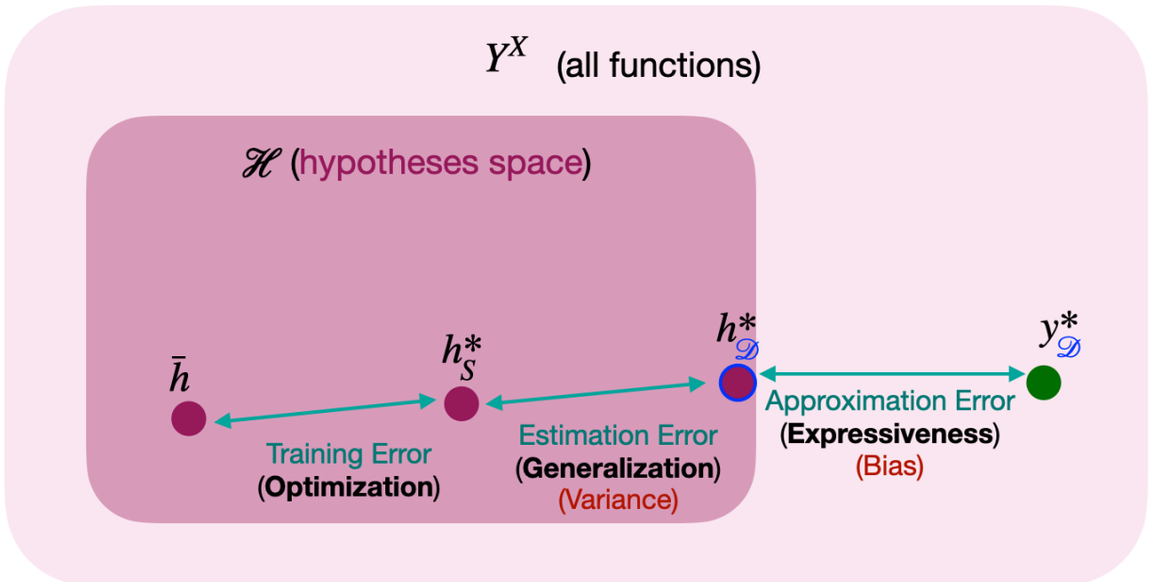
$h^*_{\mathscr{D}}$: optimal hypothesis (minimizer of population loss over $\mathscr{H}$ -infinite data sample)

h^*_S: empirically optimal hypothesis (minimizer of sample loss over $\mathscr{H}$)

$\bar{h}$: returned hypothesis

Note: For sampling to give a good proxy, we must enforce the *consistency condition* in the infinite sample size limit. Namely, $\lim_{m\to\infty} h^*_S = h^*_{\mathscr{D}}$.

Pre-Deep Learning Understanding in SLT



Optimization

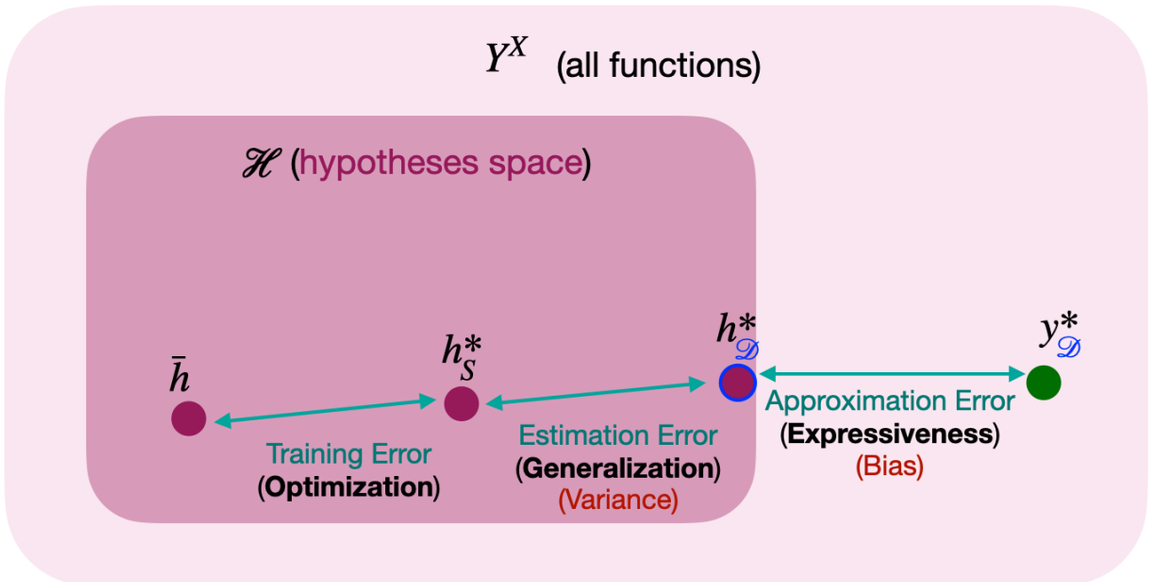
Empirical loss minimisation is a ***convex*** program: $h \approx h^*_S$ (training error ≈ 0)

Expressiveness and Generalization

Philosophical issues on "Bias-Variance trade-off"

$\mathscr{H}$	approximation error	estimation error
expand	↘	↗
shrink	↗	↘

Pre-Deep Learning Understanding in SLT



Optimization

Empirical loss minimisation is a ***convex*** program: $h \approx h_S^*$ (training error ≈ 0)

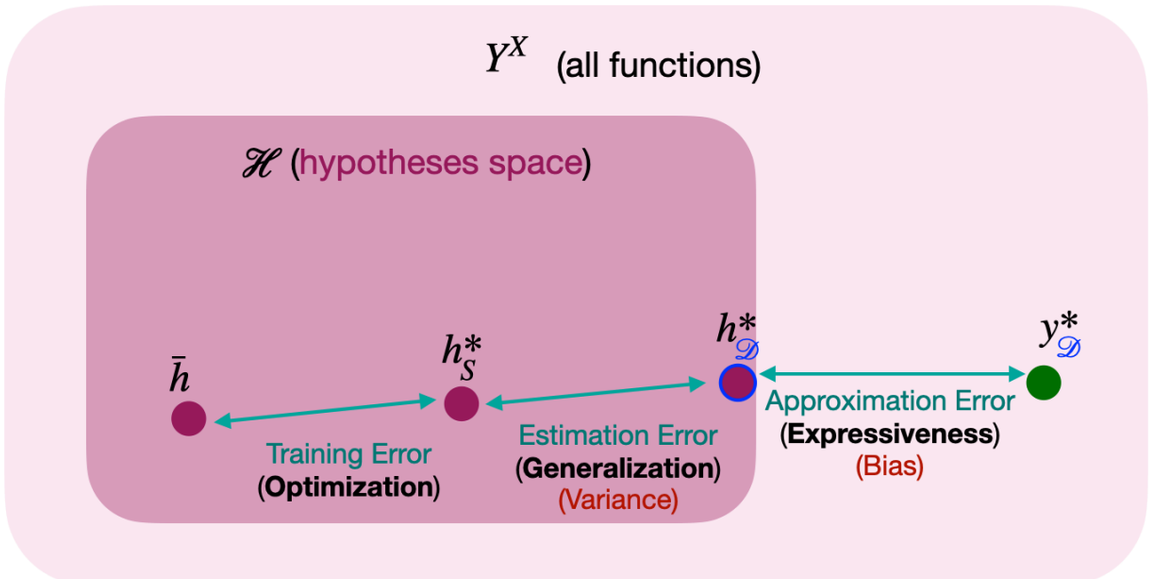
Expressiveness and Generalization

Philosophical issues on "Bias-Variance trade-off"

$\mathscr{H}$	approximation error	estimation error
expand	↘	↗
shrink	↗	↘

- High-bias: Scientific theories should be chosen from a small set of hypotheses (Occam's razor, encapsulated by VC dimensions).
- Low-variance: Theory shouldn't change much with new data coming in (stability of learning algorithm).

Deep Learning from Classical SLT Perspectives



Optimization

Empirical loss minimisation is a *non-convex* program:

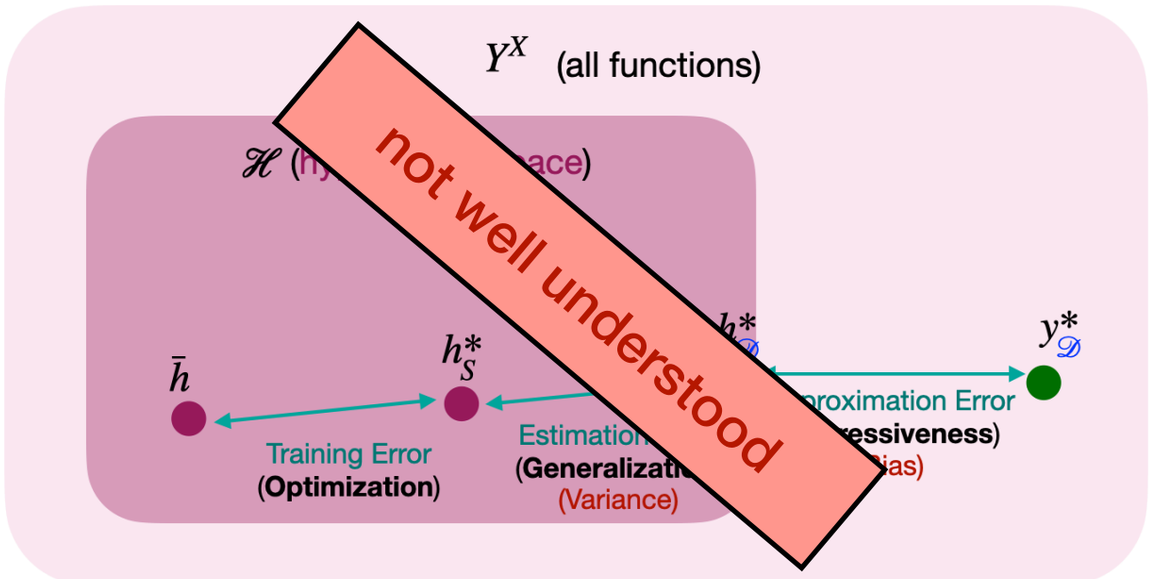
- h^*_S is not unique - many hypotheses have low training error.
- Gradient descent (GD) somehow reaches one of these.

Expressiveness and Generalization

Drastic difference from classical SLT:

- Some low training error hypotheses generalize well, others don't.
- With typical data, solution returned by GD often generalizes well.
- Expanding $\mathscr{H}$ reduces approximation error, but also estimation error!

Deep Learning from Classical SLT Perspectives



Optimization

Empirical loss minimisation is a *non-convex* program:

- h_S^* is not unique - many hypotheses have low training error
- Gradient descent (GD) somehow reaches one of these

Expressiveness and Generalization

Drastic difference from classical SLT:

- Some low training error hypotheses generalize well, others don't.
- With typical data, solution returned by GD often generalizes well.
- Expanding $\mathscr{H}$ reduces approximation error, but also estimation error!