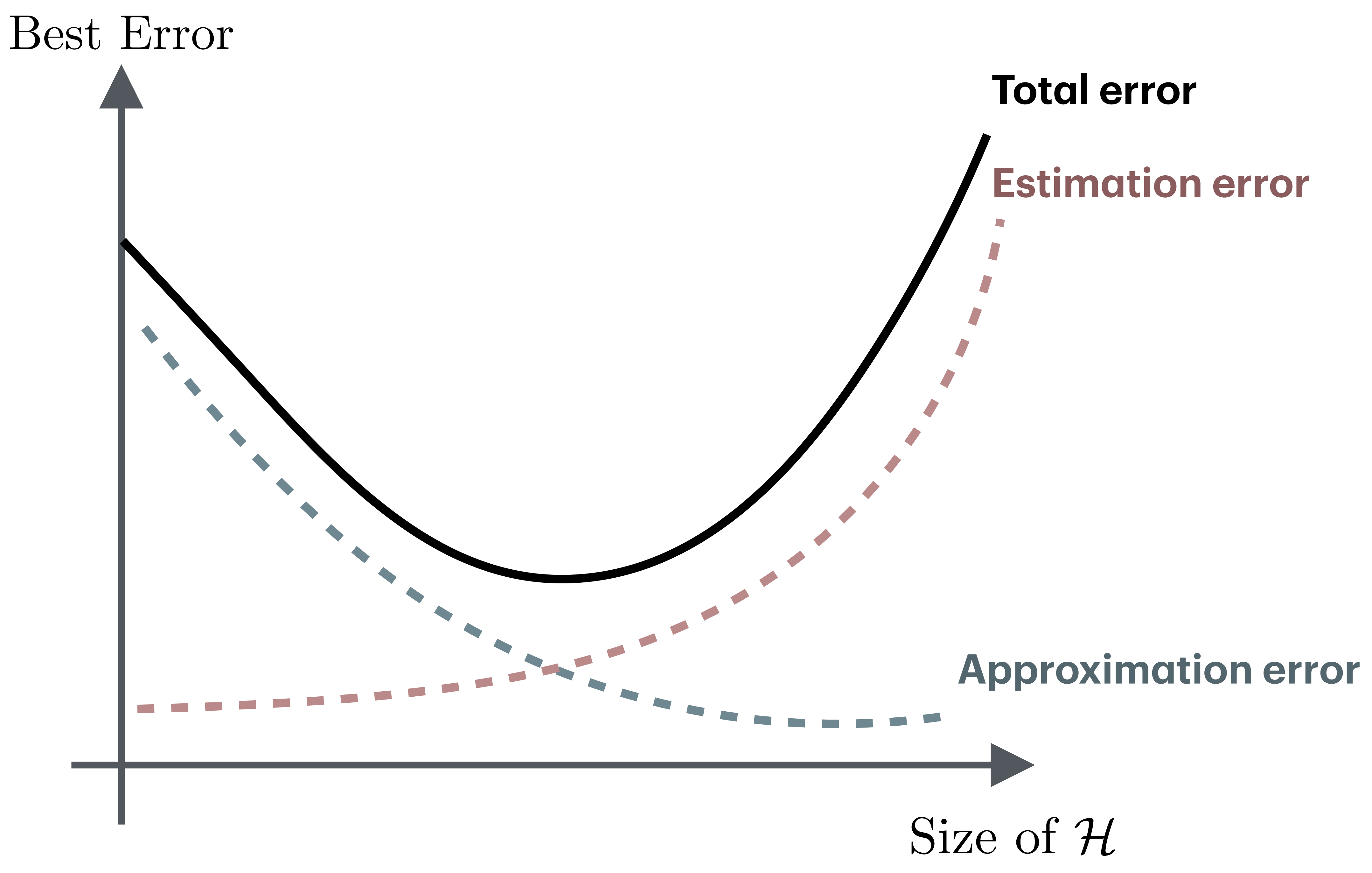
LESSON 1

INTRO TO STATISTICAL LEARNING

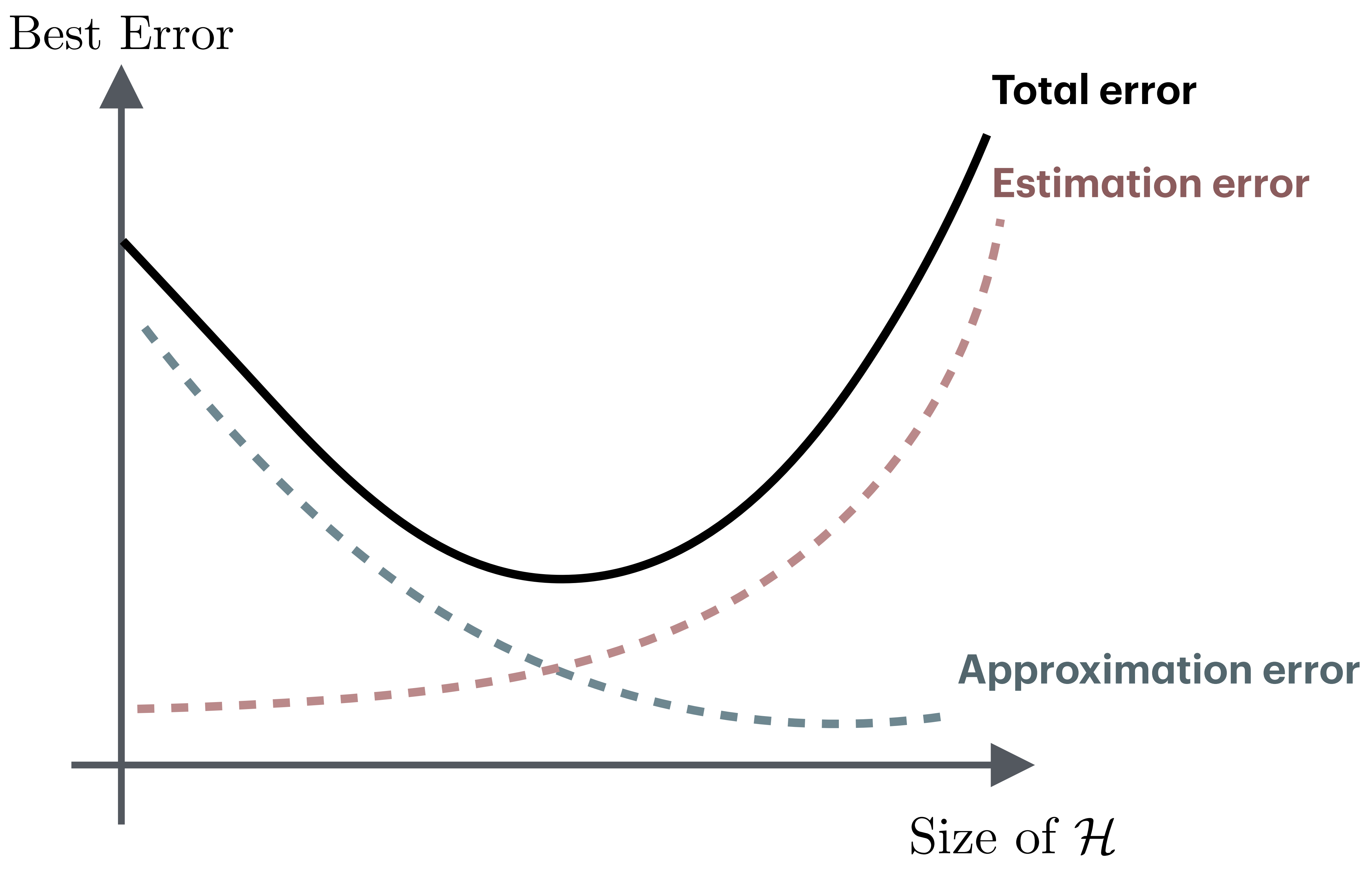


1. Supervised Learning Setting

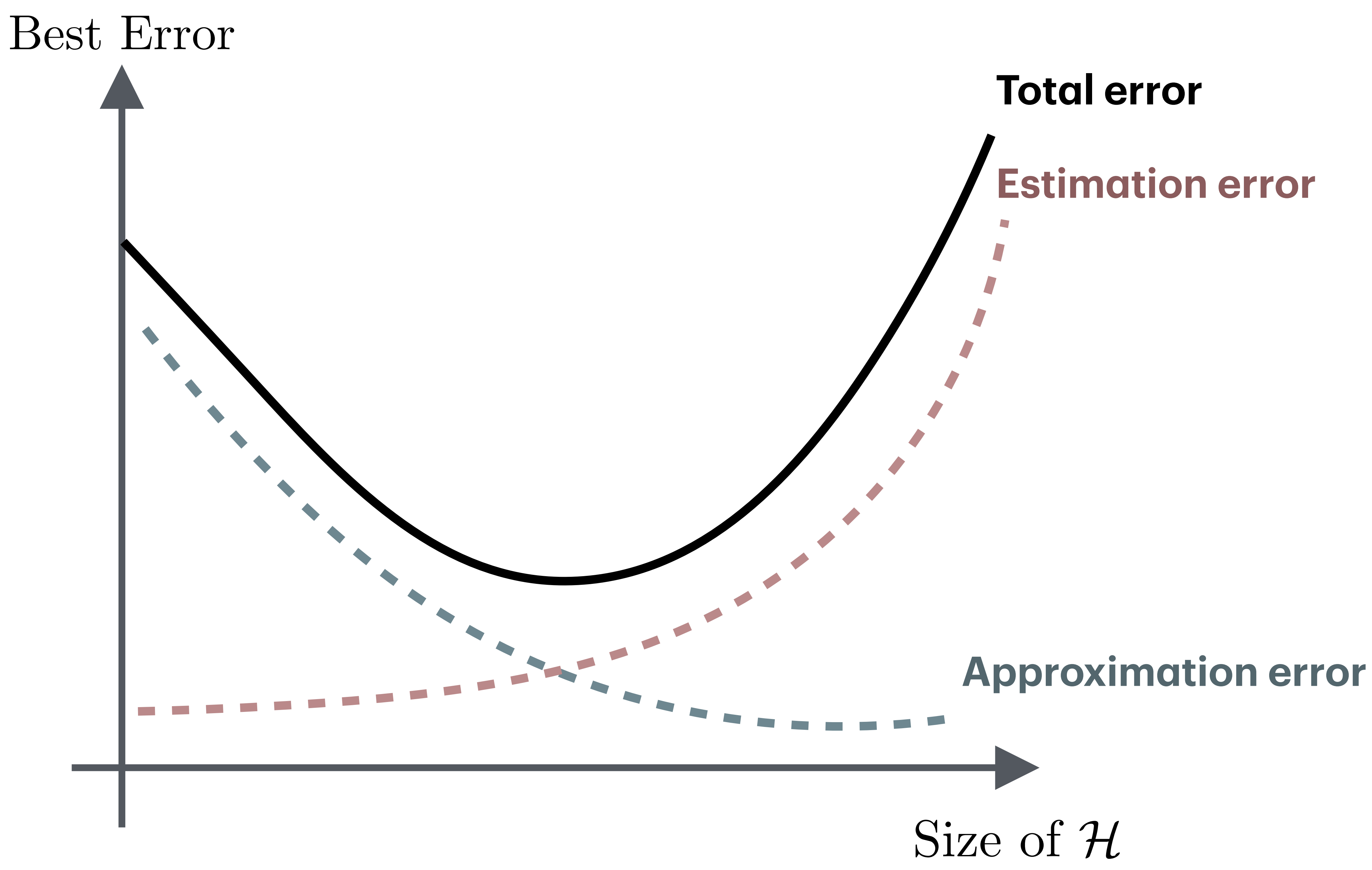
2. Estimation vs Approximation

3. Maximal inequalities

4. Rademacher Complexity



1. Supervised Learning Setting



The supervised learning setup

- General setup

The supervised learning setup

General setup

Supervised learning (**SL**) is a paradigm in machine learning where input objects and a desired output value train a model. The training data is processed, building a function that maps new data to expected output values.

The supervised learning setup

General setup

Supervised learning (**SL**) is a paradigm in machine learning where input objects and a desired output value train a model. The training data is processed, building a function that maps new data to expected output values.

Definitions

Formally, a supervised learning is a problem of the form

$$\min_{f \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\mathcal{L}(y, f(x))\right]$$

where

The supervised learning setup

General setup

Supervised learning (**SL**) is a paradigm in machine learning where input objects and a desired output value train a model. The training data is processed, building a function that maps new data to expected output values.

Definitions

Formally, a supervised learning is a problem of the form

$$\min_{f \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\mathcal{L}(y, f(x))\right]$$

where $\mathcal{H}$ is a set of functions called the **hypothesis class**

The supervised learning setup

General setup

Supervised learning (**SL**) is a paradigm in machine learning where input objects and a desired output value train a model. The training data is processed, building a function that maps new data to expected output values.

Definitions

Formally, a supervised learning is a problem of the form

$$\min_{f \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\mathcal{L}(y, f(x))\right]$$

where $\mathcal{H}$ is a set of functions called the **hypothesis class**

$f \in \mathcal{H}$ is called an **estimator**

The supervised learning setup

General setup

Supervised learning (**SL**) is a paradigm in machine learning where input objects and a desired output value train a model. The training data is processed, building a function that maps new data to expected output values.

Definitions

Formally, a supervised learning is a problem of the form

$$\min_{f \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\mathcal{L}(y, f(x))\right]$$

where $\mathcal{H}$ is a set of functions called the **hypothesis class**

$f \in \mathcal{H}$ is called an **estimator**

$\mathcal{D}$ is a **data distribution** from which we can sample **input-output pairs** $x, y \in \mathcal{X} \times \mathcal{Y}$

The supervised learning setup

General setup

Supervised learning (**SL**) is a paradigm in machine learning where input objects and a desired output value train a model. The training data is processed, building a function that maps new data to expected output values.

Definitions

Formally, a supervised learning is a problem of the form

$$\min_{f \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\mathcal{L}(y, f(x))\right]$$

where $\mathcal{H}$ is a set of functions called the **hypothesis class**

$f \in \mathcal{H}$ is called an **estimator**

$\mathcal{D}$ is a **data distribution** from which we can sample **input-output pairs** $x, y \in \mathcal{X} \times \mathcal{Y}$

$\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is an **error function** aimed at penalizing the discrepency between the ground truth y and the estimate $f(x)$.

The supervised learning setup

General setup

Supervised learning (**SL**) is a paradigm in machine learning where input objects and a desired output value train a model. The training data is processed, building a function that maps new data to expected output values.

Definitions

Formally, a supervised learning is a problem of the form

$$\min_{f \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\mathcal{L}(y, f(x))\right]$$

where $\mathcal{H}$ is a set of functions called the **hypothesis class**

$f \in \mathcal{H}$ is called an **estimator**

$\mathcal{D}$ is a **data distribution** from which we can sample **input-output pairs** $x, y \in \mathcal{X} \times \mathcal{Y}$

$\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is an **error function** aimed at penalizing the discrepency between the ground truth y and the estimate $f(x)$.

$f : \mathcal{X} \to \mathcal{Y}$ is called an **estimator**.

The supervised learning setup

General setup

Supervised learning (**SL**) is a paradigm in machine learning where input objects and a desired output value train a model. The training data is processed, building a function that maps new data to expected output values.

Definitions

Formally, a supervised learning is a problem of the form

$$\min_{f \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\mathcal{L}(y, f(x))\right]$$

where $\mathcal{H}$ is a set of functions called the **hypothesis class**

$f \in \mathcal{H}$ is called an **estimator**

$\mathcal{D}$ is a **data distribution** from which we can sample **input-output pairs** $x, y \in \mathcal{X} \times \mathcal{Y}$

$\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is an **error function** aimed at penalizing the discrepency between the ground truth y and the estimate $f(x)$.

$f : \mathcal{X} \to \mathcal{Y}$ is called an **estimator**.

Remarks :

- When $\mathcal{Y}$ is finite, this above problem is called a **classification problem**.

The supervised learning setup

General setup

Supervised learning (**SL**) is a paradigm in machine learning where input objects and a desired output value train a model. The training data is processed, building a function that maps new data to expected output values.

Definitions

Formally, a supervised learning is a problem of the form

$$\min_{f \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\mathcal{L}(y, f(x))\right]$$

where $\mathcal{H}$ is a set of functions called the **hypothesis class**

$f \in \mathcal{H}$ is called an **estimator**

$\mathcal{D}$ is a **data distribution** from which we can sample **input-output pairs** $x, y \in \mathcal{X} \times \mathcal{Y}$

$\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is an **error function** aimed at penalizing the discrepency between the ground truth y and the estimate $f(x)$.

$f : \mathcal{X} \to \mathcal{Y}$ is called an **estimator**.

Remarks :

- When $\mathcal{Y}$ is finite, this above problem is called a **classification problem**.
- When $\mathcal{Y}$ is **not** finite, this above problem is called a **regression problem**.

The supervised learning setup

General setup

Supervised learning (**SL**) is a paradigm in machine learning where input objects and a desired output value train a model. The training data is processed, building a function that maps new data to expected output values.

Definitions

Formally, a supervised learning is a problem of the form

$$\min_{f \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\mathcal{L}(y, f(x))\right]$$

where $\mathcal{H}$ is a set of functions called the **hypothesis class**

$f \in \mathcal{H}$ is called an **estimator**

$\mathcal{D}$ is a **data distribution** from which we can sample **input-output pairs** $x, y \in \mathcal{X} \times \mathcal{Y}$

$\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is an **error function** aimed at penalizing the discrepency between the ground truth y and the estimate $f(x)$.

$f : \mathcal{X} \to \mathcal{Y}$ is called an **estimator**.

Remarks :

- When $\mathcal{Y}$ is finite, this above problem is called a **classification problem**.
- When $\mathcal{Y}$ is **not** finite, this above problem is called a **regression problem**.

Empirical Risk Minimization

In practice, one often cannot access the whole distribution $\mathcal{D}$, but only a finite sample of it, $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, often called the **training set**.

The supervised learning setup

General setup

Supervised learning (**SL**) is a paradigm in machine learning where input objects and a desired output value train a model. The training data is processed, building a function that maps new data to expected output values.

Definitions

Formally, a supervised learning is a problem of the form

$$\min_{f \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\mathcal{L}(y, f(x))\right]$$

where $\mathcal{H}$ is a set of functions called the **hypothesis class**

$f \in \mathcal{H}$ is called an **estimator**

$\mathcal{D}$ is a **data distribution** from which we can sample **input-output pairs** $x, y \in \mathcal{X} \times \mathcal{Y}$

$\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is an **error function** aimed at penalizing the discrepency between the ground truth y and the estimate $f(x)$.

$f : \mathcal{X} \to \mathcal{Y}$ is called an **estimator**.

Remarks :

- When $\mathcal{Y}$ is finite, this above problem is called a **classification problem**.
- When $\mathcal{Y}$ is **not** finite, this above problem is called a **regression problem**.

Empirical Risk Minimization

In practice, one often cannot access the whole distribution $\mathcal{D}$, but only a finite sample of it, $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, often called the **training set**.

Definition

Empirical minimization consists in replacing the general supervized learning problem by a minimization ove the training set. Formally, we solve

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(y_i, f(x_i)\right)$$

The supervised learning setup

General setup

Supervised learning (**SL**) is a paradigm in machine learning where input objects and a desired output value train a model. The training data is processed, building a function that maps new data to expected output values.

Definitions

Formally, a supervised learning is a problem of the form

$$\min_{f \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\mathcal{L}(y, f(x))\right]$$

where $\mathcal{H}$ is a set of functions called the **hypothesis class**

$f \in \mathcal{H}$ is called an **estimator**

$\mathcal{D}$ is a **data distribution** from which we can sample **input-output pairs** $x, y \in \mathcal{X} \times \mathcal{Y}$

$\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is an **error function** aimed at penalizing the discrepency between the ground truth y and the estimate $f(x)$.

$f : \mathcal{X} \to \mathcal{Y}$ is called an **estimator**.

Remarks :

- When $\mathcal{Y}$ is finite, this above problem is called a **classification problem**.
- When $\mathcal{Y}$ is **not** finite, this above problem is called a **regression problem**.

Empirical Risk Minimization

In practice, one often cannot access the whole distribution $\mathcal{D}$, but only a finite sample of it, $\{(x_1, y_1), \dots, (x_n, y_n)\}$, often called the **training set**.

Definition

Empirical minimization consists in replacing the general supervized learning problem by a minimization ove the training set. Formally, we solve

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(y_i, f(x_i)\right)$$

Remarks :

- The hypothesis class is typically defined as a parameterized family of functions :

$$\mathcal{H} = \left\{x \mapsto f(w, x), w \in \mathbb{R}^d\right\}$$

The supervised learning setup

General setup

Supervised learning (**SL**) is a paradigm in machine learning where input objects and a desired output value train a model. The training data is processed, building a function that maps new data to expected output values.

Definitions

Formally, a supervised learning is a problem of the form

$$\min_{f \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\mathcal{L}(y, f(x))\right]$$

where $\mathcal{H}$ is a set of functions called the **hypothesis class**

$f \in \mathcal{H}$ is called an **estimator**

$\mathcal{D}$ is a **data distribution** from which we can sample **input-output pairs** $x, y \in \mathcal{X} \times \mathcal{Y}$

$\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is an **error function** aimed at penalizing the discrepency between the ground truth y and the estimate $f(x)$.

$f : \mathcal{X} \to \mathcal{Y}$ is called an **estimator**.

Remarks :

- When $\mathcal{Y}$ is finite, this above problem is called a **classification problem**.
- When $\mathcal{Y}$ is **not** finite, this above problem is called a **regression problem**.

Empirical Risk Minimization

In practice, one often cannot access the whole distribution $\mathcal{D}$, but only a finite sample of it, $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, often called the **training set**.

Definition

Empirical minimization consists in replacing the general supervized learning problem by a minimization ove the training set. Formally, we solve

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(y_i, f(x_i)\right)$$

Remarks :

- The hypothesis class is typically defined as a parameterized family of functions :

  $\mathcal{H} = \left\{x \mapsto f(w, x), w \in \mathbb{R}^d\right\}$

- The above problem can thus be replace by an optimization problem in $\mathbb{R}^d$:

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(y_i, f(w, x_i)\right)$$

The supervised learning setup

General setup

Supervised learning (**SL**) is a paradigm in machine learning where input objects and a desired output value train a model. The training data is processed, building a function that maps new data to expected output values.

Definitions

Formally, a supervised learning is a problem of the form

$$\min_{f \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\mathcal{L}(y, f(x))\right]$$

where $\mathcal{H}$ is a set of functions called the **hypothesis class**

$f \in \mathcal{H}$ is called an **estimator**

$\mathcal{D}$ is a **data distribution** from which we can sample **input-output pairs** $x, y \in \mathcal{X} \times \mathcal{Y}$

$\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is an **error function** aimed at penalizing the discrepency between the ground truth y and the estimate $f(x)$.

$f : \mathcal{X} \to \mathcal{Y}$ is called an **estimator**.

Remarks :

- When $\mathcal{Y}$ is finite, this above problem is called a **classification problem**.
- When $\mathcal{Y}$ is **not** finite, this above problem is called a **regression problem**.

Empirical Risk Minimization

In practice, one often cannot access the whole distribution $\mathcal{D}$, but only a finite sample of it, $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, often called the **training set**.

Definition

Empirical minimization consists in replacing the general supervized learning problem by a minimization ove the training set. Formally, we solve

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(y_i, f(x_i)\right)$$

Remarks :

- The hypothesis class is typically defined as a parameterized family of functions :

  $\mathcal{H} = \left\{x \mapsto f(w, x), w \in \mathbb{R}^d\right\}$

- The above problem can thus be replace by an optimization problem in $\mathbb{R}^d$:

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(y_i, f(w, x_i)\right)$$

- Hence, we can use the gradient descent, with respect to the hyperparameter w,

$$w_{t+1} = w_t - \alpha \frac{\partial}{\partial w}\left[\frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(y_i, f(w, x_i))\right](w_t)$$

  to learn the optimal parameter $w^\star$, and thus the optimal predictor $x \mapsto f(w^\star, x)$.

2 classical examples

- Linear regression

2 classical examples

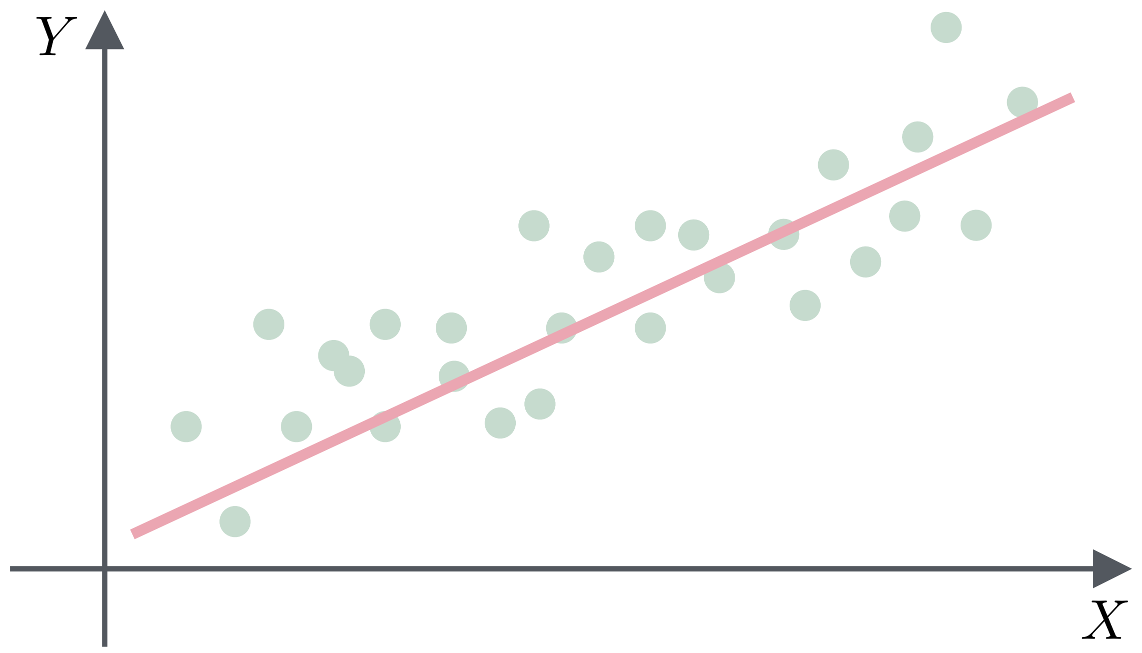
- Linear regression

  Linear regression seeks to estimate a linear relationship between a **continuous** scalar response Y, a one or more explanatory variables captured in a vector X.

2 classical examples

- Linear regression

Linear regression seeks to estimate a linear relationship between a **continuous** scalar response Y, a one or more explanatory variables captured in a vector X.



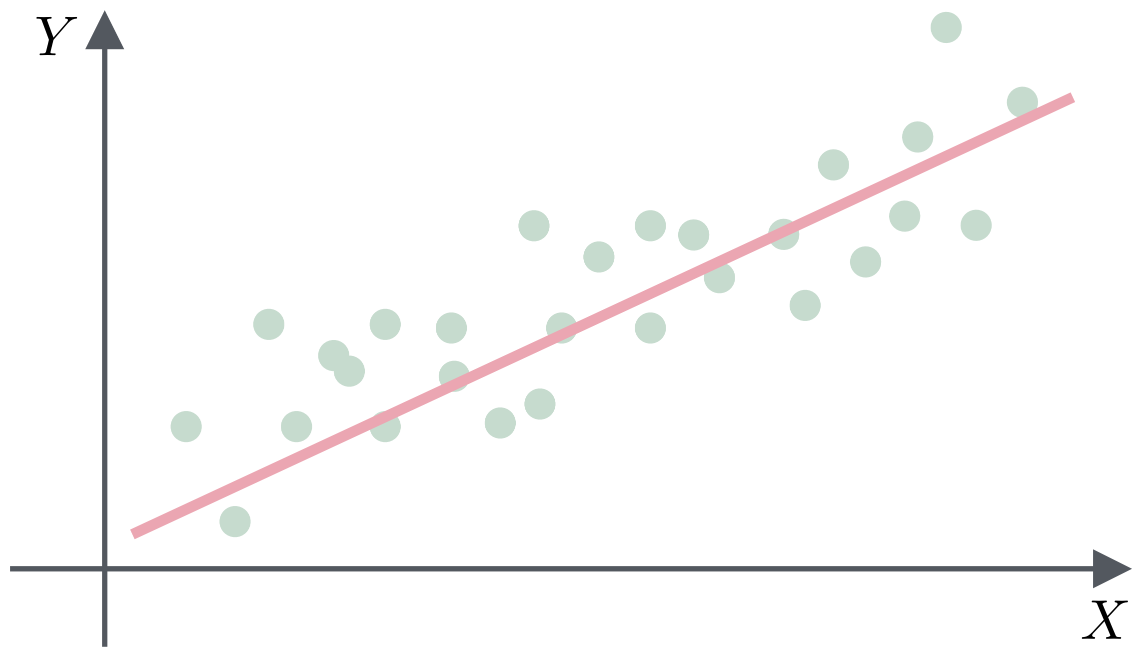
Ordinary Least Squares

$$\min_{w \in \mathbb{R}^2} \mathbb{E}\left[\|Y - Xw\|^2\right]$$

2 classical examples

- Linear regression

Linear regression seeks to estimate a linear relationship between a **continuous** scalar response Y, a one or more explanatory variables captured in a vector X.



Ordinary Least Squares

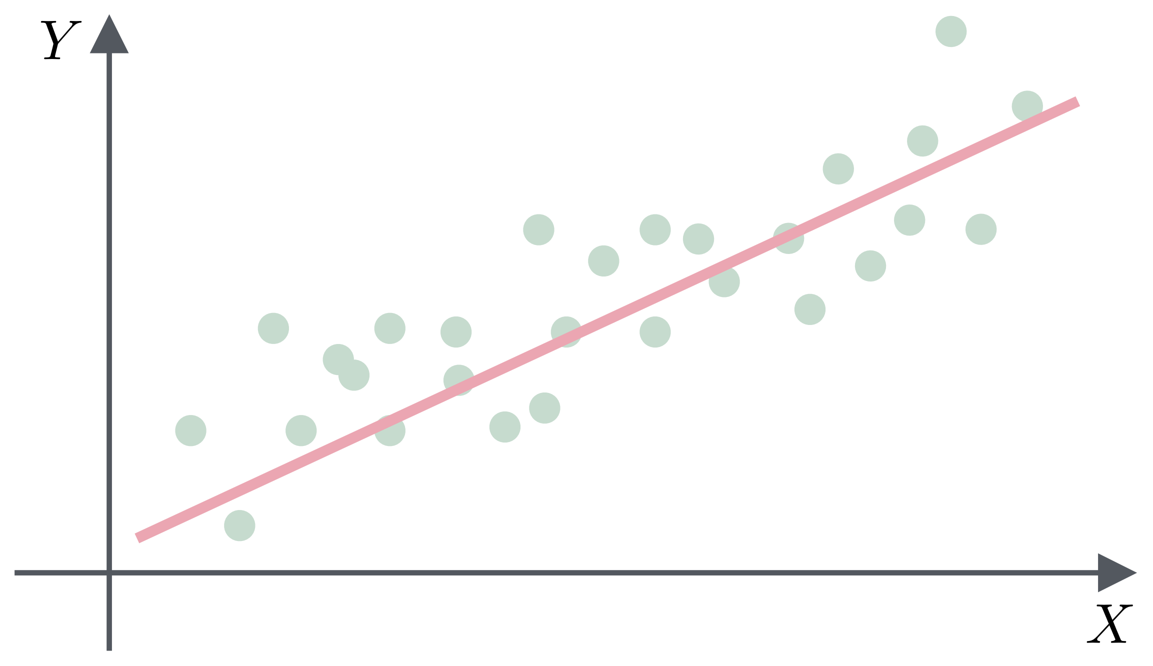
$$\min_{w \in \mathbb{R}^2} \mathbb{E}\left[\|Y - Xw\|^2\right]$$

$$\min_{f \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathcal{L}(y, f(x))\right]$$

2 classical examples

- Linear regression

Linear regression seeks to estimate a linear relationship between a **continuous** scalar response Y, a one or more explanatory variables captured in a vector X.



Ordinary Least Squares

$$\min_{w \in \mathbb{R}^2} \mathbb{E}\left[\|Y - Xw\|^2\right]$$

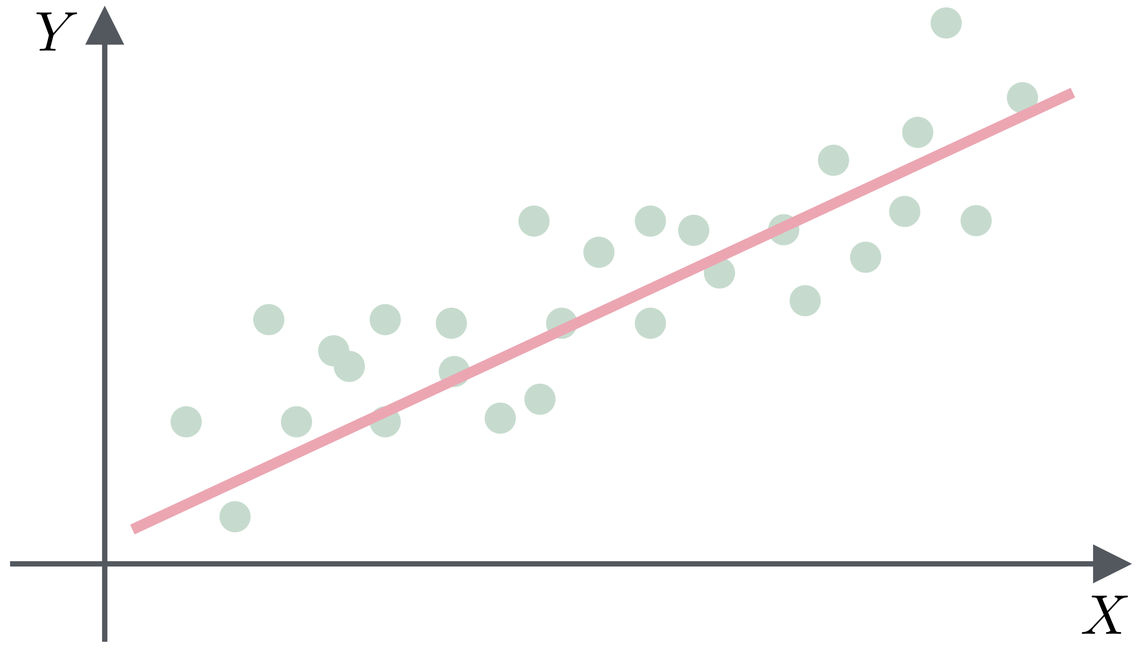
$$\min_{f \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathcal{L}(y, f(x))\right]$$

$\mathcal{H} = \mathbb{R}^2$

2 classical examples

- Linear regression

Linear regression seeks to estimate a linear relationship between a **continuous** scalar response Y, a one or more explanatory variables captured in a vector X.



Ordinary Least Squares

$$\min_{w \in \mathbb{R}^2} \mathbb{E}\left[\|Y - Xw\|^2\right]$$

$$\min_{f \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathcal{L}(y, f(x))\right]$$

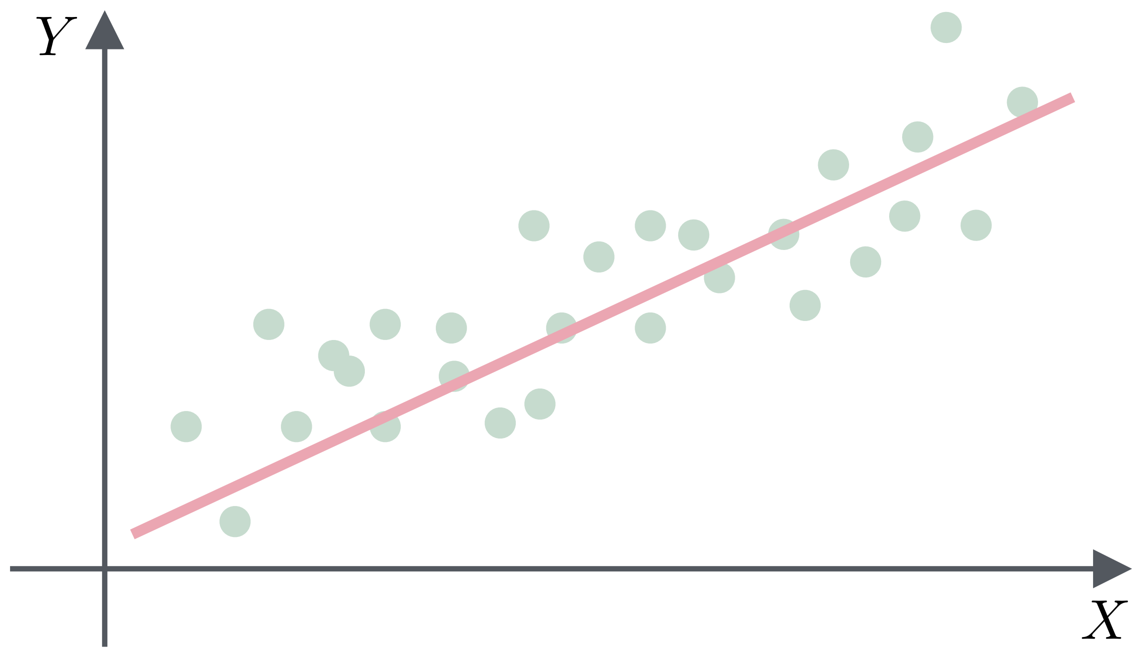
$\mathcal{L}(y, \hat{y}) = (y - \hat{y})^2$

$\mathcal{H} = \mathbb{R}^2$

2 classical examples

- Linear regression

Linear regression seeks to estimate a linear relationship between a **continuous** scalar response Y, a one or more explanatory variables captured in a vector X.



Ordinary Least Squares

$$\min_{w \in \mathbb{R}^2} \mathbb{E}\left[\|Y - Xw\|^2\right]$$

$$\min_{f \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathcal{L}(y, f(x))\right]$$

$\mathcal{L}(y, \hat{y}) = (y - \hat{y})^2$

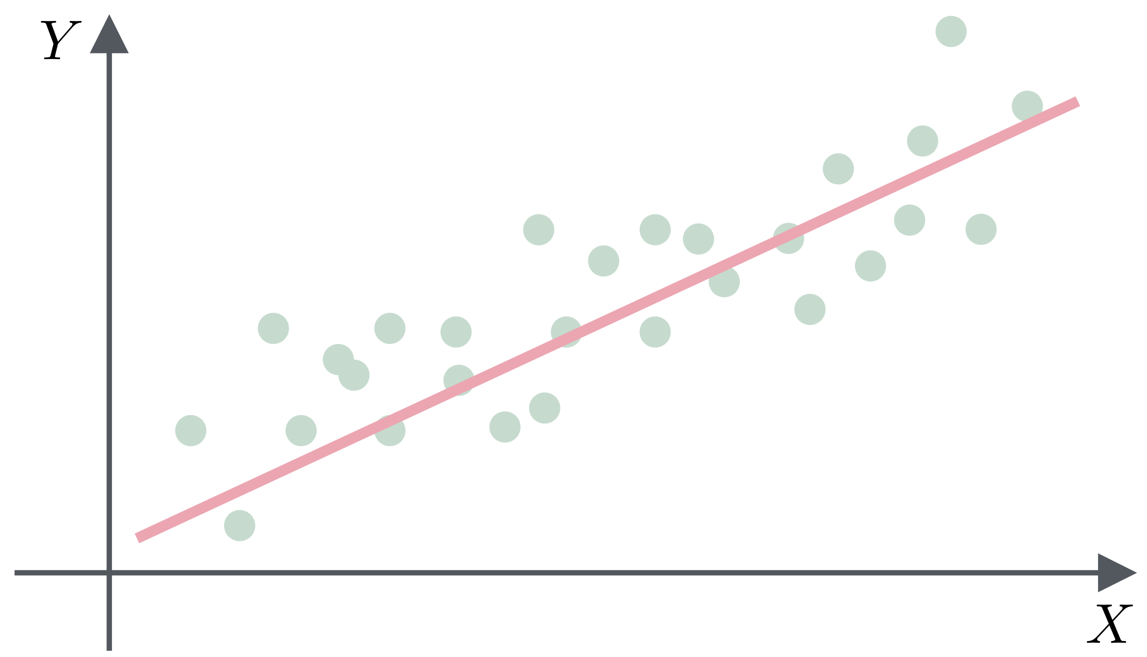
$f(x) = w^\top x$

$\mathcal{H} = \mathbb{R}^2$

2 classical examples

- Linear regression

  Linear regression seeks to estimate a linear relationship between a **continuous** scalar response Y, a one or more explanatory variables captured in a vector X.



Ordinary Least Squares

$$\min_{w \in \mathbb{R}^2} \mathbb{E}\left[\|Y - Xw\|^2\right]$$

$$\min_{f \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathcal{L}(y, f(x))\right]$$

$\mathcal{L}(y, \hat{y}) = (y - \hat{y})^2$

$f(x) = w^\top x$

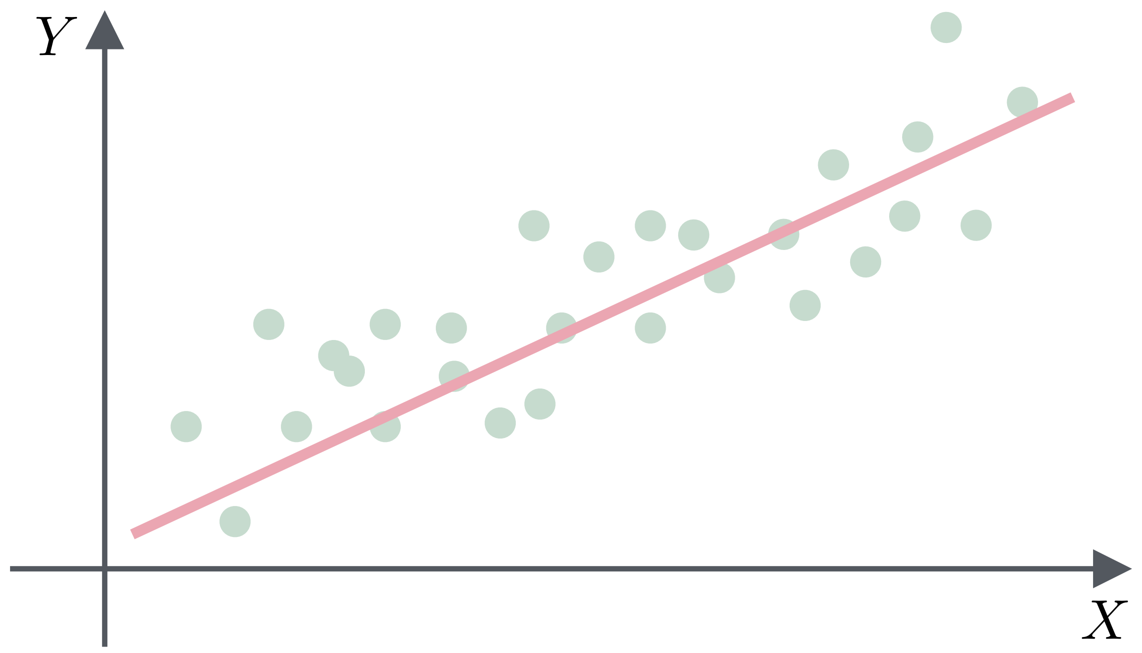
$\mathcal{H} = \mathbb{R}^2$

- Logistic regression

2 classical examples

Linear regression

Linear regression seeks to estimate a linear relationship between a **continuous** scalar response Y, a one or more explanatory variables captured in a vector X.



Ordinary Least Squares

$$\min_{w \in \mathbb{R}^2} \mathbb{E}\left[\|Y - Xw\|^2\right]$$

$$\min_{f \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathcal{L}(y, f(x))\right]$$

$$\mathcal{L}(y, \hat{y}) = (y - \hat{y})^2$$

$$f(x) = w^\top x$$

$$\mathcal{H} = \mathbb{R}^2$$

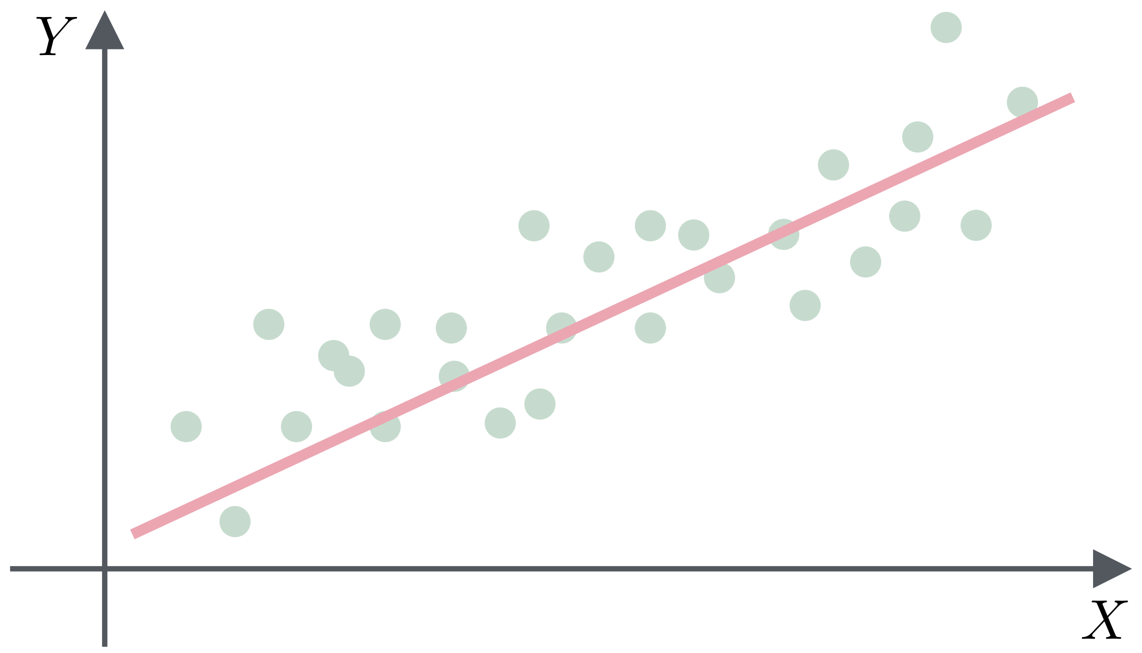
Logistic regression

Logistic regression seeks to a linear relationship between a discrete response Y taking values in some finite space $\mathcal{Y}$, and one or more explanatory variables captured in a vector X.

2 classical examples

Linear regression

Linear regression seeks to estimate a linear relationship between a **continuous** scalar response Y, a one or more explanatory variables captured in a vector X.



Ordinary Least Squares

$$\min_{w \in \mathbb{R}^2} \mathbb{E}\left[\|Y - Xw\|^2\right]$$

$$\min_{f \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathcal{L}(y, f(x))\right]$$

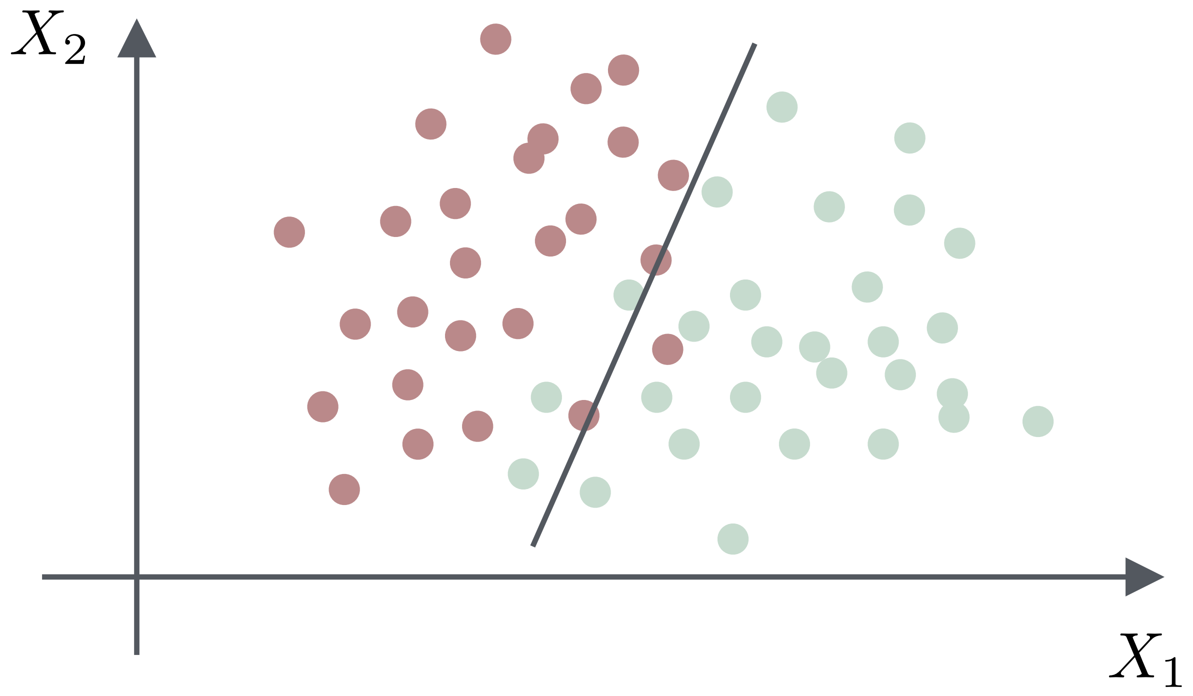
$$\mathcal{L}(y, \hat{y}) = (y - \hat{y})^2$$

$$f(x) = w^\top x$$

$$\mathcal{H} = \mathbb{R}^2$$

Logistic regression

Logistic regression seeks to a linear relationship between a discrete response Y taking values in some finite space $\mathcal{Y}$, and one or more explanatory variables captured in a vector X.



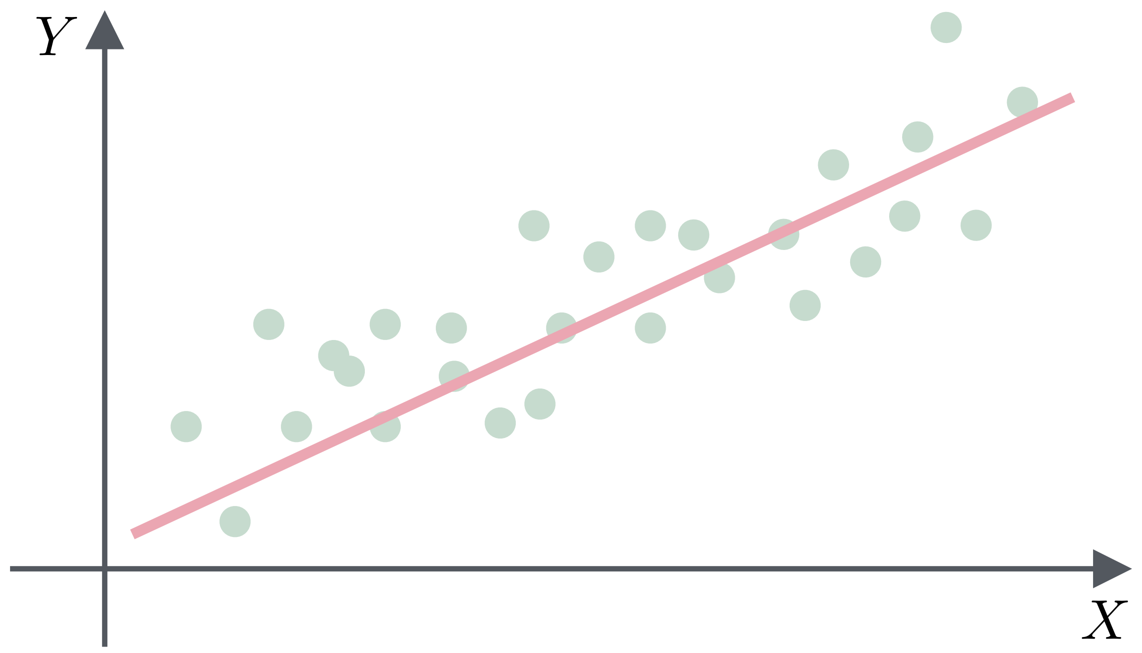
Logistic Regression

$$\min_{w \in \mathbb{R}^3} \mathbb{E}\left[\log(1 + \exp(-y\, w^T X))\right]$$

2 classical examples

- Linear regression

Linear regression seeks to estimate a linear relationship between a **continuous** scalar response Y, a one or more explanatory variables captured in a vector X.



Ordinary Least Squares

$$\min_{w \in \mathbb{R}^2} \mathbb{E}\left[\|Y - Xw\|^2\right]$$

$$\min_{f \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathcal{L}(y, f(x))\right]$$

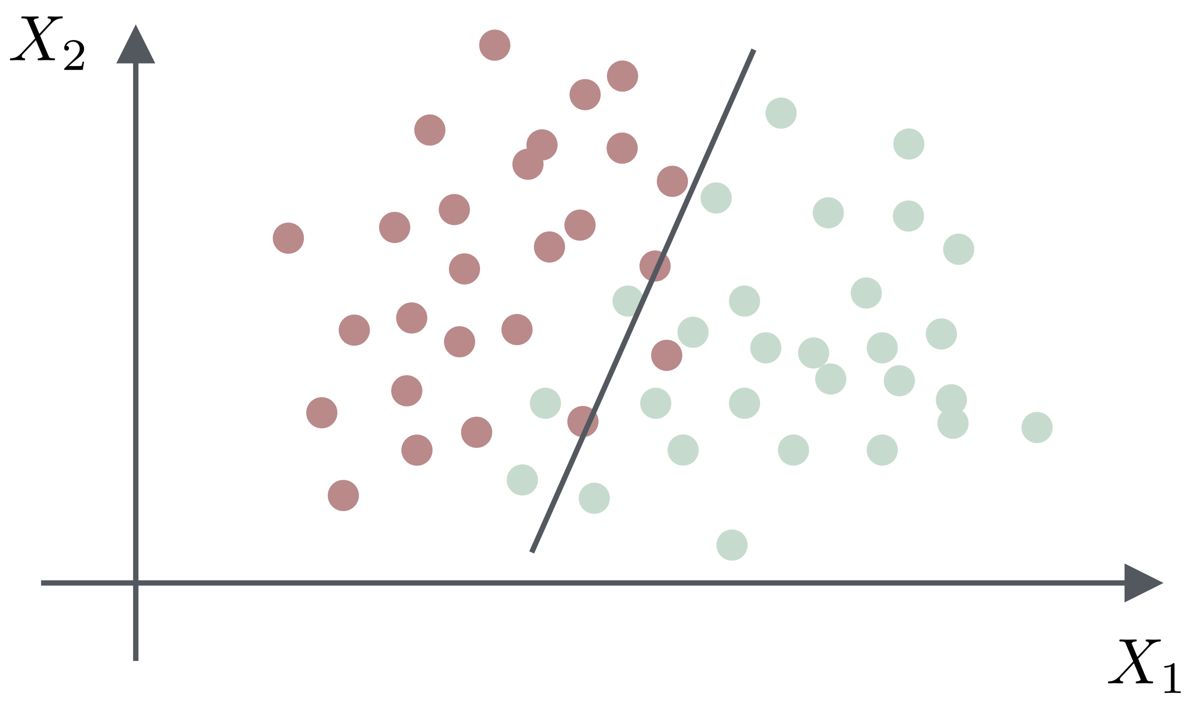
$\mathcal{L}(y, \hat{y}) = (y - \hat{y})^2$

$f(x) = w^\top x$

$\mathcal{H} = \mathbb{R}^2$

- Logistic regression

Logistic regression seeks to a linear relationship between a discrete response Y taking values in some finite space $\mathcal{Y}$, and one or more explanatory variables captured in a vector X.



Logistic Regression

$$\min_{w \in \mathbb{R}^3} \mathbb{E}\left[\log(1 + \exp(-y\, w^T X))\right]$$

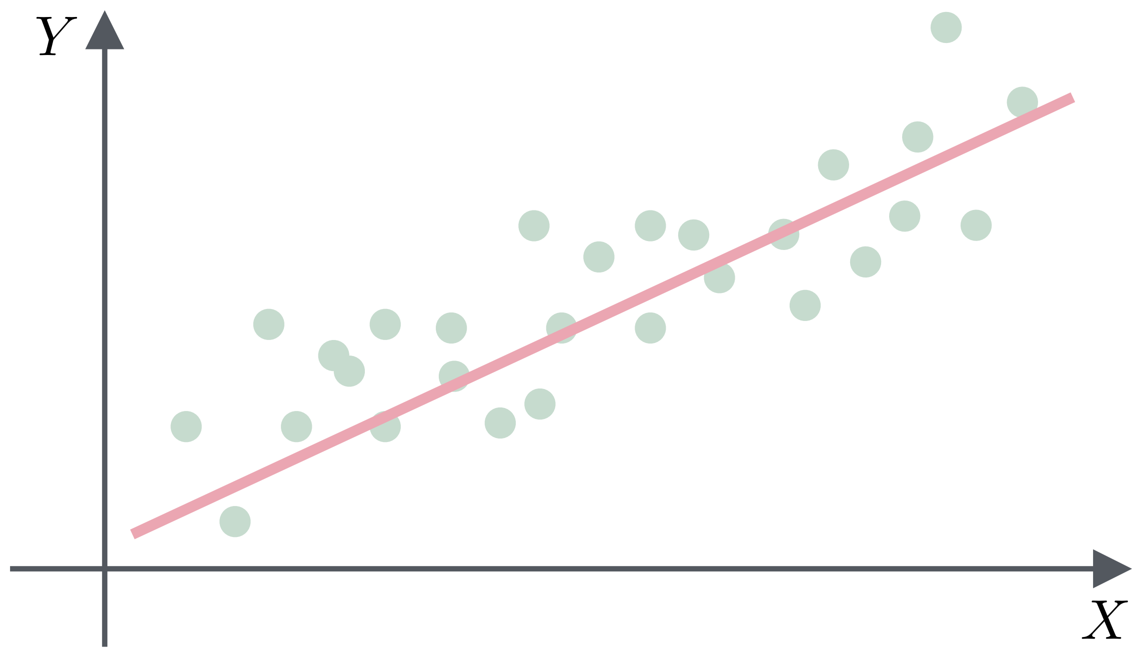
$$\min_{f \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathcal{L}(y, f(x))\right]$$

$\mathcal{H} = \mathbb{R}^3$

2 classical examples

Linear regression

Linear regression seeks to estimate a linear relationship between a **continuous** scalar response Y, a one or more explanatory variables captured in a vector X.



Ordinary Least Squares

$$\min_{w \in \mathbb{R}^2} \mathbb{E}\left[\|Y - Xw\|^2\right]$$

$$\mathcal{L}(y, \hat{y}) = (y - \hat{y})^2$$

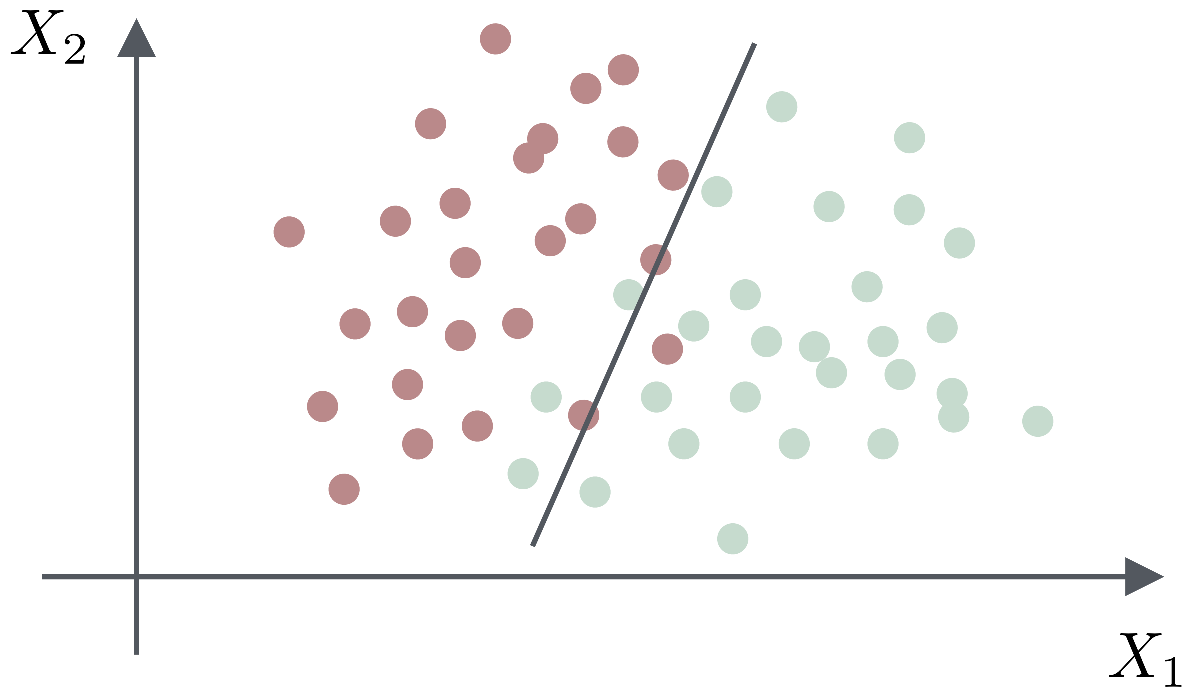
$$\min_{f \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathcal{L}(y, f(x))\right]$$

$$f(x) = w^\top x$$

$$\mathcal{H} = \mathbb{R}^2$$

Logistic regression

Logistic regression seeks to a linear relationship between a discrete response Y taking values in some finite space $\mathcal{Y}$, and one or more explanatory variables captured in a vector X.



Logistic Regression

$$\min_{w \in \mathbb{R}^3} \mathbb{E}\left[\log(1 + \exp(-y\, w^T X))\right]$$

$$\mathcal{L}(y, \hat{y}) = \log\left(1 + \exp\left(-y\, \hat{y}\right)\right)$$

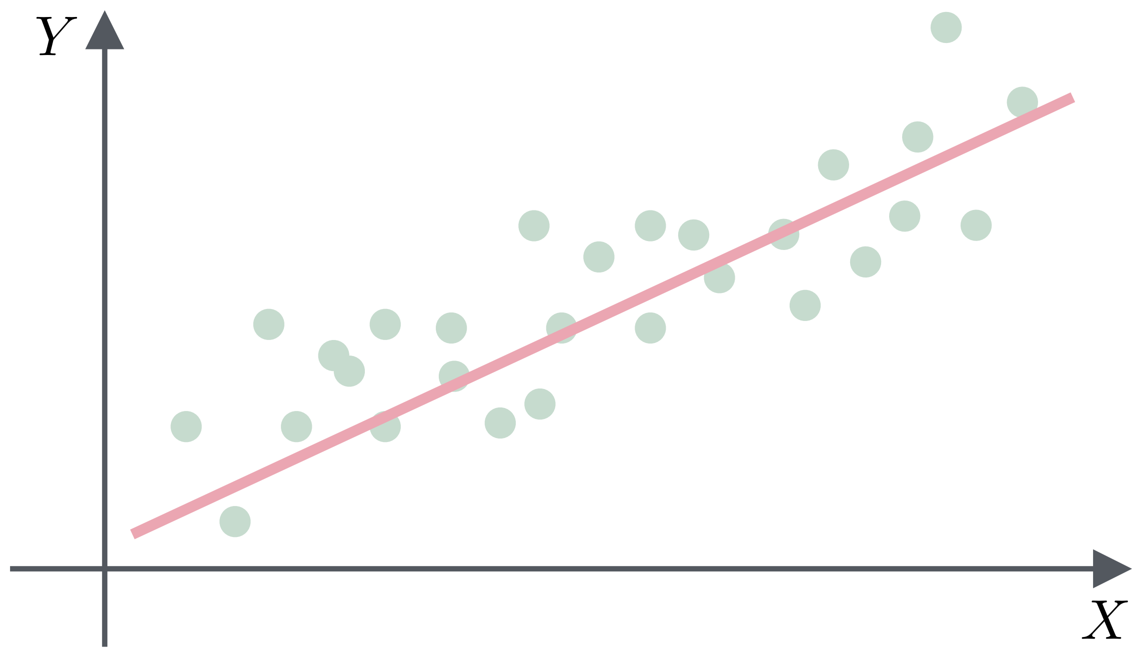
$$\min_{f \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathcal{L}(y, f(x))\right]$$

$$\mathcal{H} = \mathbb{R}^3$$

2 classical examples

Linear regression

Linear regression seeks to estimate a linear relationship between a **continuous** scalar response Y, a one or more explanatory variables captured in a vector X.



Ordinary Least Squares

$$\min_{w \in \mathbb{R}^2} \mathbb{E}\left[\|Y - Xw\|^2\right]$$

$$\min_{f \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathcal{L}(y, f(x))\right]$$

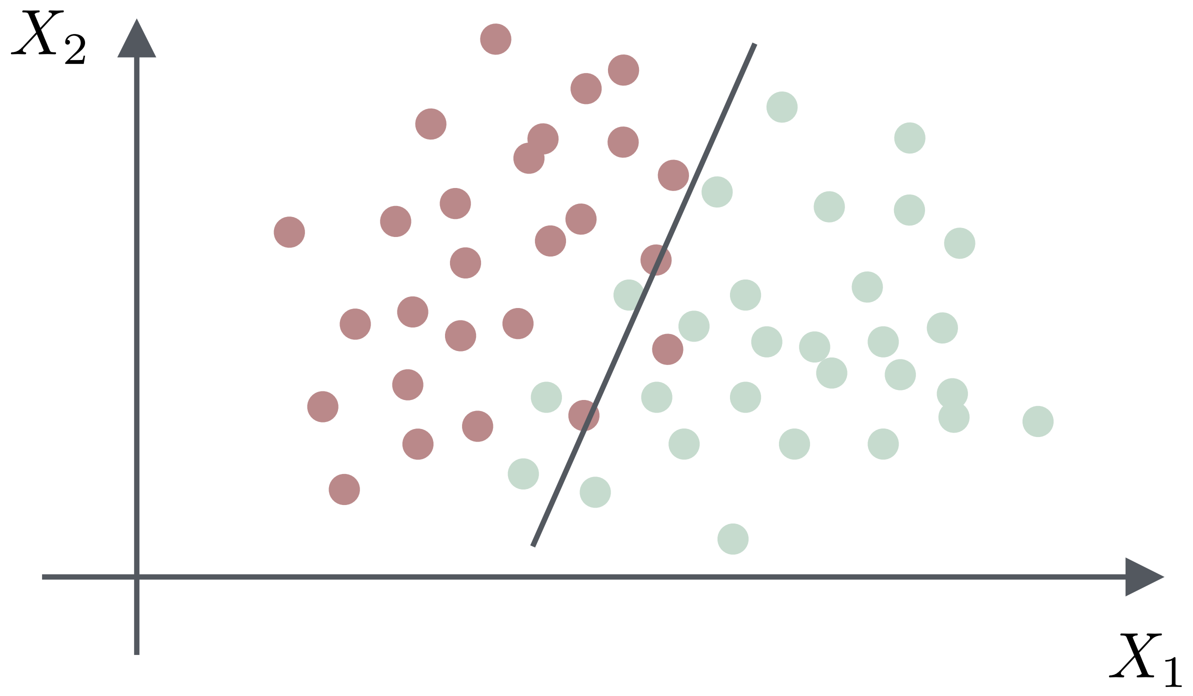
$\mathcal{L}(y, \hat{y}) = (y - \hat{y})^2$

$f(x) = w^\top x$

$\mathcal{H} = \mathbb{R}^2$

Logistic regression

Logistic regression seeks to a linear relationship between a discrete response Y taking values in some finite space $\mathcal{Y}$, and one or more explanatory variables captured in a vector X.



Logistic Regression

$$\min_{w \in \mathbb{R}^3} \mathbb{E}\left[\log(1 + \exp(-y\, w^T X))\right]$$

$$\min_{f \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathcal{L}(y, f(x))\right]$$

$\mathcal{L}(y, \hat{y}) = \log\left(1 + \exp\left(-y\, \hat{y}\right)\right)$

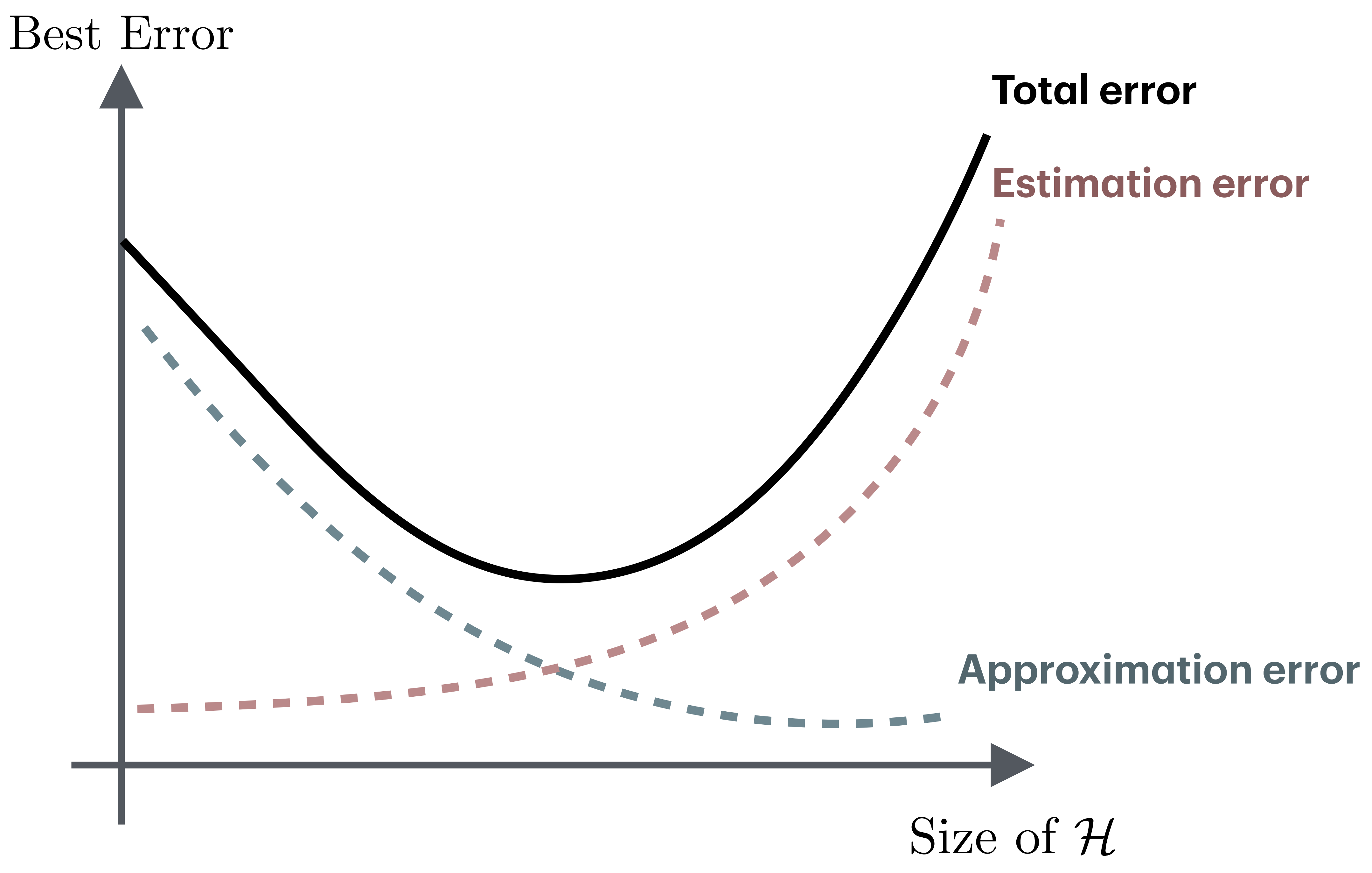
$f(x) = w^\top x$

$\mathcal{H} = \mathbb{R}^3$

Implementation

1. Supervised Learning Setting

2. Estimation vs Approximation



The supervised learning setup

- Excess Risk

The supervised learning setup

- Excess Risk

Definition

For any estimator $f \in \mathcal{H}$, we define the **excess risk** of f, denoted $r(f)$ as

$$r(f) = \mathbb{E}_{(x,y)\sim D}\left[\ell(f(x), y)\right].$$

The supervised learning setup

Excess Risk

Definition

For any estimator $f \in \mathcal{H}$, we define the **excess risk** of f, denoted $r(f)$ as

$$r(f) = \mathbb{E}_{(x,y)\sim D}\left[\ell(f(x), y)\right].$$

Remarks :

- Thus, supervised learning consists in solving $\inf_{f \in \mathcal{H}} r(f)$.

The supervised learning setup

Excess Risk

Definition

For any estimator $f \in \mathcal{H}$, we define the **excess risk** of f, denoted $r(f)$ as

$$r(f) = \mathbb{E}_{(x,y)\sim D}\left[\ell(f(x), y)\right].$$

Remarks :

- Thus, supervised learning consists in solving $\inf_{f\in\mathcal{H}} r(f)$.
- Let's introduce as well $f^\star$, and $f^{\star\star}$ as

$$f^\star \in \operatorname*{argmin}_{f\in\mathcal{H}} r(f) \qquad f^{\star\star} \in \operatorname*{argmin}_{f\in\mathcal{F}} r(f)$$

where $\mathcal{F} = \{f : \mathcal{X} \to \mathcal{Y}\}$.

The supervised learning setup

Excess Risk

Definition

For any estimator $f \in \mathcal{H}$, we define the **excess risk** of f, denoted $r(f)$ as

$$r(f) = \mathbb{E}_{(x,y)\sim D}\left[\ell(f(x), y)\right].$$

Remarks :

- Thus, supervised learning consists in solving $\inf_{f \in \mathcal{H}} r(f)$.
- Let's introduce as well $f^\star$, and $f^{\star\star}$ as

$$f^\star \in \operatorname*{argmin}_{f \in \mathcal{H}} r(f) \qquad f^{\star\star} \in \operatorname*{argmin}_{f \in \mathcal{F}} r(f)$$

where $\mathcal{F} = \{f : \mathcal{X} \to \mathcal{Y}\}$.

- Clearly, we have $r(f^{\star\star}) \leq r(f^\star)$.

The supervised learning setup

Excess Risk

Definition

For any estimator $f \in \mathcal{H}$, we define the **excess risk** of f, denoted $r(f)$ as

$$r(f) = \mathbb{E}_{(x,y)\sim D}\left[\ell(f(x), y)\right].$$

Remarks :

- Thus, supervised learning consists in solving $\inf_{f\in\mathcal{H}} r(f)$.
- Let's introduce as well $f^\star$, and $f^{\star\star}$ as

$$f^\star \in \operatorname*{argmin}_{f\in\mathcal{H}} r(f) \qquad f^{\star\star} \in \operatorname*{argmin}_{f\in\mathcal{F}} r(f)$$

  where $\mathcal{F} = \{f : \mathcal{X} \to \mathcal{Y}\}$.

- Clearly, we have $r(f^{\star\star}) \leq r(f^\star)$.

Proposition : Bayes decision rule

We have, for all $x \in \mathcal{X}$,

$$f^{\star\star}(x) = \operatorname*{argmin}_{\hat{y}\in\mathcal{Y}} \mathbb{E}\left[\ell(\hat{y}, Y)|X = x\right]$$

The supervised learning setup

Excess Risk

Definition

For any estimator $f \in \mathcal{H}$, we define the **excess risk** of f, denoted $r(f)$ as

$$r(f) = \mathbb{E}_{(x,y)\sim D}\left[\ell(f(x), y)\right].$$

Remarks :

- Thus, supervised learning consists in solving $\inf_{f \in \mathcal{H}} r(f)$.
- Let's introduce as well $f^\star$, and $f^{\star\star}$ as

$$f^\star \in \underset{f \in \mathcal{H}}{\operatorname{argmin}}\, r(f) \qquad f^{\star\star} \in \underset{f \in \mathcal{F}}{\operatorname{argmin}}\, r(f)$$

  where $\mathcal{F} = \{f : \mathcal{X} \to \mathcal{Y}\}$.

- Clearly, we have $r(f^{\star\star}) \leq r(f^\star)$.

Proposition : Bayes decision rule

We have, for all $x \in \mathcal{X}$,

$$f^{\star\star}(x) = \underset{\hat{y} \in \mathcal{Y}}{\operatorname{argmin}}\, \mathbb{E}\left[\ell(\hat{y}, Y) | X = x\right]$$

Proof : on the black board.

The supervised learning setup

Excess Risk

Definition

For any estimator $f \in \mathcal{H}$, we define the **excess risk** of f, denoted $r(f)$ as

$$r(f) = \mathbb{E}_{(x,y)\sim D}\left[\ell(f(x), y)\right].$$

Remarks :

- Thus, supervised learning consists in solving $\inf_{f\in\mathcal{H}} r(f)$.
- Let's introduce as well $f^\star$, and $f^{\star\star}$ as

$$f^\star \in \operatorname*{argmin}_{f\in\mathcal{H}} r(f) \qquad f^{\star\star} \in \operatorname*{argmin}_{f\in\mathcal{F}} r(f)$$

  where $\mathcal{F} = \{f : \mathcal{X} \to \mathcal{Y}\}$.

- Clearly, we have $r(f^{\star\star}) \leq r(f^\star)$.

Proposition : Bayes decision rule

We have, for all $x \in \mathcal{X}$,

$$f^{\star\star}(x) = \operatorname*{argmin}_{\hat{y}\in\mathcal{Y}} \mathbb{E}\left[\ell(\hat{y}, Y)|X = x\right]$$

Proof : on the black board.

Remarks :

- $f^{\star\star}$ is often called **bayes estimator**.

The supervised learning setup

Excess Risk

Definition

For any estimator $f \in \mathcal{H}$, we define the **excess risk** of f, denoted $r(f)$ as

$$r(f) = \mathbb{E}_{(x,y)\sim D}\left[\ell(f(x), y)\right].$$

Remarks :

- Thus, supervised learning consists in solving $\inf_{f\in\mathcal{H}} r(f)$.
- Let's introduce as well $f^\star$, and $f^{\star\star}$ as

$$f^\star \in \operatorname*{argmin}_{f\in\mathcal{H}} r(f) \qquad f^{\star\star} \in \operatorname*{argmin}_{f\in\mathcal{F}} r(f)$$

  where $\mathcal{F} = \{f : \mathcal{X} \to \mathcal{Y}\}$.

- Clearly, we have $r(f^{\star\star}) \leq r(f^\star)$.

Proposition : Bayes decision rule

We have, for all $x \in \mathcal{X}$,

$$f^{\star\star}(x) = \operatorname*{argmin}_{\hat{y}\in\mathcal{Y}} \mathbb{E}\left[\ell(\hat{y}, Y)|X = x\right]$$

Proof : on the black board.

Remarks :

- $f^{\star\star}$ is often called **bayes estimator**.
- Computing $f^{\star\star}$ requires the knowledge of the conditional distribution $\mathbb{P}_{Y|X}$. This is often impossible to obtain.

The supervised learning setup

Excess Risk

Definition

For any estimator $f \in \mathcal{H}$, we define the **excess risk** of f, denoted $r(f)$ as

$$r(f) = \mathbb{E}_{(x,y)\sim D}\left[\ell(f(x), y)\right].$$

Remarks :

- Thus, supervised learning consists in solving $\inf_{f \in \mathcal{H}} r(f)$.
- Let's introduce as well $f^\star$, and $f^{\star\star}$ as

$$f^\star \in \operatorname*{argmin}_{f \in \mathcal{H}} r(f) \qquad f^{\star\star} \in \operatorname*{argmin}_{f \in \mathcal{F}} r(f)$$

  where $\mathcal{F} = \{f : \mathcal{X} \to \mathcal{Y}\}$.

- Clearly, we have $r(f^{\star\star}) \leq r(f^\star)$.

Proposition : Bayes decision rule

We have, for all $x \in \mathcal{X}$,

$$f^{\star\star}(x) = \operatorname*{argmin}_{\hat{y} \in \mathcal{Y}} \mathbb{E}\left[\ell(\hat{y}, Y) | X = x\right]$$

Proof : on the black board.

Remarks :

- $f^{\star\star}$ is often called **bayes estimator**.
- Computing $f^{\star\star}$ requires the knowledge of the conditional distribution $\mathbb{P}_{Y|X}$. This is often impossible to obtain.
- In general, we suffer from an **estimation/approximation** tradeoff :

$$r(f) - \inf_{g \in \mathcal{F}} r(g) = r(f) - \inf_{h \in \mathcal{H}} r(h) + \inf_{h \in \mathcal{H}} r(h) - \inf_{g \in \mathcal{F}} r(g)$$

The supervised learning setup

Excess Risk

Definition

For any estimator $f \in \mathcal{H}$, we define the **excess risk** of f, denoted $r(f)$ as

$$r(f) = \mathbb{E}_{(x,y)\sim D}\left[\ell(f(x), y)\right].$$

Remarks :

- Thus, supervised learning consists in solving $\inf_{f \in \mathcal{H}} r(f)$.
- Let's introduce as well $f^\star$, and $f^{\star\star}$ as

$$f^\star \in \underset{f \in \mathcal{H}}{\operatorname{argmin}}\, r(f) \qquad f^{\star\star} \in \underset{f \in \mathcal{F}}{\operatorname{argmin}}\, r(f)$$

  where $\mathcal{F} = \{f : \mathcal{X} \to \mathcal{Y}\}$.

- Clearly, we have $r(f^{\star\star}) \leq r(f^\star)$.

Proposition : Bayes decision rule

We have, for all $x \in \mathcal{X}$,

$$f^{\star\star}(x) = \underset{\hat{y} \in \mathcal{Y}}{\operatorname{argmin}}\, \mathbb{E}\left[\ell(\hat{y}, Y) | X = x\right]$$

Proof : on the black board.

Remarks :

- $f^{\star\star}$ is often called **bayes estimator**.
- Computing $f^{\star\star}$ requires the knowledge of the conditional distribution $\mathbb{P}_{Y|X}$. This is often impossible to obtain.
- In general, we suffer from an **estimation/approximation** tradeoff :

$$r(f) - \inf_{g \in \mathcal{F}} r(g) = \underbrace{r(f) - \inf_{h \in \mathcal{H}} r(h)}_{\text{Estimation error}} + \inf_{h \in \mathcal{H}} r(h) - \inf_{g \in \mathcal{F}} r(g)$$

The supervised learning setup

Excess Risk

Definition

For any estimator $f \in \mathcal{H}$, we define the **excess risk** of f, denoted $r(f)$ as

$$r(f) = \mathbb{E}_{(x,y)\sim D}\left[\ell(f(x), y)\right].$$

Remarks :

- Thus, supervised learning consists in solving $\inf_{f\in\mathcal{H}} r(f)$.
- Let's introduce as well $f^\star$, and $f^{\star\star}$ as

$$f^\star \in \operatorname*{argmin}_{f\in\mathcal{H}} r(f) \qquad f^{\star\star} \in \operatorname*{argmin}_{f\in\mathcal{F}} r(f)$$

  where $\mathcal{F} = \{f : \mathcal{X} \to \mathcal{Y}\}$.

- Clearly, we have $r(f^{\star\star}) \leq r(f^\star)$.

Proposition : Bayes decision rule

We have, for all $x \in \mathcal{X}$,

$$f^{\star\star}(x) = \operatorname*{argmin}_{\hat{y}\in\mathcal{Y}} \mathbb{E}\left[\ell(\hat{y}, Y)|X = x\right]$$

Proof : on the black board.

Remarks :

- $f^{\star\star}$ is often called **bayes estimator**.
- Computing $f^{\star\star}$ requires the knowledge of the conditional distribution $\mathbb{P}_{Y|X}$. This is often impossible to obtain.
- In general, we suffer from an **estimation/approximation** tradeoff :

$$r(f) - \inf_{g\in\mathcal{F}} r(g) = \underbrace{r(f) - \inf_{h\in\mathcal{H}} r(h)}_{\text{Estimation error}} + \underbrace{\inf_{h\in\mathcal{H}} r(h) - \inf_{g\in\mathcal{F}} r(g)}_{\text{Approximation error}}$$

The supervised learning setup

Excess Risk

Definition

For any estimator $f \in \mathcal{H}$, we define the **excess risk** of f, denoted $r(f)$ as

$$r(f) = \mathbb{E}_{(x,y)\sim D}\left[\ell(f(x), y)\right].$$

Remarks :

- Thus, supervised learning consists in solving $\inf_{f\in\mathcal{H}} r(f)$.
- Let's introduce as well $f^\star$, and $f^{\star\star}$ as

$$f^\star \in \operatorname*{argmin}_{f\in\mathcal{H}} r(f) \qquad f^{\star\star} \in \operatorname*{argmin}_{f\in\mathcal{F}} r(f)$$

  where $\mathcal{F} = \{f : \mathcal{X} \to \mathcal{Y}\}$.

- Clearly, we have $r(f^{\star\star}) \leq r(f^\star)$.

Proposition : Bayes decision rule

We have, for all $x \in \mathcal{X}$,

$$f^{\star\star}(x) = \operatorname*{argmin}_{\hat{y}\in\mathcal{Y}} \mathbb{E}\left[\ell(\hat{y}, Y)|X = x\right]$$

Proof : on the black board.

Remarks :

- $f^{\star\star}$ is often called **bayes estimator**.
- Computing $f^{\star\star}$ requires the knowledge of the conditional distribution $\mathbb{P}_{Y|X}$. This is often impossible to obtain.
- In general, we suffer from an **estimation/approximation** tradeoff :

$$r(f) - \inf_{g\in\mathcal{F}} r(g) = \underbrace{r(f) - \inf_{h\in\mathcal{H}} r(h)}_{\text{Estimation error}} + \underbrace{\inf_{h\in\mathcal{H}} r(h) - \inf_{g\in\mathcal{F}} r(g)}_{\text{Approximation error}}$$

- And we need to work within the ERM framework, which leads us to work with the **empirical excess risk** $R(f)$, defined as

$$R(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i)$$

The supervised learning setup

Excess Risk

Definition

For any estimator $f \in \mathcal{H}$, we define the **excess risk** of f, denoted $r(f)$ as

$$r(f) = \mathbb{E}_{(x,y)\sim D}\left[\ell(f(x), y)\right].$$

Remarks :

- Thus, supervised learning consists in solving $\inf_{f\in\mathcal{H}} r(f)$.
- Let's introduce as well $f^\star$, and $f^{\star\star}$ as

$$f^\star \in \operatorname*{argmin}_{f\in\mathcal{H}} r(f) \qquad f^{\star\star} \in \operatorname*{argmin}_{f\in\mathcal{F}} r(f)$$

  where $\mathcal{F} = \{f : \mathcal{X} \to \mathcal{Y}\}$.

- Clearly, we have $r(f^{\star\star}) \leq r(f^\star)$.

Proposition : Bayes decision rule

We have, for all $x \in \mathcal{X}$,

$$f^{\star\star}(x) = \operatorname*{argmin}_{\hat{y}\in\mathcal{Y}} \mathbb{E}\left[\ell(\hat{y}, Y)|X = x\right]$$

Proof : on the black board.

Remarks :

- $f^{\star\star}$ is often called **bayes estimator**.
- Computing $f^{\star\star}$ requires the knowledge of the conditional distribution $\mathbb{P}_{Y|X}$. This is often impossible to obtain.
- In general, we suffer from an **estimation/approximation** tradeoff :

$$r(f) - \inf_{g\in\mathcal{F}} r(g) = \underbrace{r(f) - \inf_{h\in\mathcal{H}} r(h)}_{\text{Estimation error}} + \underbrace{\inf_{h\in\mathcal{H}} r(h) - \inf_{g\in\mathcal{F}} r(g)}_{\text{Approximation error}}$$

- And we need to work within the ERM framework, which leads us to work with the **empirical excess risk** $R(f)$, defined as

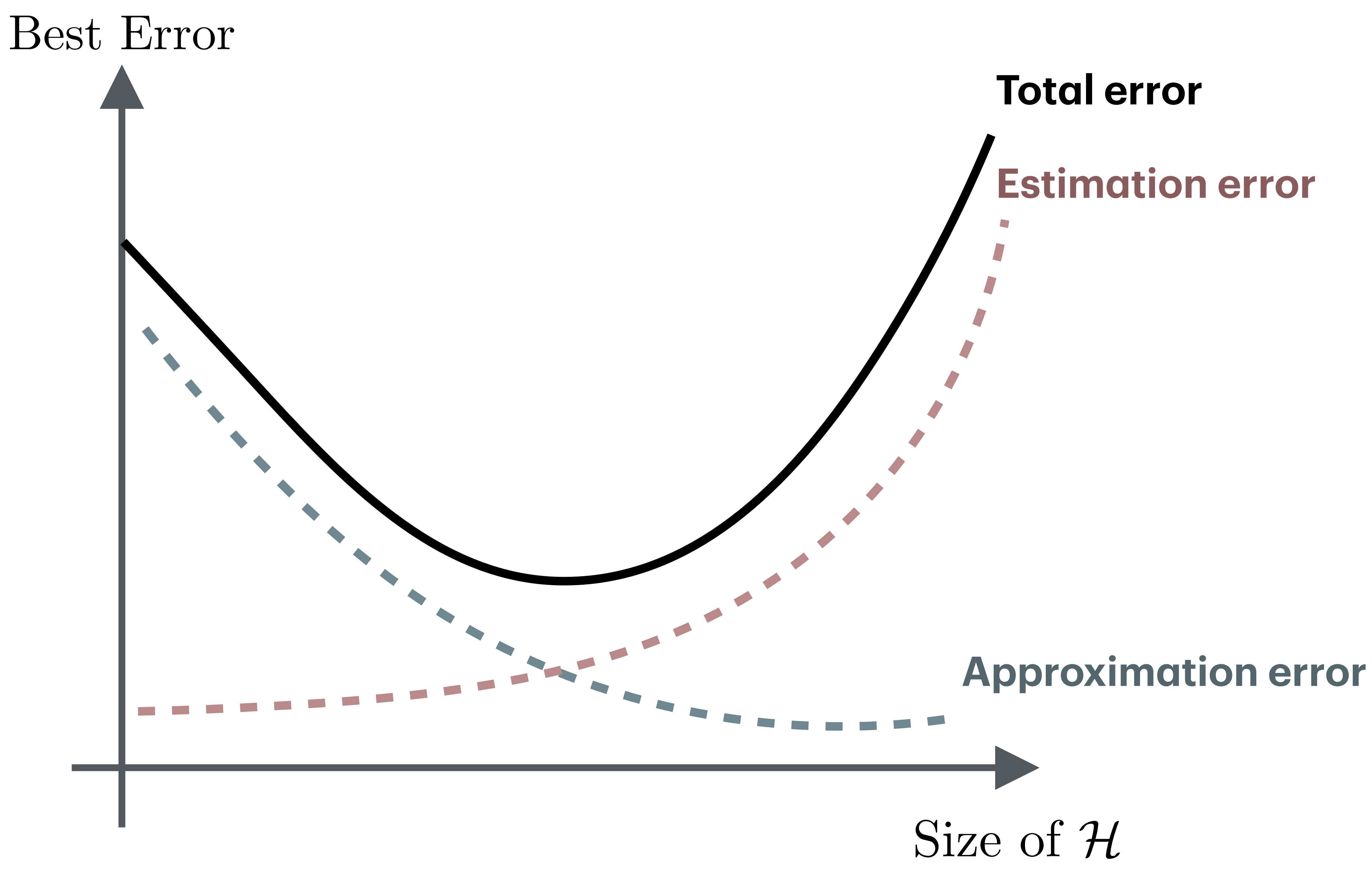
$$R(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i)$$

  Note that here, the (X_i, Y_i) are **random** samples

1. Supervised Learning Setting

2. Estimation vs Approximation

3. Maximal inequalities



Excess risk decompositon and maximal inequalities

- Excess Risk Decomposition

Proposition : Excess Risk Decompositon

For any $f \in \mathcal{H},$ we have

$$r(f) - r(f^\star) \leq R(f) - R(\tilde{f}^\star) + \sup_{f \in \mathcal{H}} \{R(f) - r(f)\} + \sup_{f \in \mathcal{H}} \{r(f) - R(f)\}$$

where $\tilde{f}^\star = \operatorname*{argmin}_{f \in \mathcal{H}} R(f) = \operatorname*{argmin}_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{N} \ell(f(x_i), y_i).$

Excess risk decompositon and maximal inequalities

Excess Risk Decomposition

Proposition : Excess Risk Decompositon

For any $f \in \mathcal{H}$, we have

$$r(f) - r(f^\star) \leq R(f) - R(\tilde{f}^\star) + \sup_{f \in \mathcal{H}} \{R(f) - r(f)\} + \sup_{f \in \mathcal{H}} \{r(f) - R(f)\}$$

where $\tilde{f}^\star = \operatorname*{argmin}_{f \in \mathcal{H}} R(f) = \operatorname*{argmin}_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{N} \ell(f(x_i), y_i)$.

Remarks :

- $R(f) - R(\tilde{f}^\star)$ may be reduced through optimization methods like gradient descent.

Excess risk decompositon and maximal inequalities

Excess Risk Decomposition

Proposition : Excess Risk Decompositon

For any $f \in \mathcal{H}$, we have

$$r(f) - r(f^\star) \leq R(f) - R(\tilde{f}^\star) + \sup_{f \in \mathcal{H}} \{R(f) - r(f)\} + \sup_{f \in \mathcal{H}} \{r(f) - R(f)\}$$

where $\tilde{f}^\star = \underset{f \in \mathcal{H}}{\operatorname{argmin}}\, R(f) = \underset{f \in \mathcal{H}}{\operatorname{argmin}}\, \frac{1}{n} \sum_{i=1}^{N} \ell(f(x_i), y_i)$.

Remarks :

- $R(f) - R(\tilde{f}^\star)$ may be reduced through optimization methods like gradient descent.
- In this chapter, we will focus on the next terms, $\sup_{f \in \mathcal{H}} \{R(f) - r(f)\}$, and $\sup_{f \in \mathcal{H}} \{r(f) - R(f)\}$.

Excess risk decompositon and maximal inequalities

Excess Risk Decomposition

Proposition : Excess Risk Decompositon

For any $f \in \mathcal{H}$, we have

$$r(f) - r(f^\star) \leq R(f) - R(\tilde{f}^\star) + \sup_{f \in \mathcal{H}} \{R(f) - r(f)\} + \sup_{f \in \mathcal{H}} \{r(f) - R(f)\}$$

where $\tilde{f}^\star = \operatorname*{argmin}_{f \in \mathcal{H}} R(f) = \operatorname*{argmin}_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{N} \ell(f(x_i), y_i)$.

Remarks :

- $R(f) - R(\tilde{f}^\star)$ may be reduced through optimization methods like gradient descent.
- In this chapter, we will focus on the next terms,
  $\sup_{f \in \mathcal{H}} \{R(f) - r(f)\}$, and $\sup_{f \in \mathcal{H}} \{r(f) - R(f)\}$.
- Our goal is to derive **maximal inequalities**, i.e. inequalities of the form

$$\mathbb{E}\left[\sup_{f \in \mathcal{H}} r(f) - R(f)\right] \leq \frac{\psi(\mathcal{H})}{n^\alpha}$$

where ψ, and α are to determine.

Excess risk decompositon and maximal inequalities

Excess Risk Decomposition

Proposition : Excess Risk Decompositon

For any $f \in \mathcal{H}$, we have

$$r(f) - r(f^\star) \leq R(f) - R(\tilde{f}^\star) + \sup_{f \in \mathcal{H}} \{R(f) - r(f)\} + \sup_{f \in \mathcal{H}} \{r(f) - R(f)\}$$

where $\tilde{f}^\star = \underset{f \in \mathcal{H}}{\operatorname{argmin}}\, R(f) = \underset{f \in \mathcal{H}}{\operatorname{argmin}}\, \frac{1}{n} \sum_{i=1}^{N} \ell(f(x_i), y_i)$.

Remarks :

- $R(f) - R(\tilde{f}^\star)$ may be reduced through optimization methods like gradient descent.
- In this chapter, we will focus on the next terms,

  $\sup_{f \in \mathcal{H}} \{R(f) - r(f)\}$, and $\sup_{f \in \mathcal{H}} \{r(f) - R(f)\}$.
- Our goal is to derive **maximal inequalities**, i.e. inequalities of the form

$$\mathbb{E}\left[\sup_{f \in \mathcal{H}} r(f) - R(f)\right] \leq \frac{\psi(\mathcal{H})}{n^{\alpha}}$$

where ψ, and α are to determine.

The case $\mathcal{H}$ finite.

- Let us first consider the case $\mathcal{H}$ finite.

Excess risk decompositon and maximal inequalities

Excess Risk Decomposition

Proposition : Excess Risk Decompositon

For any $f \in \mathcal{H}$, we have

$$r(f) - r(f^\star) \leq R(f) - R(\tilde{f}^\star) + \sup_{f\in\mathcal{H}} \{R(f) - r(f)\} + \sup_{f\in\mathcal{H}} \{r(f) - R(f)\}$$

where $\tilde{f}^\star = \operatorname{argmin}_{f\in\mathcal{H}} R(f) = \operatorname{argmin}_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{N} \ell(f(x_i), y_i)$.

Remarks :

- $R(f) - R(\tilde{f}^\star)$ may be reduced through optimization methods like gradient descent.
- In this chapter, we will focus on the next terms,
  $\sup_{f\in\mathcal{H}} \{R(f) - r(f)\}$, and $\sup_{f\in\mathcal{H}} \{r(f) - R(f)\}$.
- Our goal is to derive **maximal inequalities**, i.e. inequalities of the form
  $$\mathbb{E}\left[\sup_{f\in\mathcal{H}} r(f) - R(f)\right] \leq \frac{\psi(\mathcal{H})}{n^\alpha}$$
  where ψ, and α are to determine.

The case $\mathcal{H}$ finite.

- Let us first consider the case $\mathcal{H}$ finite.
- As a consequence of the law of large number, we know that $R(f) \underset{n\to\infty}{\longrightarrow} r(f)$ almost surely, for all $f \in \mathcal{H}$.

Excess risk decompositon and maximal inequalities

Excess Risk Decomposition

Proposition : Excess Risk Decompositon

For any $f \in \mathcal{H}$, we have

$$r(f) - r(f^\star) \leq R(f) - R(\tilde{f}^\star) + \sup_{f\in\mathcal{H}} \{R(f) - r(f)\} + \sup_{f\in\mathcal{H}} \{r(f) - R(f)\}$$

where $\tilde{f}^\star = \operatorname*{argmin}_{f\in\mathcal{H}} R(f) = \operatorname*{argmin}_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{N} \ell(f(x_i), y_i)$.

Remarks :

- $R(f) - R(\tilde{f}^\star)$ may be reduced through optimization methods like gradient descent.
- In this chapter, we will focus on the next terms,

  $\sup_{f\in\mathcal{H}} \{R(f) - r(f)\}$, and $\sup_{f\in\mathcal{H}} \{r(f) - R(f)\}$.
- Our goal is to derive **maximal inequalities**, i.e. inequalities of the form

$$\mathbb{E}\left[\sup_{f\in\mathcal{H}} r(f) - R(f)\right] \leq \frac{\psi(\mathcal{H})}{n^{\alpha}}$$

where ψ, and α are to determine.

The case $\mathcal{H}$ finite.

- Let us first consider the case $\mathcal{H}$ finite.
- As a consequence of the law of large number, we know that $R(f) \xrightarrow[n\to\infty]{} r(f)$ almost surely, for all $f \in \mathcal{H}$.
- Here we seek to establish a **non-asymptotic** bound that holds uniformly over $\mathcal{H}$.

Excess risk decompositon and maximal inequalities

Excess Risk Decomposition

Proposition : Excess Risk Decompositon

For any $f \in \mathcal{H}$, we have

$$r(f) - r(f^\star) \leq R(f) - R(\tilde{f}^\star) + \sup_{f \in \mathcal{H}} \{R(f) - r(f)\} + \sup_{f \in \mathcal{H}} \{r(f) - R(f)\}$$

where $\tilde{f}^\star = \underset{f \in \mathcal{H}}{\operatorname{argmin}}\, R(f) = \underset{f \in \mathcal{H}}{\operatorname{argmin}}\, \frac{1}{n} \sum_{i=1}^{N} \ell(f(x_i), y_i)$.

Remarks :

- $R(f) - R(\tilde{f}^\star)$ may be reduced through optimization methods like gradient descent.
- In this chapter, we will focus on the next terms,

  $\sup_{f \in \mathcal{H}} \{R(f) - r(f)\}$, and $\sup_{f \in \mathcal{H}} \{r(f) - R(f)\}$.
- Our goal is to derive **maximal inequalities**, i.e. inequalities of the form

$$\mathbb{E}\left[\sup_{f \in \mathcal{H}} r(f) - R(f)\right] \leq \frac{\psi(\mathcal{H})}{n^\alpha}$$

  where ψ, and α are to determine.

The case $\mathcal{H}$ finite.

- Let us first consider the case $\mathcal{H}$ finite.
- As a consequence of the law of large number, we know that $R(f) \underset{n \to \infty}{\longrightarrow} r(f)$ almost surely, for all $f \in \mathcal{H}$.
- Here we seek to establish a **non-asymptotic** bound that holds uniformly over $\mathcal{H}$.

Proposition : Hoeffding's inequality

Let X be a real-valued raudan sariable sech that $a \leq X - \mathbb{E}[X] \leq b$. Then for any $\lambda \in \mathbb{R}$, we have

$$\mathbb{E}\left[e^{\lambda(X - \mathbb{E}[X])}\right] \leqslant e^{\lambda^2 \frac{(b-a)^2}{8}}.$$

Excess risk decompositon and maximal inequalities

Excess Risk Decomposition

Proposition : Excess Risk Decompositon

For any $f \in \mathcal{H}$, we have

$$r(f) - r(f^\star) \leq R(f) - R(\tilde{f}^\star) + \sup_{f\in\mathcal{H}} \{R(f) - r(f)\} + \sup_{f\in\mathcal{H}} \{r(f) - R(f)\}$$

where $\tilde{f}^\star = \operatorname{argmin}_{f\in\mathcal{H}} R(f) = \operatorname{argmin}_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{N} \ell(f(x_i), y_i)$.

Remarks :

- $R(f) - R(\tilde{f}^\star)$ may be reduced through optimization methods like gradient descent.
- In this chapter, we will focus on the next terms,

  $\sup_{f\in\mathcal{H}} \{R(f) - r(f)\}$, and $\sup_{f\in\mathcal{H}} \{r(f) - R(f)\}$.
- Our goal is to derive **maximal inequalities**, i.e. inequalities of the form

$$\mathbb{E}\left[\sup_{f\in\mathcal{H}} r(f) - R(f)\right] \leq \frac{\psi(\mathcal{H})}{n^\alpha}$$

where ψ, and α are to determine.

The case $\mathcal{H}$ finite.

- Let us first consider the case $\mathcal{H}$ finite.
- As a consequence of the law of large number, we know that $R(f) \xrightarrow[n\to\infty]{} r(f)$ almost surely, for all $f \in \mathcal{H}$.
- Here we seek to establish a **non-asymptotic** bound that holds uniformly over $\mathcal{H}$.

Proposition : Hoeffding's inequality

Let X be a real-valued raudan sariable sech that $a \leq X - \mathbb{E}[X] \leq b$. Then for any $\lambda \in \mathbb{R}$, we have

$$\mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right] \leqslant e^{\lambda^2 \frac{(b-a)^2}{8}}.$$

Proof : on the black board.

Excess risk decompositon and maximal inequalities

Excess Risk Decomposition

Proposition : Excess Risk Decompositon

For any $f \in \mathcal{H}$, we have

$$r(f) - r(f^\star) \leq R(f) - R(\tilde{f}^\star) + \sup_{f\in\mathcal{H}} \{R(f) - r(f)\} + \sup_{f\in\mathcal{H}} \{r(f) - R(f)\}$$

where $\tilde{f}^\star = \operatorname*{argmin}_{f\in\mathcal{H}} R(f) = \operatorname*{argmin}_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{N} \ell(f(x_i), y_i)$.

Remarks :

- $R(f) - R(\tilde{f}^\star)$ may be reduced through optimization methods like gradient descent.
- In this chapter, we will focus on the next terms,
  $\sup_{f\in\mathcal{H}} \{R(f) - r(f)\}$, and $\sup_{f\in\mathcal{H}} \{r(f) - R(f)\}$.
- Our goal is to derive **maximal inequalities**, i.e. inequalities of the form

$$\mathbb{E}\left[\sup_{f\in\mathcal{H}} r(f) - R(f)\right] \leq \frac{\psi(\mathcal{H})}{n^\alpha}$$

where ψ, and α are to determine.

The case $\mathcal{H}$ finite.

- Let us first consider the case $\mathcal{H}$ finite.
- As a consequence of the law of large number, we know that $R(f) \xrightarrow[n\to\infty]{} r(f)$ almost surely, for all $f \in \mathcal{H}$.
- Here we seek to establish a **non-asymptotic** bound that holds uniformly over $\mathcal{H}$.

Proposition : Hoeffding's inequality

Let X be a real-valued raudan sariable sech that $a \leq X - \mathbb{E}[X] \leq b$.
Then for any $\lambda \in \mathbb{R}$, we have

$$\mathbb{E}\left[e^{\lambda(X - \mathbb{E}[X])}\right] \leqslant e^{\lambda^2 \frac{(b-a)^2}{8}}.$$

Proof : on the black board.

Remarks :

- Hoeffding's inequality is a classical example of concentration inequality.