Super Learners

and their oracle properties

Jinyang Liu (sqf320)

Department of Mathematical Sciences
University of Copenhagen

June 2023

Introduction

Binary regression

Let $O = (Y, X)$ be an observation for $Y \in \{0, 1\}$ and $X \in \mathcal{X}$ for $\mathcal{X} \subseteq \mathbb{R}^d$. We assume that $O \sim P$ for some $P \in \mathcal{P}$.

Let $\Theta = \{\theta \mid \theta : \mathcal{X} \to [0, 1] \text{ measurable}\}$ be the set of **regression functions**. We would like to **estimate** a function $\theta \in \Theta$ such that the mean squared error (MSE) or risk

$$R(\theta, P) = \int L(O, \theta) dP = \int (Y - \theta(X))^2 dP$$

is minimized. It turns out that the conditional expectation

$$x \mapsto E(Y \mid X = x) = P(Y \mid X = x)$$

is what minimizes the MSE. We refer to it as the **regression**.

Example of a regression function

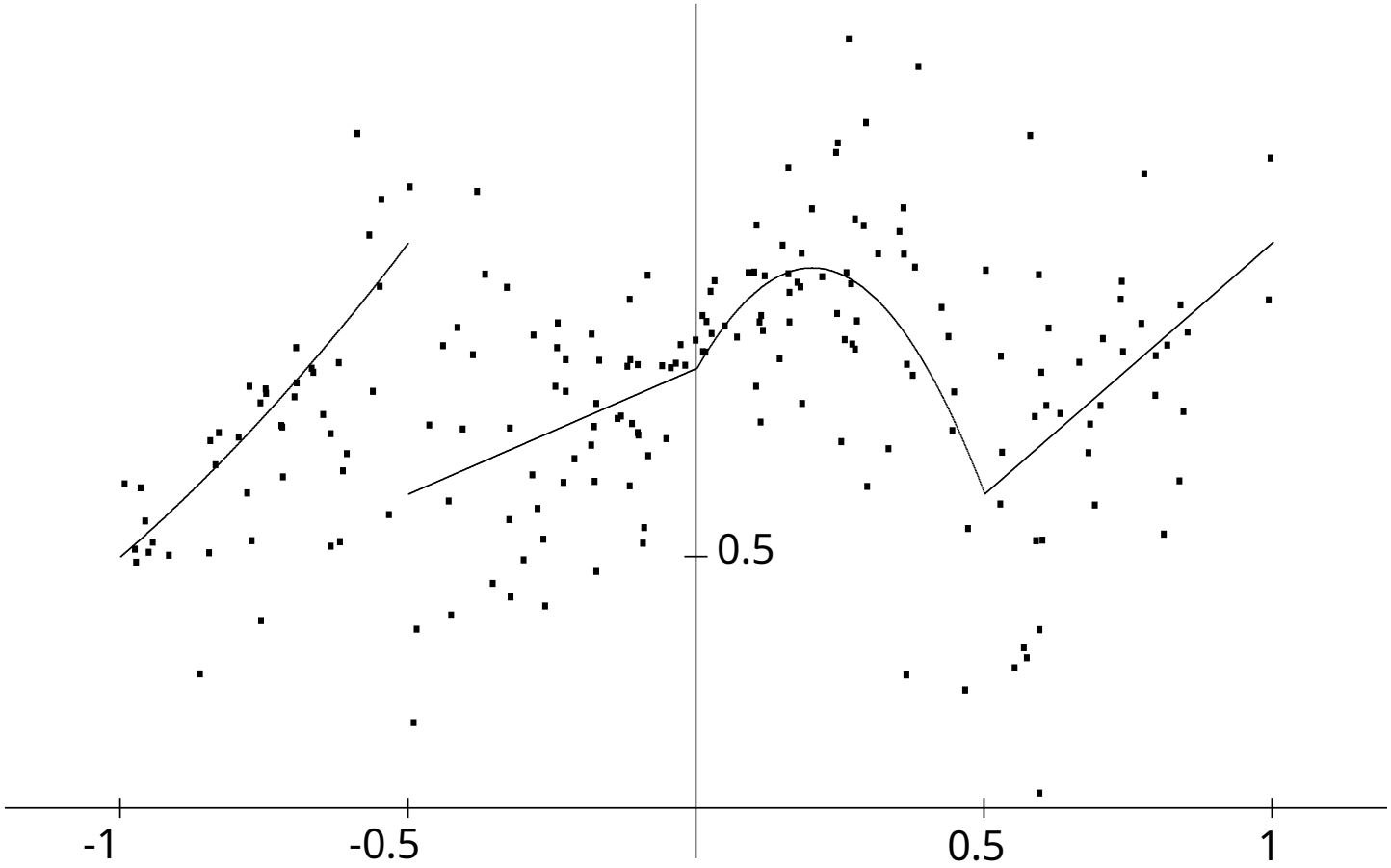

Figure: Example of a pathological regression that can be difficult to learn using parametric techniques. Here a continuous outcome Y is plotted against a single continuous covariate X (Györfi et al., 2002).

Linear approximation

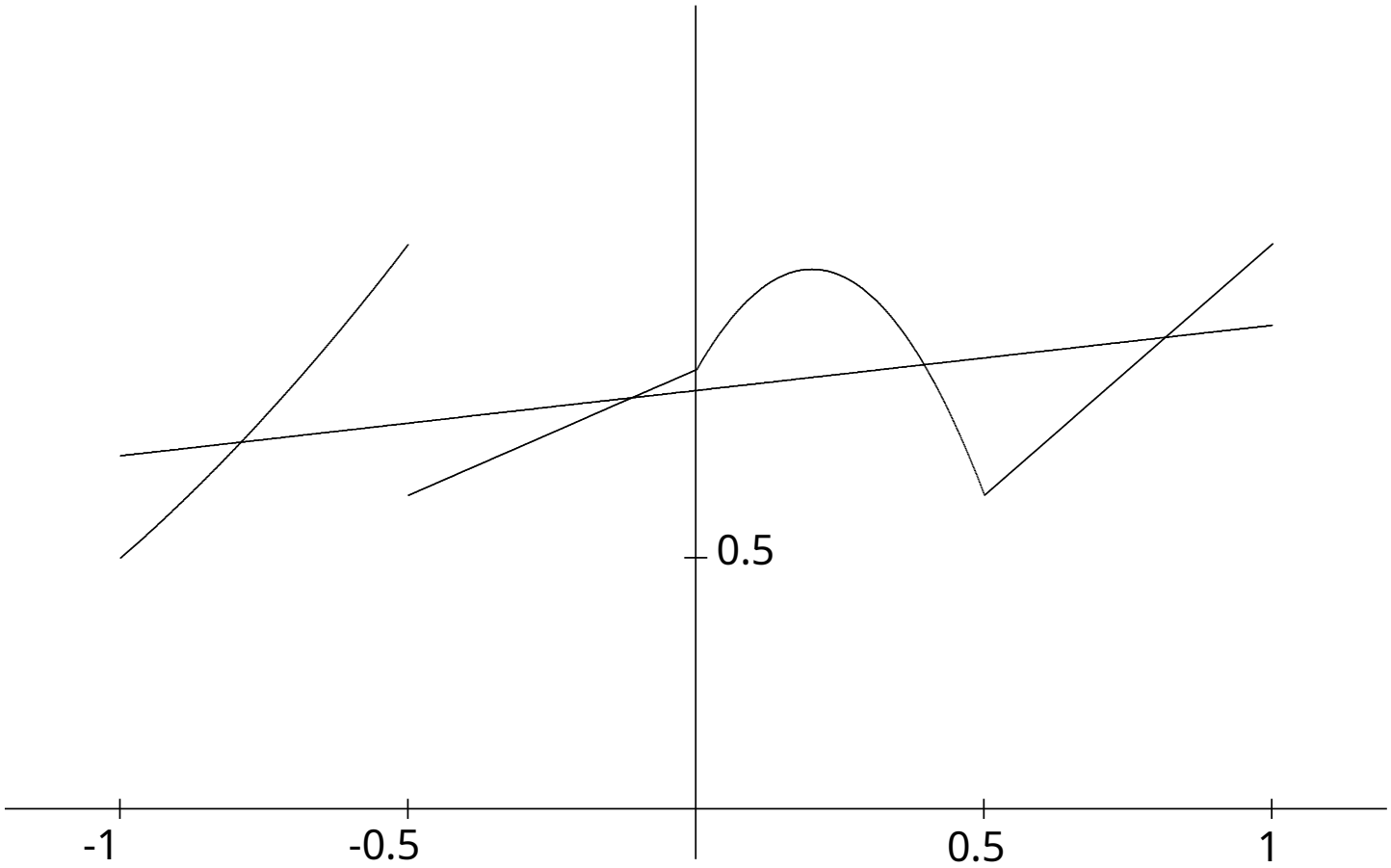

Figure: Approximating the regression using linear regression, which is very biased (Györfi et al., 2002).

Terminology

We observe $D_n = (O_1, \ldots, O_n)$, upon which we apply our **learning algorithms**

Definition (Learning algorithm)

A learning algorithm is a measurable map $\psi : \mathcal{O}^n \to \Theta$ for $n \in \mathbb{N}$.

We assume that the ψ is well-defined for all $n \in \mathbb{N}$ and that the ordering of the observations does not matter.

Definition (Learner or fitted learner)

Let ψ be a learning algorithm, a learner is the outcome of applying ψ to our data D_n denoted as $\psi(D_n)$, which is a map in Θ.

We usually have a **library of learning algorithms**,

$$\Psi = \{\psi_q \mid 1 \leq q \leq k\},$$

for which we can use to estimate the regression.

K-fold Cross-validation

There is a one-to-one correspondance between our data D_n and the empirical measures over n observations

$$P_n = \sum_{i=1}^{n} \delta_{O_i},$$

where δ_{O_i} is the Dirac measure over O_i. K-fold cross-validation splits D_n into **validation** and **training** sets. Index the validation sets by $s \in \{1, \ldots, K\}$ and denote the empirical measure over the validation set s as

$$P^1_{n,s} := \frac{1}{n_1} \sum_{i:s(i)=s} \delta_{O_i}, \qquad P^0_{n,s} := \frac{1}{n_0} \sum_{i:s(i)\neq s} \delta_{O_i}.$$

Here $s(i)$ denotes whether O_i is in the validation set s, and n_1, n_0 are the number of observations in the validation and training sets respectively.

K-fold Cross-validation Procedure

Cross-validation is used to evaluate each algorithm, and is the central idea of the **super learner**:

1. Randomly split D_n into K disjoint and exhaustive validation sets
2. For each $s \in \{1, \ldots, K\}$ fit each $\psi \in \Psi$ on the training data $P^0_{n,s}$ and obtain $\psi(P^0_{n,s})$
3. Use $\psi(P^0_{n,s})$ to predict on the validation set to obtain **level-1 covariates** $Z_i = \left(\psi_1(P^0_{n,s(i)})(X_i), \ldots, \psi_k(P^0_{n,s(i)})(X_i)\right)$
4. Calculate the MSE of ψ on the validation set for $s \in \{1, \ldots, K\}$

$$R(\psi(P^0_{n,s}), P^1_{n,s}) = \frac{1}{n_1} \sum_{i:s(i)=s} \left(Y_i - \psi(P^0_{n,s})(X_i)\right)^2$$

Discrete Super Learner

Cross-validation allows us to select the algorithm with the lowest **empirical risk**

$$\hat{\psi}_n := \arg\min_{\psi \in \Psi} \frac{1}{K} \sum_{s=1}^{K} R(\psi(P^0_{n,s}), P^1_{n,s}),$$

also known as the **cross-validation selected algorihm**. The **discrete super learner** is simply the cross-validation selected algorithm fitted on the entire dataset

$$X \mapsto \hat{\psi}_n(P_n)(X).$$

It is compared to the **oracle selected learning algorithm** that has the true minimum risk

$$\tilde{\psi}_n := \arg\min_{\psi \in \Psi} \frac{1}{K} \sum_{s=1}^{K} R(\psi(P^0_{n,s}), P).$$

Discrete Super Learner: Oracle Equivalence

Corollary (Asymptotic equivalence)

If there exists an $\varepsilon > 0$ *such that*

$$E_{D_n} \frac{1}{K} \sum_{s=1}^{K} R(\tilde{\psi}_n(P^0_{n,s}), P) > \varepsilon \qquad \text{for all } n \in \mathbb{N},$$

and if $n_1 = f(n)$ *for some polynomial function* f*, then the risk of the super learner is asymptotically equivalent with the risk of the oracle selected learner, that is*

$$\lim_{n \to \infty} \frac{E_{D_n} E_{S^n} R(\hat{\psi}_n(P^0_{n,S^n}), P)}{E_{D_n} E_{S^n} R(\tilde{\psi}_n(P^0_{n,S^n}), P)} = 1.$$