



**A Bachelor of Science thesis**

# **Super Learners**

and their oracle properties

Jinyang Liu

Supervised by Prof. Thomas Gerds  
Co-supervised by Prof. Niels Richard Hansen  
Department of Mathematical Sciences  
University of Copenhagen, Denmark

Submitted: March 24, 2023

**Contents**

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>The discrete super learner, dSL</b>	<b>4</b>
2.1	Risks and selectors . . . . .	6
2.2	Oracle inequalities . . . . .	7
<b>3</b>	<b>The ensemble super learner, eSL</b>	<b>7</b>
<b>4</b>	<b>Simulation results</b>	<b>7</b>
<b>5</b>	<b>Discussion</b>	<b>7</b>

# 1 Introduction

Our setup closely models what is described in [VDL06] and [LD03]. Let  $O_1, \dots, O_n$  be  $n$ -i.i.d. observations distributed according to  $P \in \mathcal{P}$  on some measurable space  $(\mathcal{O}, \mathcal{A})$  where  $O_i \in \mathcal{O}$  for each  $i$  and  $\mathcal{P}$  is our statistical model. For a parameter set  $\Theta$  we define the corresponding loss function  $L : \mathcal{O} \times \Theta \rightarrow [0, \infty)$  as a measurable map such that our goal is to find an estimator  $\hat{\theta}$  that minimizes the true risk function  $R : \Theta \rightarrow \mathbb{R}$  given as

$$R(\theta) = \int L(x, \theta) dP(x) = EL(O_1)$$

The parameter set  $\Theta$  can be Euclidean, but for the focus of this thesis we will consider it as a collection of functions of the form  $\theta : \mathcal{X} \rightarrow \mathbb{R}$ .

*Example 1* (Regression functions  $\Theta$ ). Let  $O_1 = (Y_1, X_1), \dots, O_n = (Y_n, X_n) \in \mathcal{O} = \mathbb{R} \times \mathcal{X}$  be i.i.d. observations distributed according to some  $P \in \mathcal{P}$  such that they satisfy the model

$$Y_1 = \theta_0(X_1) + \varepsilon,$$

for an unobservable stochastic error term  $\varepsilon$ . The goal is to estimate an unknown **regression function**  $\theta_0 \in \Theta$  where  $\Theta = \{\theta \mid \theta : \mathcal{X} \rightarrow \mathbb{R}\}$ , is the set of possible regression functions each having  $\mathcal{X}$  as their domain. [VDL06]

*Example 2* (Parametric family). Consider the initial setup from example 1. If  $Y_i$  is  $\mathcal{B}(\mathbb{R}) - \mathcal{B}(\mathbb{R})$  measurable and  $X_i$  is  $\mathcal{F} - \mathcal{B}(\mathbb{R})$  measurable for some sigma-algebra  $\mathcal{F}$  on  $\mathcal{X}$ , then a **generalized regression model** could be considered as parametrized family of distributions,  $\mathcal{P} = \{P_\theta \mid \theta \in \Theta\}$ , given that  $\Theta$  is finite-dimensional.

We can parametrize the conditional probability distributions for  $Y_1$  given  $X_1 = x$  as  $\mathcal{Q} = \{Q_{\theta(x)} \mid \theta \in \Theta\}$  such that  $Q_{\theta(x)}$  is a valid probability distribution on  $\mathcal{B}(\mathbb{R})$  for each  $x \in \mathcal{X}$  and  $\theta \in \Theta$ . For a given  $P_\theta \in \mathcal{P}$  there will exist a  $Q_\theta \in \mathcal{Q}$  such that

$$P_\theta(Y \in A \mid X = x) = Q_{\theta(x)}(A) \quad \text{for all } A \in \mathcal{B}(\mathbb{R}).$$

If we assume that  $X_1$  is distributed according to some  $H_0$  on  $\mathcal{X}$ , then the distribution  $P_\theta$  over our observations (the joint over  $Y$  and  $X$ ) will be

$$P_\theta(X \in A, Y \in B) = \int_A Q_{\theta(x)}(B) dH_0(x)$$

for every  $A \in \mathcal{F}$  and  $B \in \mathcal{B}(\mathbb{R})$ .

*Example 3* (Logistic regression model). Let  $O_1 = (Y_1, X_1), \dots, O_n = (Y_n, X_n) \in \mathcal{O} = \{0, 1\} \times \mathcal{X}$  be i.i.d. observations from some distribution  $P_{\theta_0} \in \mathcal{P}$ , where  $Y_i$  is binary and  $\mathcal{X} \subseteq \mathbb{R}^k$ . We would like to estimate the parameter function  $\theta_0 \in \Theta$

$$\theta_0(x) = E(Y_1 \mid X_1 = x) = P_{\theta_0}(Y_1 = 1 \mid X_1 = x),$$

In logistic regression we assume that  $\Theta = \{x \mapsto \text{expit}(\beta x) \mid \beta \in \mathbb{R}^k\}$ , so  $\theta_0(x) = \text{expit}(\beta_0 x)$ , then the goal becomes to estimate the  $k$ -dimensional parameter  $\beta_0$ , in this case the  $\mathbb{R}^k$  parameter  $\beta_0$  completely determines  $\theta_0$ , so  $\Theta$  is also  $k$ -dimensional. The conditional distributions of  $Y_1$  given  $X_1 = x$  are Bernoulli distributions and can be parametrized as  $\mathcal{Q} = \{\text{Ber}(\text{expit}(\beta x)) \mid \beta \in \mathbb{R}^k\}$ . Now from example 2 we know that the statistical model,  $\mathcal{P}$ , can be parametrized through  $\beta$ , in particular we have

$$\begin{aligned} P_\beta(Y_1 = 1, X_1 \in A) &= \int_A Q_{\theta(x)}(\{1\}) dH_0(x) \\ &= \int_A \text{expit}(\beta x) dH_0(x) \end{aligned}$$

If  $H_0$  has density  $f$  w.r.t. Lebesgue measure, we can write

$$P_\beta(Y_1 = 1, X_1 \in A) = \int_A \text{expit}(\beta x) f(x) dm(x)$$

We will now turn our attention to statistical estimators. Statistical literature commonly write that an estimator is stochastic variable taking values in our parameter space  $\hat{\theta} \in \Theta$ . An estimator is achieved by considering i.i.d. observations  $O_1, \dots, O_n \in \mathcal{O}$  distributed according to some measure  $P$  from some statistical model  $\mathcal{P}$ . We leave the model unspecified as it can be both parametric or nonparametric. Now let  $h : \mathcal{O}^n \rightarrow \Theta$  be a measurable map, an estimator created from  $h$  is the random variable  $T = h(O_1, \dots, O_n)$ . For  $\Theta \subseteq \mathbb{R}^k$  the canonical  $\sigma$ -algebra on  $\Theta$  is the Borel algebra, but when the parameter set is a set of functions, the  $\sigma$ -algebra can only be chosen after careful consideration of constraints on  $\Theta$ .

## 2 The discrete super learner, dSL

In the following section we introduce the terminology “estimator algorithm” which corresponds to the measurable map  $h$  from our finite sample observation space to our parameter space.

**Definition 1** (Estimator algorithm  $h$ ). An estimator algorithm is a measurable map  $h : \mathcal{O}^n \rightarrow \Theta$  for  $n \in \mathbb{N}$ .

**Definition 2** (Statistical Estimator  $\hat{\theta}$ ). Let  $O_1, \dots, O_n \in \mathcal{O}$  be i.i.d. observations distributed according to some  $P \in \mathcal{P}$  for a statistical model  $\mathcal{P}$  on  $\mathcal{O}$ . Let  $h : \mathcal{O}^n \rightarrow \Theta$  be an estimator algorithm. An estimator is the random variable  $\hat{\theta} = h(O_1, \dots, O_n) \in \Theta$ .

There is a one-to-one correspondence between the tuples of i.i.d. observations  $(O_1, \dots, O_n) \in \mathcal{O}^n$  and the empirical measures over  $n$  observations on  $(\mathcal{O}, \mathcal{A})$  defined as

$$P_n(A) = \frac{1}{n} \sum_{i=1}^n \delta_{O_i}(A) \quad \text{for } A \in \mathcal{A}.$$

Note that the empirical measure is a random variable. Thus, we can write  $h(P_n)$  as an alternative representation of the estimator  $h(O_1, \dots, O_n)$ , by adjusting the notation without introducing ambiguity.

*Example 4* (Prediction algorithm). Consider the setup from example 3, where we have i.i.d. observations  $O_1 = (Y_1, X_1), \dots, O_n = (Y_n, X_n)$  such that  $Y_i \in \{0, 1\}$  and  $\mathcal{X} \in \mathbb{R}^k$  and our goal is to estimate the probability  $\theta(x) = P_\theta(Y_1 = 1 \mid X_1 = x) \dots$

We would now like to consider the scenario where we have a library (set) of learner algorithms,  $h_1, \dots, h_n$ . From these algorithms, we can define the set of learners  $\{\hat{\theta}_q = h_q(P_n) \mid 1 \leq q \leq p\}$ , where our goal is to find  $\hat{\theta}_{\hat{q}}(P_n)$ , which denotes the learner that minimizes  $R$  and  $\hat{q}$  may depend on the observations.

In order to find  $\hat{q}$  we have to proceed via cross validation. In cross validation, we randomly split our data into a training set and a test set. Let the random binary vector  $S = (S_1, \dots, S_n) \in \{0, 1\}^n$  be independent of  $X_1, \dots, X_n$  such that  $S_i = 0$  indicates that  $X_i$  should be in the training set and  $S_i = 1$  indicates that  $X_i$  belongs to the test set. We can define the empirical distributions over these two subsets,  $P_{n,S}^0$  and  $P_{n,S}^1$  as

$$P_{n,S}^0 = \frac{1}{n_0} \sum_{i:S_i=0} \delta_{X_i}$$

$$P_{n,S}^1 = \frac{1}{1 - n_0} \sum_{i:S_i=1} \delta_{X_i}$$

Where  $n_0$  would be the number of  $S_i$ 's that are marked 0.

*Example 5* (Random splits). For  $n = 9$  observations one could for example define the distribution of the random vector  $S$  as

$$P(S = (0, 0, 0, 0, 0, 0, 1, 1, 1)) = \frac{1}{3}$$

$$P(S = (0, 0, 0, 1, 1, 1, 0, 0, 0)) = \frac{1}{3}$$

$$P(S = (1, 1, 1, 0, 0, 0, 0, 0, 0)) = \frac{1}{3},$$

i.e. 3-fold cross-validation.

In general for  $n$  observations we have  $2^n$  ways of choosing which observations should be in the training set and in the validation set. It might not be desirable to define the discrete probabilities for  $S$  over  $\{0, 1\}^n$  simply as  $\frac{1}{2^n}$  for each possible combination of training/validation data, since that would also include the combination where  $n_1 = 0$ . To ensure that we always have  $n_1 > 0$ , then let  $n_1$  be given, then we see that there are  $\binom{n}{n_1}$  ways of choosing both the validation and training set. We can therefore define the distribution of  $S$  as

$$P(S = s) = \binom{n}{n_1}^{-1} \quad \text{for each } s \in \{0, 1\}^n \text{ where } \sum_i s_i = n_1$$

## 2.1 Risks and selectors

**Definition 3** (True risk of  $q$ 'th learner averaged over splits). Given the data  $O_1, \dots, O_n \in \mathcal{O}$  and a set of learners  $\{\theta_q(P_{n,S}^0) \mid 1 \leq q \leq p\}, p \in \mathbb{N}$  applied to our training data  $P_{n,S}^0$ . The risks of these learners averaged over some split-variable  $S$  is given as a function of  $q$

$$q \mapsto E_S \int L(o, \theta_q(P_{n,S}^0)) dP(o) = E_S R(\theta_q(P_{n,S}^0))$$

Where  $P$  is the true distribution for our data  $X$ .

**Definition 4** (Oracle selector). The oracle selector is a function  $\tilde{q} : \mathcal{O}^n \rightarrow \{1, \dots, p\}$  which finds the learner that minimizes the true risk given our data  $O_1, \dots, O_n \in \mathcal{O}$ .

$$\tilde{q}(O_1, \dots, O_n) = \arg \min_{1 \leq q \leq p} E_S R(\theta_q(P_{n,S}^0))$$

Where  $P_{n,S}^0$  is the empirical distribution over the training set of  $O_1, \dots, O_n$  as specified by some split-variable  $S$ .

In similar manner to the above the definitions, we can define the cross-validation risk and the cross-validation selector for our learners

**Definition 5** (Cross-validation risk of  $i$ 'th learner averaged over splits). Given the data  $O_1, \dots, O_n \in \mathcal{O}$  and a set of learners  $\{\theta_q(P_{n,S}^1) \mid 1 \leq q \leq p\}, p \in \mathbb{N}$ . The cross-validation risks of these learners averaged over some split-variable  $S$  is given as a function of  $q$

$$q \mapsto E_S \int L(o, \theta_q(P_{n,S}^1)) dP_{n,S}^1(o) = E_S \hat{R}(\theta_q(P_{n,S}^1))$$

Where  $P_{n,S}^1$  is the empirical distribution over the validation of  $O_1, \dots, O_n$ . We write  $\hat{R}$  for the empirical risk over the validation set.

**Definition 6** (Cross-validation selector). The cross-validation selector is a function  $\hat{q} : \mathcal{O}^n \rightarrow \{1, \dots, p\}$  which finds the learner that minimizes the cross-validation risk given our data  $O_1, \dots, O_n \in \mathcal{O}$ .

$$\hat{q}(O_1, \dots, O_n) = \arg \min_{1 \leq q \leq p} E_S \hat{R}(\theta_q(P_{n,S}^1))$$

Where  $\hat{R}$  is the empirical risk over the validation set and  $P_{n,S}^0$  is the empirical distribution over the training set of  $O_1, \dots, O_n$  as specified by some split-variable  $S$ .

We are interested in the risk difference between the cross-validation selector and the oracle selector, we remark that the optimal risk is attained at the true value  $\theta_0$

$$R(\theta_0) = \int L(o, \theta_0) dP(o),$$

and clearly it is the case that  $R(\theta_0) \leq R(\theta)$  for any learner  $\theta$  of  $\theta_0$ . Given a set of learners we define the centered conditional risk as the difference

$$\begin{aligned} \Delta_S(\theta_{\hat{q}}, \theta_0) &= R(\theta_{\hat{q}}(P_{n,S}^1)) - R(\theta_0) \\ &= E_S \int L(o, \theta_{\hat{q}}(P_{n,S}^1)) - L(o, \theta_0) dP(o) \end{aligned}$$

## 2.2 Oracle inequalities

We introduce the notation  $Pf$  for the integral  $\int f dP$  of an integrable function  $f$  with respect to  $P$ . Additionally, if  $P_n$  represents the empirical measure of  $O_1, \dots, O_n$ , we denote the empirical process indexed over an appropriate class of functions  $\mathcal{F}$  as  $G_n f = \sqrt{n}(P_n f - Pf)$ . Furthermore, we extend this notation to  $G_{n,S}^i f = \sqrt{n}(P_{n,S}^i - Pf)$  for the empirical processes that correspond to applying the empirical measure over either the training sample or validation sample.

**Lemma 7** ([VDL06]). *For  $\delta > 0$  it holds that*

$$\begin{aligned} E_S \int L(o, \theta_{\hat{q}}(P_{n,S}^0)) dP(o) &\leq (1 + 2\delta) E_S \int L(o, \theta_{\tilde{q}}(P_{n,S}^0)) dP(o) \\ &\quad + \frac{1}{\sqrt{n_1}} E_S \max_{1 \leq q \leq p} \int L(o, \theta_q(P_{n,S}^0)) d((1 + \delta)G_{n,S}^1 - \delta\sqrt{n_1}P)(o) \\ &\quad - \frac{1}{\sqrt{n_1}} E_S \max_{1 \leq q \leq p} \int L(o, \theta_q(P_{n,S}^0)) d((1 + \delta)G_{n,S}^1 + \delta\sqrt{n_1}P)(o) \end{aligned}$$

*Proof.* See appendix □

The following result is due to [LD03]:

**Theorem 8** (Asymptotic equality). *The cross validation selector  $\hat{q}$  performs asymptotically as well as the oracle selector  $\tilde{q}$  in the sense that*

$$\frac{\Delta_S(\theta_{\hat{q}}, \theta_0)}{\Delta_S(\theta_{\tilde{q}}, \theta_0)} \rightarrow 1 \quad \text{in probability for } n \rightarrow \infty$$

## 3 The ensemble super learner, eSL

## 4 Simulation results

## 5 Discussion

Pellentesque tincidunt sodales risus, vulputate iaculis odio dictum vitae. Ut ligula tortor, porta a consequat ac, commodo non risus. Nullam sagittis luctus pretium. Integer vel nibh at justo convallis imperdiet sit amet ut lorem. Sed in gravida turpis. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Sed in massa vitae ligula pellentesque feugiat vitae in risus. Cras iaculis tempus mi, sit amet viverra nulla viverra pellentesque.