University of California, Berkeley

U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2016 *Paper* 355

Online Cross-Validation-Based Ensemble Learning

David Benkeser* Samuel D. Lendle†

Cheng Ju‡ Mark J. van der Laan**

*Division of Biostatistics, University of California, Berkeley, benkeser@berkeley.edu

†Pandora Media Inc., Oakland, CA, lendle@stat.berkeley.edu

‡Division of Biostatistics, University of California, Berkeley, cju@berkeley.edu

**Division of Biostatistics, University of California, Berkeley, laan@berkeley.edu

Online Cross-Validation-Based Ensemble Learning

David Benkeser, Samuel D. Lendle, Cheng Ju, and Mark J. van der Laan

Abstract

Online estimators update a current estimate with a new incoming batch of data without having to revisit past data thereby providing streaming estimates that are scalable to big data. We develop flexible, ensemble-based online estimators of an infinite-dimensional target parameter, such as a regression function, in the setting where data are generated sequentially by a common conditional data distribution given summary measures of the past. This setting encompasses a wide range of time-series models and as special case, models for independent and identically distributed data. Our estimator considers a large library of candidate online estimators and uses online cross-validation to identify the algorithm with the best performance. We show that by basing estimates on the cross-validation-selected algorithm, we are asymptotically guaranteed to perform as well as the true, unknown best-performing algorithm. We provide extensions of this approach including online estimation of the optimal ensemble of candidate online estimators. We illustrate the practical performance of our methods using simulations and a real data example where we make streaming predictions of infectious disease incidence using data from a large database.

1 Introduction

Currently the size of data sets is growing faster than the speed of processors. It is now common to encounter data on the order of millions or even billions of observations. In these situations, statistical learning is limited more by computation time than sample size. This has led to increased interest in online estimation. Online estimators update a current estimator with a new incoming batch of data without revisiting past data, thereby avoiding the computational limitations associated with big data. As a motivating example, consider a database that records the incidence of an infectious disease over time. Researchers may be interested in developing an algorithm that accurately predicts future incidence of disease based on past incidence and other regional characteristics. However, the scale of the data may be such that re-computing the prediction algorithm with each batch of incoming data would be prohibitively slow. Online algorithms offer a way to ensure fast updating of predictions as new data is accrued.

There is a growing body of literature describing online algorithms, but little in the literature guides how to select from amongst these algorithms in practice. In the setting of small-scale, independent and identically distributed (i.i.d.) data, cross-validation can be used to objectively compare the performance of a library of candidate estimators. Theoretical results guarantee that the estimator that exhibits the best estimated cross-validated performance is asymptotically equivalent with the unknown best estimator in the library (van der Laan and Dudoit, 2003; van der Vaart et al., 2006; van der Laan et al., 2006). These results extend to the best ensemble (i.e., weighted combination) of candidate estimators. Due to this theoretical property, these estimators have been referred to as super learners (van der Laan et al., 2007; Polley et al., 2012). In practice, super learning has been shown to be effective in many settings, including prediction of mortality among the elderly (Rose, 2013), of mortality in the ICU (Pirracchio et al., 2015), and of health care costs (Rose, 2016). However, in the setting of big or streaming data, the existing super learning approach is limited by the computational expense of performing cross-validated model selection with each incoming batch of data. Furthermore, the approach is not applicable in dependent data settings.

In this work, we extend the super learning framework to settings with large-scale dependent data. We propose an online form of cross-validation that is used to identify the best candidate online algorithm in a library of candidate algorithms. We show that under mild conditions the performance of the estimator based on the cross-validation-selected best algorithm is asymptotically equivalent with the performance of the best candidate estimator. This allows researchers to posit many different algorithms for estimation, learn in real time which algorithm is best, and base future estimates on this algorithm. We also propose an online method for identifying the best ensemble of the candidate online estimators. We provide a further extension relevant for i.i.d. observations.

The outline of the remainder is as follows. We formulate the general statistical estimation problem in Section 2 and review key concepts from online literature in Section 3. In Section 4, we introduce online cross-validation and discuss how it can be used to identify the best-performing online estimator from many candidate estimators. We discuss the optimality of the estimator that is based on this cross-validation-selected estimator by comparing its performance to that of the unknown best-performing algorithm. In this section, we also extend our estimator to allow for online estimation of the optimal ensemble of all candidate online algorithms. In Section 5, we conduct a short simulation study and in Section 6, we apply our methods to a data set where the goal is streaming prediction of infectious disease. We conclude with a short discussion.

2 Formulation of the estimation problem

2.1 Statistical model

Suppose at each time t_i, we observe a random variable $O(i)$, $i = 1, \ldots, n$. For example, consider an infectious disease database that records the incidence of an infectious disease in one or several geographic regions. The observed data might consist of $O(i) = \{W(i), Y(i)\}$, where $W(i)$ corresponds to characteristics of a region at time i such as vaccination rates, while $Y(i)$ corresponds to the incidence rate of the infectious disease. Let P_0^n be the true probability distribution of $O(1), \ldots, O(n)$, and let p_0^n be its density with respect to a dominating measure μ^n. The likelihood of an observation $o = \{o(1), \ldots, o(n)\}$ can be factorized according to time-ordering as follows:

$$p_0^n(o) = \prod_{i=1}^{n} p_{0,i}(o(i) \mid \bar{O}(i-1) = \bar{o}(i-1)) \ ,$$

where we defined $\bar{O}(i-1) = \{O(1), \ldots, O(i-1)\}$. If we make no further assumptions about P_0^n, the statistical estimation problem is intractable – we only have a single observation from P_0^n, which limits our ability to learn about the underlying data generating process. The problem could be greatly simplified by making the usual i.i.d. assumption, which would allow us to write

$$\begin{aligned} p_0^n(o) &= \prod_{i=1}^{n} p_{0,i}(o(i) \mid \bar{O}(i-1) = \bar{o}(i-1)) \\ &= \prod_{i=1}^{n} \bar{p}_0(o(i)) \ , \end{aligned} \tag{1}$$

where $\bar{p}_0$ is an unconditional density common to each observation. However, in many settings such an independence assumption is not justified. For example, in the infectious disease setting the incidence at a given time $Y(i)$ might depend on past disease

incidence and past vaccination rates. For some diseases, it may be reasonable to assume that the incidence $Y(i)$ is independent of past data conditional on the previous k measurements, $Z(i) = \{O(j) : j = i-1, i-2, \ldots, i-k\}$. In general, we expect to encounter settings where an observation $O(i)$ is independent of past observations given some fixed-dimension summary of the past data, $Z(i) = f_i(\bar{O}(i-1))$,

$$\begin{aligned} p_0^n(o) &= \prod_{i=1}^{n} p_{0,i}(o(i) \mid \bar{O}(i-1) = \bar{o}(i-1)) \\ &= \prod_{i=1}^{n} p_{0,i}(o(i) \mid Z(i) = z(i)) \ . \end{aligned}$$

However, with this assumption each observation $O(i)$ still may have a unique conditional distribution. Therefore, we make a stationarity assumption – that is, we assume each observation has a common conditional distribution $\bar{P}_0$ given Z. We use $\bar{p}_0$ to denote the conditional density of $\bar{P}_0$ with respect to a dominating measure μ. We can now express the likelihood of the observed data as

$$\begin{aligned} p_0^n(o) &= \prod_{i=1}^{n} p_{0,i}(o(i) \mid \bar{O}(i-1) = \bar{o}(i-1)) \\ &= \prod_{i=1}^{n} p_{0,i}(o(i) \mid Z(i) = z(i)) \\ &= \prod_{i=1}^{n} \bar{p}_0(o(i) \mid Z(i) = z(i)) \end{aligned}$$

This expression makes clear that the conditional density $\bar{p}_0$ of each observation does not change over time, though the conditioning set Z will change. Nevertheless, with each observation we gain more information about the common conditional distribution of the data. Notice that the assumption of i.i.d. observations (1) is a special case of this assumption, where $Z(i) = \emptyset$ for all i. Another important special case is data generated by a group sequential adaptive design in which the treatment probability is a function of summary measures of the observed data on previously sampled groups (van der Laan, 2008; Chambaz and van der Laan, 2011a,b). More generally, this assumption permits a wide range of time-series models. We define our statistical model $\mathcal{M}$ as a collection of possible stationary distributions $\bar{P}$ that could have given rise to the observed data.

2.2 Statistical target parameter and loss functions

We are interested in learning about a feature of the true data distribution. To formalize this notion, we call the feature of interest the statistical target parameter and

write it as a function $\Psi : \mathcal{M} \rightarrow \mathbf{\Psi}$ that takes a distribution $\bar{P}$ from the model and maps it into the parameter space $\mathbf{\Psi}$. In some cases, we may wish to learn about the entire conditional distribution $\bar{P}_0$; however, often we are satisfied learning about a summary measure of this distribution. For example, in the infectious disease setting we are occasionally interested in the joint conditional distribution of disease and regional characteristics; however, in many cases we are interested in a summary of this distribution, such as the conditional mean of disease incidence given current regional characteristics and past measurements.

Our method for estimation should reflect the choice of the statistical target parameter. For example, to learn about the conditional mean of disease incidence, we could estimate the joint conditional distribution of regional characteristics and disease incidence, which would imply an estimate of the conditional mean. However, such a procedure is not targeted towards the goal of estimating the conditional mean. To ensure parsimony between our estimation procedure and our target parameter, we introduce the notion of a loss function. Suppose we are interested in estimating the target parameter $\psi_0 = \Psi(\bar{P}_0)$. We call $(Z, O, \psi) \rightarrow L(\psi)(Z, O)$ a loss function for ψ_0 if for all z, $E_0\{L(\psi_0)(Z, O) \mid Z = z\} = \min_{\psi \in \mathbf{\Psi}} E_0\{L(\psi_0)(Z, O) \mid Z = z\}$, where we use $E_0(\cdot|Z = z)$ to denote the expectation under $\bar{P}_0$ given $Z = z$. In words, a loss function for a given parameter is defined as a function whose true conditional mean given a summary of the past is minimized by the true value of the parameter.

Returning to the infectious disease example, if we are interested in the full joint conditional density $\bar{p}_0$, we could use negative log-likelihood loss,

$$L(\bar{p})(z, o) = -\log\{\bar{p}(o \mid Z = z)\} \ .$$

For each z, $E_0[-\log\{\bar{p}(O|Z)\} \mid Z = z]$ is minimized by the true conditional density $\bar{p}_0$. If instead we are interested in the conditional mean of disease incidence given current regional characteristics and past disease incidence, we could use the squared-error loss, $L(\psi)(O, Z) = \{Y - \psi(Z, W)\}^2$. Notice that

$$E_0[\{Y - \psi(Z, W)\}^2 \mid Z = z] = E_0\big(E_0\big[\{Y - \psi(Z, W)\}^2 \ \big| \ Z = z, W\big] \ \big| \ Z = z\big) \ ,$$

where the inner expectation is taken over the conditional distribution of Y given $Z = z$ and W. For every (z, w), the inner expectation is minimized over all $\psi \in \mathbf{\Psi}$ by $\psi_0(z, w)$ the true conditional mean of Y given $Z = z$ and $W = w$.

Loss functions play an important role in the development of our methodology in two ways. First, the expectation of a loss function can be used to define a theoretical criteria for comparing an estimator and the truth, as we show in the next section. Second, the empirical mean of the loss serves as a criteria for comparing various estimators of the statistical target parameter and we use this fact to develop our estimator.

3 Online estimation

To introduce key concepts in online estimation, we consider the parametric model and i.i.d. regression setting, where $Z = \emptyset$ and we assume that the mean of Y conditional on W is described by the linear model $\psi_\beta(W) = \beta' W$. This setting has been extensively studied in the online literature in recent years (Zinkevich, 2003; Crammer et al., 2006; Bottou, 2010; Shalev-Shwartz, 2011). Suppose we are interested in $\psi_0(W)$, the conditional mean of Y given W. In the assumed parametric model, estimating ψ_0 corresponds to estimating $\beta_0 = \text{argmin}_\beta E_0\{L(\psi_\beta)(Y, W)\}$ for $\beta_0 \in \mathbb{R}^d$ and an appropriate loss function. Define $\hat{\beta}_n$ as minimizer of the empirical average of the loss function,

$$\hat{\beta}_n = \text{argmin}_\beta \frac{1}{n} \sum_{i=1}^{n} L(\psi_\beta)(O(i)) \ .$$

For example, if we assume the parametric model $\{p_\beta : \beta \in \mathbb{R}^d\}$ and let $L(\psi_\beta) = -\log p_\beta$, then $\hat{\beta}_n$ is the maximum likelihood estimator of β_0.

To study the performance of $\psi_{\hat{\beta},n}$ as an estimator of ψ_0, we can construct loss-based dissimilarity measures. The measure $d_{0n}(\psi_{\hat{\beta},n}, \psi_0) = E_0\{L(\psi_{\hat{\beta},n})(O) - L(\psi_0)(O)\}$ compares the true average loss when using $\psi_{\hat{\beta},n}$ to the true average loss when using ψ_0. This measure can be decomposed further:

$$\begin{aligned} d_{0n}(\psi_{\hat{\beta},n}, \psi_0) &= E_0\{L(\psi_{\beta,0})(O) - L(\psi_0)(O)\} - E_0\{L(\psi_{\hat{\beta},n})(O) + L(\psi_{\beta,0})(O)\} \\ &= d_{0n}(\psi_{\beta,0}, \psi_0) + d_{0n}(\psi_{\hat{\beta},n}, \psi_{\beta,0}) \ . \end{aligned}$$

The first term is sometimes referred to as the approximation error and describes the average loss incurred by estimating ψ_0 with $\psi_{\beta,0}$. The second term is referred to as the estimation error and describes the average loss incurred by minimizing empirical rather than true mean of the loss function (Bousquet and Bottou, 2008). In big data settings, computing the true minimizer $\hat{\beta}_n$ can be computationally expensive. Rather than carrying out this minimization with great accuracy, online algorithms may be formulated to approximate the minimum.

Stochastic gradient descent is one such online algorithm, which involves an iterative optimization routine that takes a small step in the direction of the negative gradient of the loss function at a randomly selected observation from the data set. We define the recursive updating step

$$\beta_{t+1} = \beta_t - \gamma_t \Gamma_t \frac{d}{d\beta_t} L(\psi_{\beta,t})(O(t)) \tag{2}$$

where γ_t is a scalar step size or learning rate, Γ_t is a $d \times d$ matrix, and $O(t)$ is the observation used at the t-th step (Bottou, 2010). In first-order SGD Γ_t is some constant times the identity matrix, while other variants replace Γ_t with an appropriate diagonal matrix (e.g., diagonal elements of the estimated inverse Hessian) (Duchi

et al., 2011; Zeiler, 2012). Second-order SGD accounts for the curvature of the loss function by using a Γ_t that approximates the inverse Hessian(Murata, 1998). However, computing and storing an estimate of this matrix is often computationally expensive for high-dimensional d and, though it is optimal, second-order SGD is rarely used in practice. There are many other methods for online optimization that have been used in a variety of contexts (Polyak and Juditsky, 1992; Xu, 2011), including settings with regularized loss functions, such as the Lasso regression and support vector machines (Fu, 1998; Langford et al., 2009; Kivinen et al., 2004; Balakrishnan and Madigan, 2008; Shalev-Shwartz et al., 2011).

Regardless of which method is chosen, after t steps we hope that the approximated minimum is sufficiently close to the true minimum. We can again use loss-based dissimilarities to study the performance of $\psi_{\beta,t}$ as an estimator of ψ_0 using

$$d_{0n}(\psi_{\beta,t}, \psi_0) = d_{0n}(\psi_{\beta,0}, \psi_0) + d_{0n}(\psi_{\hat{\beta},n}, \psi_{\beta,0}) + d_{0n}(\psi_{\beta,t}, \psi_{\hat{\beta},n}) \ ,$$

where the first two terms are again the approximation and estimation error, while the new term is the optimization error incurred by using β_t rather than the true minimizer $\hat{\beta}_n$. Existing results in the online learning literature suggest that in big data settings, the estimation and optimization error will be small (Shalev-Shwartz, 2011). Thus, the performance of an online estimator will be determined largely by the approximation error. To minimize the approximation error, we utilize the super learning framework, where we posit a library of candidate estimators for the purpose of estimating ψ_0. It is not possible a-priori to know which estimator will perform best according to our loss-based dissimilarity. However, we can estimate performance of the estimators from the data using cross validation. Cross validation is a sample-splitting technique that involves training a method (e.g., estimating parameters of a parametric model) on a portion of the data, called the training sample, and subsequently evaluating the average loss of those estimators on the withheld portion of the data, called the validation sample. In the following section, we propose an online form of cross validation that can be used to evaluate candidate online estimators in the present setting with large-scale, dependent data.

4 Online super learner

4.1 Online cross-validation for dependent data

Suppose we have K candidate estimators $\hat{\Psi}_k, \ k = 1, \ldots, K$ that can be applied to data sets $\{Z(i), O(i)\}$ for i ranging over a subset of $\{1, \ldots, n\}$. Suppose that these estimators use the first n_ℓ observations to construct initial estimators of $\hat{\Psi}_k$ and proceed with online updates thereafter. For example, if $\hat{\Psi}_k$ is based on a parametric model, maximum likelihood estimation could be used based on the first n_ℓ observations to obtain initial estimates of the model's parameters and stochastic gradient descent

used thereafter to provide online updates of the parameter estimates. To evaluate the performance of different candidate estimators, we use online cross validation to estimate the average loss of each candidate. At time $t_0 \in \{n_\ell + 1, \ldots, n\}$ we define the data received before t_0 as the training sample, and the singleton $O(t_0)$ as the validation sample. We use this sample splitting to evaluate how well an estimator trained on the past is able to predict an outcome at the next time point.

For each t_0, let P_{t_0-1} denote the empirical distribution of the training sample $\{Z(i), O(i) : i = 1, \ldots, t_0 - 1\}$ and let $\hat{\Psi}_{k,t_0-1} = \hat{\Psi}_k(P_{t_0-1})$ denote the estimator $\hat{\Psi}_k$ trained using $\{Z(i), O(i) : i = 1, \ldots, t_0 - 1\}$. Given a candidate estimator $\hat{\Psi}_k$ we define its online cross-validated risk as

$$R_{CV,n}(\hat{\Psi}_k) = \frac{1}{n - n_\ell + 1} \sum_{t_0=n_\ell+1}^{n} L(\hat{\Psi}_{k,t_0-1})(Z(t_0), O(t_0)) \ .$$

Note that if $\hat{\Psi}_k$ is an online estimator, then the online cross-validated risk is also an online estimator, computed recursively as

$$R_{CV,n+1}(\hat{\Psi}_k) = \frac{n - n_\ell + 1}{n - n_\ell + 2} R_{CV,n}(\hat{\Psi}_k) + \frac{1}{n - n_\ell + 2} L(\hat{\Psi}_{k,n})(Z(n+1), O(n+1)) \ .$$

The proposed cross validation thus proceeds as follows: with each new observation $O(t_0+1)$ create $\{Z(t_0+1), O(t_0+1)\}$; evaluate the loss $L(\hat{\Psi}_k(P_{t_0})(Z(t_0+1), O(t_0+1))$ for each k; add this loss to the current estimate of online cross-validated risk; update each online estimator $\hat{\Psi}_{k,t_0}$ into $\hat{\Psi}_{k,t_0+1}$ using $O(t_0 + 1)$. Upon receipt of the next observation $O(t_0 + 2)$, the process is repeated.

4.2 Online cross-validation selector and online oracle selector

The online cross-validated risk $R_{CV,n}(\hat{\Psi}_k)$ gives an empirical measure of performance for each estimator $k = 1, \ldots, K$. Based on this measure, we define

$$k_n = \text{argmin}_k R_{CV,n}(\hat{\Psi}_k)$$

as the online estimator with the best estimated performance, which we refer to as the online cross-validation selector. We can now define a new estimator that at each step t uses the estimates from the online cross-validation selector, $\hat{\Psi}(P_t) = \hat{\Psi}_{k_t}(P_t), t = 1, \ldots, n$. We call this estimator the discrete online super learner. Notice that over time the discrete online super learner could switch from one estimator to another. If all the candidate estimators are online estimators, then the discrete super learner is itself an online estimator and therefore is as scalable as any of the candidate estimators.

We turn to what can be said theoretically about this approach. Note that the online cross-validated risk estimates the following true online cross-validated risk:

$$\tilde{R}_{CV,n}(\hat{\Psi}_k) = \frac{1}{n - n_\ell + 1} \sum_{t_0=n_\ell+1}^{n} E_0\{L(\hat{\Psi}_{k,t_0-1})(O(t_0)) \mid Z(t_0) = z(t_0)\} \ . \qquad (3)$$

This is the sum over all times of the true average loss for the estimator $\hat{\Psi}_k$ with respect to the conditional distribution of $O(t_0)$ given $Z(t_0)$ equals the observed value $z(t_0)$ and is minimized by ψ_0. To study how the performance of a particular estimator compares to the true parameter, we can define an online loss-based dissimilarity,

$$d_{0n}(\hat{\Psi}_k, \psi_0) = \frac{1}{n - n_\ell + 1} \sum_{t_0 = n_\ell + 1}^{n} E_0\{L(\hat{\Psi}_{k,t_0-1})(O(t_0)) - L(\psi_0)(O(t_0)) \mid Z(t_0) = z(t_0)\} .$$

For certain loss functions, this measure can be made more intuitive by re-writing the comparison as a difference between the estimator and the truth. Consider the squared L^2 distance between ψ and ψ_0 under the conditional distribution $\bar{P}_0$ of $W(t_0), Y(t_0)$ given $Z(t_0) = z$,

$$d^2_{L2,z}(\psi, \psi_0) = E_0[\{\psi(W, Z) - \psi_0(W, Z)\}^2 \mid Z(t_0) = z)] .$$

When considering squared-error loss, the loss-based dissimilarity $d_{0n}(\hat{\Psi}_k, \psi_0)$ can be written

$$\begin{aligned} &\frac{1}{n - n_\ell + 1} \sum_{t_0 = n_\ell + 1}^{n} E_{W,0}\left[\left\{ \hat{\Psi}_{k,t_0-1}(W, Z(t_0)) - \psi_0(W, Z(t_0)) \right\}^2 \,\middle|\, Z(t_0) = z(t_0) \right] \\ &\quad = \frac{1}{n - n_\ell + 1} \sum_{t_0 = n_\ell + 1}^{n} d^2_{L2,z(t_0)}(\hat{\Psi}_{k,t_0-1}, \psi_0) . \end{aligned}$$

This shows that the loss-based dissimilarity is equal to the sum over the observations of the squared L^2 distance between $\hat{\Psi}_k$ and ψ_0, where at each time t_0 the average is computed with respect to the distribution of $\{W(t_0), Y(t_0)\}$ conditional on $Z(t_0) = z(t_0)$. Similarly, with binary Y and a log-likelihood loss criteria, we can show that $d_{0n}(\hat{\Psi}_k, \psi_0)$ has an interpretation as the Kullback-Leibler divergence of $\hat{\Psi}_k$ and ψ_0 under the conditional distribution of O given $Z = z$.

We have now argued that the online loss-based dissimilarity is an interesting way to compare an estimator to true value of the unknown target parameter in the online setting. We therefore can consider which of the candidate estimators minimizes this online loss-based dissimilarity and define

$$\tilde{k}_n = \arg\min_k \frac{1}{n - n_\ell + 1} \sum_{t_0 = n_\ell + 1}^{n} d_{0n}(\hat{\Psi}_k, \psi_0) .$$

We call this index the online oracle selector. Of course, the oracle selector is unknown in practice as it depends on the true conditional distribution of the data. Nevertheless, we can compare the performance of the online discrete super learner to that of the oracle selector. In Section A of the Appendix, we provide a formal theorem that

establishes a finite-sample inequality comparing the discrete online super learner to the online oracle selector. This inequality can be used to show that

$$\frac{d_{0n}(\hat{\Psi}_{k_n}, \psi_0)}{d_{0n}(\hat{\Psi}_{\tilde{k}_n}, \psi_0)} \to 1 \ ,$$

as n goes to infinity. That is, the performance of the discrete online super learner is asymptotically equivalent with the performance of the online oracle selector. The formal proofs of these results are included in Sections B-E of the Appendix. We present an additional theorem in the i.i.d. setting for a cross-validation scheme that mimics classic V-fold cross-validation in Section F of the Appendix.

An appealing feature of our results is that the number of candidate online algorithms considered can be quite large and is allowed to grow with n. For example, our results admit schemes that consider n^2 different algorithms. Thus, the number of candidate online algorithms that one can practically consider is limited far more by computational considerations than statistical considerations. In practice, these results imply that we have the ability to posit a vast number of online algorithms and allow the data to teach us which is best. For example, consider the problem of making streaming predictions about infectious disease incidence. The online super learning framework allows us to query many infectious disease experts to gather interesting ideas for how to construct online prediction algorithms. The various prediction algorithms are updated with each incoming data point and at any time we can make a prediction based on the algorithm that has given the best predictions in the past. With enough data, our results guarantee that we will be making predictions that are as good as if we had known a-priori which algorithm was best for predicting disease incidence.

4.3 Online ensemble of candidate estimators

We now consider how to create a more flexible online learner by considering an ensemble of a given set of estimators. We define $\hat{\Psi}_\alpha$ as a combination of K estimators indexed by a finite-dimensional vector of coefficients α, e.g., a convex linear combination

$$\hat{\Psi}_\alpha = \sum_{k=1}^{K} \alpha(k)\hat{\Psi}_k \text{ where } \alpha \in \left\{ x \in \mathbb{R}^K : x(k) \geq 0, \sum_{k=1}^{K} x(k) = 1 \right\} .$$

Let $R_{CV,n}(\hat{\Psi}_\alpha)$ be the online cross-validated risk given by

$$R_{CV,n}(\hat{\Psi}_\alpha) = \frac{1}{n - n_\ell + 1} \sum_{t_0=n_\ell+1}^{n} L(\hat{\Psi}_{\alpha,t_0-1})(Z(t_0), O(t_0)) \ ,$$

and let α_n be the choice of α that minimizes online cross-validated risk, $\alpha_n = \text{argmin}_\alpha R_{CV,n}(\hat{\Psi}_\alpha)$. Tracking each online estimator $\hat{\Psi}_\alpha$ for all α only involves tracking

the K online estimators $\hat{\Psi}_k$, but α_n is itself not an online estimator since it involves recomputing the minimum for each n. Therefore, we propose to approximate the minimum α_n with a stochastic gradient descent algorithm.

We define $S_{n,\alpha} = \frac{d}{d\alpha} L(\hat{\Psi}_{\alpha,n-1})$ as the score vector for α and c_n as an appropriate diagonal matrix. For example, if $L(\hat{\Psi}_\alpha)$ is twice differentiable and K is small, we could define $S^1_{n,\alpha} = \frac{d}{d\alpha} S_{n,\alpha}$ as the matrix of second derivatives and

$$c_n = \left\{ -\frac{1}{n - n_\ell + 1} \sum_{t_0 = n_\ell + 1}^{n} S^1_{t_0 - 1, \alpha_{t_0 - 1}}(Z(t_0), O(t_0)) \right\}^{-1} ,$$

as the inverse of the estimated Hessian. The stochastic gradient descent estimator approximating α_n is defined by

$$\alpha^*_{n+1} = \alpha^*_n + \frac{1}{n+1} c_n S_{n,\alpha^*}(Z(n+1), O(n+1)) \ .$$

This updating step can be refined by checking whether

$$L(\hat{\Psi}_{\alpha^*_{n+1}})(O(n+1)) \leq L(\hat{\Psi}_{\alpha^*_n})(O(n+1))$$

and if not, replacing α^*_{n+1} by a convex linear combination of α^*_n and α^*_{n+1} for which there is an actual reduction in the loss.

We refer to this estimator as the online super learner and as above, we can define an oracle selector for this class of estimators as the choice of weights that minimizes the true average of the loss-based discrepancy:

$$\tilde{\alpha}_n = \arg\min_\alpha \frac{1}{n - n_\ell + 1} \sum_{t_0 = n_\ell + 1}^{n} E_0\{ L(\hat{\Psi}_{\alpha, t_0 - 1})(O) - L(\psi_0)(O) \mid Z(t_0) = z(t_0) \} \ .$$

Our oracle results extend to this setting and we can show that

$$\frac{d_{0n}(\hat{\Psi}_{\alpha^*_n}, \psi_0)}{d_{0n}(\hat{\Psi}_{\tilde{\alpha}_n}, \psi_0)} \to 1 \ ,$$

as n goes to infinity. That is, the performance of the online super learner is asymptotically equivalent with the optimal ensemble of candidate estimators.

5 Simulation for independent identically distributed data

We studied the performance of the online super learner in the setting of i.i.d. data consisting of a binary outcome Y and seven other covariates $W = (W_1, \ldots, W_7)$. The components of W were independent and distributed as follows: $W_1 \sim \text{Uniform}(-4, 4)$,

Name	Formula
GLM1	$W_1^2 + W_2 + W_3 + W_4 + W_5 + W_6 + W_7$
GLM2	$W_1 + W_2 * W_3 + W_4 * W_6 + W_5 + W_7$
GLM3	$W_1 + W_1^2 + W_2 + W_3 + W_4 + W_5 + W_6 + W_7^3$
GLM4	$W_1 + W_1^2 + W_3 + W_4 * W_6 + W_5 + W_7$
GLM5	$W_1 + W_2 * W_3 + W_4 + W_5 + W_6 + W_7^3$
GLM6	$W_1 * W_2 * W_3 * W_4 * W_5 * W_6 * W_7$
GLM7	$W_1 + W_2 + W_3 + W_4 + W_5 + W_6 + W_7$
GLM8	$W_1^2 + W_2^2 + W_3 + W_4^2 + W_5^2 + W_6 + W_7$

Table 1: Super learner library for the simulation. The formula column shows the regression formula for each model. Here X^d to denotes the inclusion of polynomial terms for variable X up to degree d, while $X * Y$ denotes inclusion of both main effects and cross-product interaction terms for variables X and Y.

$W_2 \sim \text{Normal}(0, 1)$, $W_3 \sim \text{Bernoulli}(0.5)$, $W_4 \sim \text{Uniform}(-4, 4)$, $W_5 \sim \text{Normal}(0, 1)$, $W_6 \sim \text{Bernoulli}(0.25)$, and $W_7 \sim \text{Uniform}(0, 1)$. The true conditional mean of Y was given by

$$\text{logit}(\psi_0(W)) = -2 + 0.1W_1^2 + W_2W_3 - W_4W_6 - W_5 + 0.7\text{log}W_7$$

The candidate online algorithms used by the online super learner were first-order stochastic gradient descent algorithms used to estimate the parameters of the eight different logistic regression models shown in Table 1. Note that none of the parametric models was correctly specified, as would be expected in practice. The online super learner was constructed using negative log-likelihood loss as loss function and a logistic ensemble

$$\hat{\Psi}_\alpha = \text{expit}\left\{\sum_{k=1}^{K} \alpha(k)\text{logit}(\hat{\Psi}_k)\right\} \text{ where } \alpha \in \left\{x \in \mathbb{R}^K : x(k) \geq 0, \sum_{k=1}^{K} x(k) = 1\right\}.$$

The super learner weights were updated using a first-order stochastic gradient descent algorithm plus a projection step to ensure the sum of the weights was equal to one at each step. We considered sample sizes of $1e4$, $5e4$, $1e5$, $5e5$ and $1e6$ and performed 500 simulations for each sample size/data-generating mechanism combination. We set $n_\ell = 200$ and used the first 100 observations to obtain initial estimates of the parameters of the online SGD algorithms and the second 100 observations to obtain initial estimates of the super learner weights. For each simulation, we evaluated the algorithms on true risk calculated numerically on an independent test set of size $1e6$.

The results of the simulation are shown in Figure 1. The best performing of the candidate algorithms was GLM2, which accounted for both covariate interactions in ψ_0. However, the performance of this algorithm was notably inferior to both super learners. The online super learner had the lowest average risk across the 500 simulations, followed by the online discrete super learner.

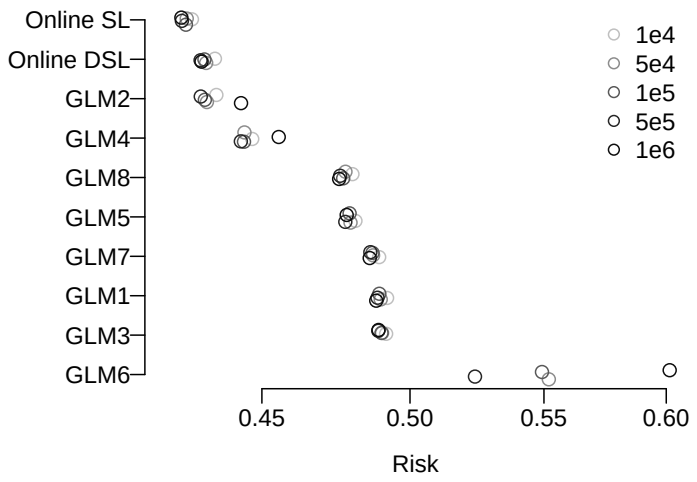

Figure 1: Results from the simulation study. The average risk across 500 simulations is shown for each online algorithm and sample size. Online SL is the online super learner while Online DSL is the online discrete super learner.

6 Online prediction of infectious disease incidence

We used the online super learner to make predictions of disease incidence using data assembled by Project Tycho (Van Panhuis et al., 2013). The Project Tycho database is freely available and includes weekly notifiable disease reports for several infectious diseases in the United States. We analyzed the standardized incidence per 100,000 population of Hepatitis A infections. These measures date back to January 1966 and include a total of 90,839 reports. We used the data from years 1966-1968 to generate initial estimates for our candidate online learners and subsequently used the online super learner to make streaming predictions of the weekly standardized incidence of Hepatitis A in each state for each week recorded from 1968-2011. At each week, we based our predictions on the incidence of disease in the previous four weeks. However, many states had at least some missing weekly incidence recordings, so at time t, we used the summary measure $Z(t) = \{M(t-i), \tilde{Y}(t-i) = M(t-i)Y(t-i) : i = 1, \ldots, 4\}$ to generate predictions, where $M(i)$ is the indicator of incidence being recorded at time i. We used the bounded negative log-likelihood loss function to evaluate our predictions,

$$L(\psi)(Y, Z) = -\frac{Y}{u}\log\left\{\frac{\psi(Z)}{u}\right\} - \left(1 - \frac{Y}{u}\right)\log\left\{1 - \frac{\psi(Z)}{u}\right\},$$

where u denotes the upper bound on disease incidence, here set to be 41.6, the maximum observed value. As a simple proof-of-concept, we considered a limited library of candidate online learners consisting of various bounded logistic regression models

Name	Formula
GLM1	1
GLM2	$M(t-1) + \tilde{Y}(t-1)$
GLM3	$M(t-1) + M(t-2) + \tilde{Y}(t-1) + \tilde{Y}(t-2)$
GLM4	$M(t-1) + M(t-2) + M(t-3) + \tilde{Y}(t-1) + \tilde{Y}(t-2) + \tilde{Y}(t-3)$
GLM5	$M(t-1) + M(t-2) + M(t-3) + M(t-4) + \tilde{Y}(t-1)$ $+\tilde{Y}(t-2) + \tilde{Y}(t-3) + \tilde{Y}(t-4)$
GLM6	$M(t-1) + M(t-2) + \tilde{Y}(t-1) + \tilde{Y}(t-2) + I(\tilde{Y}(t-2) > 0)\tilde{Y}(t-1)$
GLM7	$M(t-1) + M(t-2) + \tilde{Y}(t-1) + \tilde{Y}(t-2) + I(\tilde{Y}(t-2) > 0)\tilde{Y}(t-1)$ $+I(\tilde{Y}(t-3) > 0)\tilde{Y}(t-1)$
GLM8	$M(t-1) + M(t-2) + \tilde{Y}(t-1) + I(\tilde{Y}(t-2) > 0)$
GLM9	$M(t-1) + M(t-2) + M(t-3) + \tilde{Y}(t-1) + I(\tilde{Y}(t-2) > 0)$ $+I(\tilde{Y}(t-3) > 0)$

Table 2: Super learner library for the Tycho data analysis. The formula column shows the regression formula for each bounded logistic regression model used in the analysis. The formula "1" denotes an intercept only model.

with parameters estimated via first-order stochastic gradient descent, where we define a bounded logistic regression model as a logistic regression on the transformed outcome $Y/u \in (0, 1)$. The regression formulas for the various models are shown in Table 2.

We evaluated the performance of the candidate estimators and online super learners based on two criteria: the online cross-validated risk and the risk calculated on a validation set consisting of the final 1,000 recorded weekly reports from 2011-2012, which were withheld from the initial training. The online cross-validated risk is the average loss incurred on weekly predictions made with a given algorithm between 1968 and 2011. The out-of-sample predictive risk is an estimate of the risk of using the predictions from the final models trained using data through 2011 to make predictions of Hepatitis A incidence in the future. The results of the analysis are shown in Figure 2. The online super learner performed the best in terms of online cross-validated risk followed by GLM5 and the discrete online super learner. The discrete online super learner (GLM5) performed best in terms of validation risk followed closely by the online super learner.

7 Discussion

The online super learner can be used for estimation of any common parameter of the conditional probability distribution of $O(t)$, given $\bar{O}(t-1)$ that minimizes the conditional expectation of a loss function. Our results demonstrate that under weak conditions, this super learner will be asymptotically equivalent with the oracle-selected

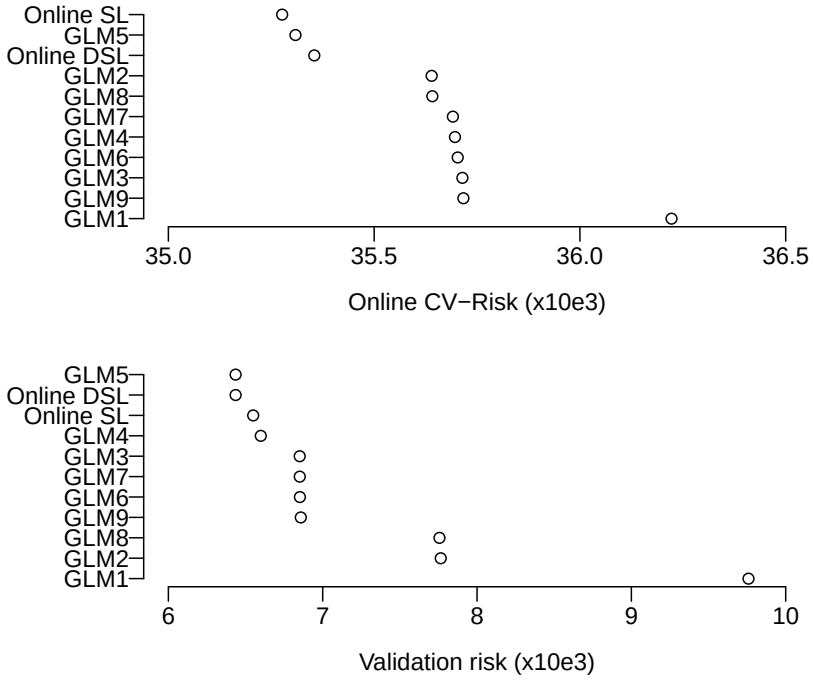

Figure 2: Results for Hepatitis A prediction. The top panel shows the ordered online cross-validated risk of the super learners and candidate online algorithms. The bottom panel shows the ordered risk on the validation data.

estimator. These results therefore provide a powerful way to optimally combine multiple estimators in the online, dependent data setting. The results have implications for the case that the statistical target parameter is a pathwise differentiable (typically, low dimensional) parameter of $\bar{P}_0$. In this case, an online asymptotically normally distributed, efficient estimator can be constructed using targeted minimum loss-based estimation (van der Laan and Rubin, 2006; van der Laan, 2008; van der Laan and Rose, 2011). Such an estimator relies on good initial estimators of certain key nuisance parameters, such as conditional means or densities. The online super learner can therefore be used to aid in construction of online estimators for pathwise differentiable parameters of nonparametric time series models of the type defined in this article.

We expect that the oracle inequality we establish will hold under weaker stationarity assumptions. In particular, depending on the target parameter, the theorem may permit the sole inclusion of stationarity assumptions on relevant portions of the conditional probability distribution of $O(t)$ given $Z(t)$. For example, in the infectious disease prediction problem, our results may allow for the conditional distribution of $W(t)$ given $Z(t)$ to change over time, so long as the conditional distribution of $Y(t)$ given $W(t)$ and $Z(t)$ remains stationary. Confirming this result is left to future work. Also left to future work is implementing the online super learning in a fast, parallelized manner. Such an implementation could ensure that the online super learner requires no more computation time than the slowest candidate online estimator. It will also be important to develop software that incorporates a large library of candidate online estimators, as the performance of the online super learner is limited only by the performance of the best of its constituent online algorithms. The super learner could also be used to select tuning parameters for a single online algorithm, such as Lasso regression or support vector machines. However, as there is unlikely to be a single online algorithm that performs well in every setting and we expect superior performance by considering a large and diverse set of candidate online learners.

Acknowledgment

This work is funded by NIH-grant 5R01AI074345-07 and Bill and Melinda Gates Foundation Grant OPP1147962.

References

Suhrid Balakrishnan and David Madigan. Algorithms for sparse linear classifiers in the massive data setting. *Journal of Machine Learning Research*, 9(Feb):313–337, 2008.

Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

Olivier Bousquet and Léon Bottou. The tradeoffs of large scale learning. In *Advances in neural information processing systems*, pages 161–168, 2008.

A. Chambaz and M.J. van der Laan. Targeting the optimal design in randomized clinical trials with binary outcomes and no covariate. Technical Report 258, Division of Biostatistics, University of California, Berkeley, 2010. Available at `http://biostats.bepress.com/ucbbiostat/paper258/`.

A. Chambaz and M.J. van der Laan. Targeting the optimal design in randomized clinical trials with binary outcomes and no covariate, theoretical study. *Int J Biostat*, 7(1):1–32, 2011a. Working paper 258, www.bepress.com/ucbbiostat.

A. Chambaz and M.J. van der Laan. Targeting the optimal design in randomized clinical trials with binary outcomes and no covariate, simulation study. *Int J Biostat*, 7(1):33–, 2011b. Working paper 258,www.bepress.com/ucbbiostat.

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7 (Mar):551–585, 2006.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 999999:2121–2159, 2011.

Wenjiang J Fu. Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3):397–416, 1998.

Jyrki Kivinen, Alexander J Smola, and Robert C Williamson. Online learning with kernels. *IEEE Transactions on Signal Processing*, 52(8):2165–2176, 2004.

John Langford, Lihong Li, and Tong Zhang. Sparse online learning via truncated gradient. *Journal of Machine Learning Research*, 10(Mar):777–801, 2009.

Noboru Murata. A statistical study of on-line learning. *Online Learning and Neural Networks. Cambridge University Press, Cambridge, UK*, 1998.

Romain Pirracchio, Maya L Petersen, Marco Carone, Matthieu Resche Rigon, Sylvie Chevret, and Mark J van der Laan. Mortality prediction in intensive care units with the super icu learner algorithm (sicula): a population-based study. *The Lancet Respiratory Medicine*, 3(1):42–52, 2015.

E.C. Polley, Sherri Rose, and M.J. van der Laan. Super learning. In M.J. van der Laan and S. Rose, editors, *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, New York Dordrecht Heidelberg London, 2012.

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

Sherri Rose. Mortality risk score prediction in an elderly population using machine learning. *American Journal of Epidemiology*, 177(5):443–452, 2013.

Sherri Rose. A machine learning framework for plan payment risk adjustment. *Health services research*, 2016.

Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.

Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical programming*, 127(1): 3–30, 2011.

M.J. van der Laan. Estimation based on case-control designs with known prevalance probability. *The International Journal of Biostatistics*, page http://www.bepress.com/ijb/vol4/iss1/17/, 2008.

M.J. van der Laan and S. Dudoit. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. Technical report, Division of Biostatistics, University of California, Berkeley, November 2003.

M.J. van der Laan and S. Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, New York, 2011.

M.J. van der Laan and D. Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.

M.J. van der Laan, S. Dudoit, and A.W. van der Vaart. The cross-validated adaptive epsilon-net estimator. *Statistics and Decisions*, 24(3):373–395, 2006.

M.J. van der Laan, E. Polley, and A. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(25), 2007. ISSN 1.

A.W. van der Vaart, S. Dudoit, and M.J. van der Laan. Oracle inequalities for multi-fold cross-validation. *Statistics and Decisions*, 24(3):351–371, 2006.

Willem G Van Panhuis, John Grefenstette, Su Yon Jung, Nian Shong Chok, Anne Cross, Heather Eng, Bruce Y Lee, Vladimir Zadorozhny, Shawn Brown, Derek Cummings, et al. Contagious diseases in the united states from 1888 to the present. *The New England Journal of Medicine*, 369(22):2152, 2013.

Wei Xu. Towards optimal one pass large scale learning with averaged stochastic gradient descent. *CoRR*, abs/1107.2490, 2011. URL `http://arxiv.org/abs/1107.2490`.

Matthew D Zeiler. Adadelta: An adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. 2003.

Appendix

A Oracle inequalities

For loss functions that yield a quadratic loss-based dissimilarity measure, we have the following formal theorem comparing the online cross-validation selector with the corresponding oracle selector.

Theorem 1 *Consider the model $\mathcal{M}$ for the distribution $\bar{P}_0$ of O given Z and the definition of the target parameter $\Psi : \mathcal{M} \to \mathbf{\Psi}$ as the minimizer of a particular loss function for all z. Consider also the above defined online cross-validation selector k_n and online oracle selector $\tilde{k}_n$. Under assumptions* **A1-A4** *explicitly stated in the Appendix B, for any $\delta > 0$, there exists a constant $C(\delta, M_1, M_2) < \infty$ universal in n and choice of candidate estimators such that*

$$d_{0n}(\hat{\Psi}_{k_n}, \psi_0) \leq (1 + 2\delta) d_{0n}(\hat{\Psi}_{\tilde{k}_n}, \psi_0) + Z_n \ ,$$

where $Z_n = Z_{n1} + Z_{n2}$, $P_0^n(Z_{n2} = 0) \to 1$ as $n \to \infty$, and for $n > n_1$ for some $n_1 < \infty$,

$$E_0 Z_{n1} \leq C(\delta, M_1, M_2) \frac{M_{3n}^2[1 + \log\{K(n)\}]}{n} \ .$$

If Assumption **A4** *does not hold, then*

$$d_{0n}(\hat{\Psi}_{k_n}, \psi_0) = o_P(n^{-1} M_{3n}^3) + o_P(n^{-1} M_{3n}^2 \{1 + \log K(n)\}) \ .$$

For loss functions that yield non-quadratic loss-based dissimilarities, we have the following theorem.

Theorem 2 *Consider the model $\mathcal{M}$ for the distribution $\bar{P}_0$ of O given Z and the definition of the target parameter $\Psi : \mathcal{M} \to \mathbf{\Psi}$ defined as the minimizer of a particular loss function for all z. Consider also the above defined online cross-validation selector k_n and online oracle selector $\tilde{k}_n$. Under assumption* **A5** *explicitly stated in the Appendix B, there exists a constant $C(M_1) < \infty$ universal in n and choice of candidate estimators such that*

$$E_0 d_{0n}(\hat{\Psi}_{k_n}, \psi_0) \leq E_0 d_{0n}(\hat{\Psi}_{\tilde{k}_n}, \psi_0) + C(M_1) \left[\frac{\log\{1 + K(n)\}}{n} \right]^{1/2} \ .$$

B Assumptions for Theorems and Discussion

Assumptions for Theorem 1.
A1. There exist an $M_1 < \infty$ so that

$$\sup_{\psi \in \mathbf{\Psi}} \sup_{i, O(i), Z(i)} \mid L(\psi)(Z(i), O(i)) - L(\psi_0)(Z(i), O(i)) \mid \leq M_1,$$

where the supremum over $Z(i), O(i)$ is taken over a support of the distribution $Z(i), O(i)$.
A2. There exist an $M_2 < \infty$ so that with probability 1

$$\sup_{\psi \in \mathbf{\Psi}} \frac{\bar{P}_{0,z(i)}\{L(\psi) - L(\psi_0)\}^2}{\bar{P}_{0,z(i)}\{L(\psi) - L(\psi_0)\}} \leq M_2 < \infty. \tag{4}$$

A3. There exists a slowly increasing sequence $M_{3n} < \infty$ (e.g., $M_{3n} = \log n$) so that with probability tending to 1, for both $\bar{k}_n = k_n$ and $\bar{k}_n = \tilde{k}_n$, we have

$$\frac{1}{M_{3n}} < \frac{d_{0n}(\hat{\Psi}_{\bar{k}_n}, \psi_0)}{E_0 d_{0n}(\hat{\Psi}_{\bar{k}_n}, \psi_0)} < M_{3n}.$$

A4.

$$n M_{3n}^{-3} \min_k E_0 d_{0n}(\hat{\Psi}_k, \psi_0) \to \infty \text{ as } n \to \infty.$$

Assumption A1 states that the loss function is uniformly bounded by some constant M_1, uniformly in all possible realizations of $O(i), Z(i)$ and the candidate estimators of ψ_0. Assumption A2 is an assumption one expects to hold for quadratic uniformly bounded loss functions, as shown in van der Laan, Dudoit (2003). Assumption A3 states that the mean one random variable $d_{0n}(\hat{\Psi}_{k_n}, \psi_0)/E_0 d_{0n}(\hat{\Psi}_{k_n}, \psi_0)$ (and similarly for $\tilde{k}_n$) falls with probability tending to 1 in an interval slowly growing towards its full support $(0, \infty)$. We anticipate that this assumption will hold for any sequence M_{3n} that converges to infinity such as $M_{3n} = \log n$. Assumption A3 is approximately equivalent with assuming that the mean zero random variable $\log d_{0n}(\hat{\Psi}_{k_n}, \psi_0) - E_0 \log d_{0n}(\hat{\Psi}_{k_n}, \psi_0)$ falls with probability tending to 1 in an interval $[-\log n, \log n]$. Assumption A4 only affects the precise statement of the result. Given that M_{3n} is a sequence that grows arbitrarily slow to infinity, assumption A4 states that the oracle selected estimator converges to ψ_0 at a rate slower than the rate $1/n$ of a maximum likelihood estimator for a correctly specified parametric model. Therefore assumption A4 will typically hold, but if somehow one of the candidate

estimators converges to the truth at the parametric rate $1/n$, then the online super learner converges at an almost equally fast rate $\log(n)/n$.

Assumptions for Theorem 2.
A5. There exist an $M_1 < \infty$ such that

$$\sup_{\psi \in \mathbf{\Psi}} \sup_{i, O(i), Z(i)} \mid L(\psi)(Z(i), O(i)) - L(\psi_0)(Z(i), O(i)) \mid \leq M_1,$$

where the supremum over $Z(i), O(i)$ is taken over a support of the distribution $Z(i), O(i)$.

C Preliminary Material for Proof of Theorems

In this section, we show that the difference between the online cross-validated risk and its desired target is a discrete martingale. We then present a theorem from the literature that provides an exponential inequality for the tail probability of such a discrete martingale. Hence, we will make use of the shorthand notation $\bar{P}_{0,z} f = \int f(z, o) d\bar{P}_z(o)$ to denote the expectation of the function $f(z, o)$ under $\bar{P}$ given $Z = z$.

C.1 Centered online cross-validated risk is a discrete martingale

The difference between the online cross-validated risk and the online cross-validated true risk (minimized by oracle selector) can be written as a martingale as follows:

$$\begin{aligned}
&(n - n_\ell + 1)\{R_{CV,n}(\hat{\Psi}_k) - \tilde{R}_{CV,n}(\hat{\Psi}_k)\} \\
&= \sum_{t_0 = n_\ell + 1}^{n} \{L(\hat{\Psi}_k(P_{t_0 - 1}))(Z(t_0), O(t_0)) - L(\psi_0)(Z(t_0), O(t_0))\} \\
&\quad - \sum_{t_0 = n_\ell + 1}^{n} \bar{P}_{0, z(t_0)} \{L(\hat{\Psi}_k(P_{t_0 - 1})) - L(\psi_0)\} \\
&= \sum_{t_0 = n_\ell + 1}^{n} \{f(t_0, \bar{O}(t_0 - 1), O(t_0)) - E_0(f(t_0, \bar{O}(t_0 - 1), O(t_0)) \mid \bar{O}(t_0 - 1))\} \\
&= M_n(f),
\end{aligned}$$

where

$$f(t_0, \bar{O}(t_0 - 1), O(t_0)) = L(\hat{\Psi}_k(P_{t_0 - 1}))(Z(t_0), O(t_0)) \ .$$

For $k < n$, $E_0(M_n(f) \mid \bar{O}(k)) = M_k(f)$, which proves that $(M_n(f) : n = n_\ell + 1, \ldots)$ is a discrete martingale in n.

C.2 Martingale Exponential inequality for tail probability

In order to establish an oracle inequality for the online cross-validation selector based on data $O(1), \ldots, O(n)$, we require an exponential inequality for tail-probabilities of Martingale sums $M_n(f)$. For that purpose, we refer to Theorem 8 (page 40) in (Chambaz and van der Laan, 2010) for the following exponential inequality for Martingales:

Theorem 3 *(Proposition A2 in van Handel, 2009) For the sake of this theorem, let* $M_n(f) = \sum_{i=1}^n f(i, O(i), \bar{O}(i-1)) - \bar{P}_{0,z(i)} f$, $\bar{P}_{0,z(i)}$ *denoting the conditional probability distribution of* $O(i)$*, given* $Z(i)$*, and let* $\mathcal{F}$ *be a set of such functions* f*. Fix* $K > 0$ *and define, for all* $f \in \mathcal{F}$, $n \geq 1$,

$$\tilde{R}_{n,K}(f) = \frac{2K^2}{n} \sum_{i=1}^n \bar{P}_{0,z(i)} \phi \left(\frac{| f |}{K} \right),$$

where $\phi(x) = \exp(x) - x - 1$*. There exists a universal constant* $C > 0$ *(e.g.,* $C = 100$*) such that, for any* $n \geq 1$, $R > 0$,

$$P \left(\sup_{f \in \mathcal{F}} I(\tilde{R}_{n,K}(f) \leq R) \frac{M_n(f)}{n} \geq x \right) \leq 2 \exp \left\{ - \frac{nx^2}{C^2(c_1 + 1)R} \right\}$$

for any $x, c_0, c_1 > 0$ *satisfying* $c_0^2 \geq C^2(c_1 + 1)$ *and*

$$\frac{c_0}{\sqrt{n}} \int_0^{\sqrt{R}} \sqrt{H(\mathcal{F}, \| \cdot \|_\infty, \epsilon)} d\epsilon \leq x \leq \frac{c_1 R}{K}.$$

Here $H(\mathcal{F}, \| \cdot \|_\infty, \epsilon) = \log(1 + N(\mathcal{F}, \| \cdot \|_\infty, \epsilon))$ *is the so called entropy function with respect to supremum norm and* $N(\mathcal{F}, \| \cdot \|_\infty, \epsilon)$ *is the covering number defined as the number of balls with radius* ϵ *that is needed to cover* $\mathcal{F}$*.*

For a specified c_0 and c_1, satisfying $c_0^2 \geq C^2(c_1 + 1)$, R, for x larger than $c_0 E/\sqrt{n}$ and smaller than $c_1 R/K$, the above exponential inequality applies, where

$$E = \int_0^{\sqrt{R}} \sqrt{H(\mathcal{F}, \| \cdot \|_\infty, \epsilon)} d\epsilon \ .$$

On this interval of x-values we have $x \leq c_1 R/K$, which implies $c_1 \geq xK/R$. Therefore, we can restate the above result as follows: For a specified R, c_0, c_1 satisfying $c_0^2 \geq C^2(c_1 + 1)$, and $x \in (c_0/\sqrt{n} E, c_1 R/K)$, we have,

$$P \left(\sup_{f \in \mathcal{F}} I(\tilde{R}_{n,K}(f) \leq R) \frac{M_n(f)}{n} \geq x \right) \leq 2 \exp \left\{ - \frac{nx^2}{C^2(Kx + R)} \right\}.$$

In words, one can conclude that the above inequality shows that for x of the order $1/\sqrt{n}$, the tail probability behaves as $\exp(-nx^2)$, while for large x, it behaves as $\exp(-nx)$.

Specifically, for a single f, we obtain the following corollary.

Corollary 1 *For any* $c_0, c_1 \geq 0$ *satisfying* $c_0^2 \geq C^2(c_1+1)$ *and* $x \in (c_0/\sqrt{n}\sqrt{R}, c_1 R/K)$, *we have*

$$P\left(I(\tilde{R}_{n,K}(f) \leq R)\frac{M_n(f)}{n} \geq x\right) \leq 2\exp\left\{-\frac{nx^2}{C^2(Kx+R)}\right\}. \tag{5}$$

In our proof $L(\psi) - L(\psi_0)$ plays the role of f. Regarding bounding $\tilde{R}_{n,K}(f)$, note also that if $\| f \|_\infty < C$ is uniformly bounded, then $\tilde{R}_{n,K}(f)$ is bounded by a constant depending on C. In our proof for quadratic loss functions we require a bound on $\tilde{R}_{n,K}(f)$ in terms of $\frac{1}{n}\sum_{i=1}^n \bar{P}_{0,z(i)} f$. For that purpose we use the following lemma.

Lemma 1 *Let* $L^0(\hat{\Psi})(\bar{O}(i)) = L(\hat{\Psi}(P_{i-1}))(Z(i), O(i)) - L(\psi_0)(Z(i), O(i))$. *Suppose that with probability 1,* $\sup_{\psi \in \mathbf{\Psi}} \mid L^0(\psi)(Z(i), O(i)) \mid < M_1 < \infty$, *and*

$$\sup_{\psi \in \mathbf{\Psi}} \frac{\bar{P}_{0,z(i)}\{L^0(\psi)\}^2}{\bar{P}_{0,z(i)} L^0(\psi)} \leq M_2 < \infty.$$

Then,

$$\begin{aligned} \tilde{R}_{n,K}(L^0(\hat{\Psi})) &= \frac{2K^2}{n}\sum_{i=1}^n \bar{P}_{0,z(i)}\phi\left(\frac{\mid L^0(\hat{\Psi}(P_{i-1})) \mid}{K}\right) \\ &\leq 2M_2(1/2K^2 + 1/6 M_1 K \exp(M_1/K))\frac{1}{n}\sum_{i=1}^n \bar{P}_{0,z(i)} L^0(\hat{\Psi}(P_{i-1})). \end{aligned}$$

Proof: A third order Taylor expansion for $\exp(x)$ yields $\phi(x) = x^2/2! + \exp(\xi(x))x^3/3!$ for some $\xi(x)$. This can be bounded by

$$x^2(1/2 + 1/6\exp(M_1/K)M_1/K)$$

by using that $\mid x \mid < M_1/K$. As a consequence, we can bound $\bar{P}_{0,z(i)}\phi(\mid L^0(\psi) \mid /K)$ by $(1/2 + 1/6 M_1/K \exp(M_1/K))\bar{P}_{0,z(i)}\{L^0(\psi)\}^2$, which, by assumption, can be bounded by $M_2(1/2 + 1/6 M_1/K \exp(M_1/K))\bar{P}_{0,z(i)} L^0(\psi)$. This proves the lemma.□

D Proof of Theorem 1

For notational convenience, we let $n = (n - n_\ell + 1)$ and let the sum over t_0 run from 1 to n. We have

$$\begin{aligned}
0 \;\leq\; & d_{0n}(\hat{\Psi}_{k_n}, \psi_0) \\
= \;& \frac{1}{n}\sum_{t_0} \bar{P}_{0,z(t_0)}\{L(\hat{\Psi}_{k_n}(P_{t_0-1})) - L(\psi_0)\} \\
& -(1+\delta)\frac{1}{n}\sum_{t_0}\{L(\hat{\Psi}_{k_n}(P_{t_0-1})) - L(\psi_0)\}(O(t_0), Z(t_0)) \\
& +(1+\delta)\frac{1}{n}\sum_{t_0}\{L(\hat{\Psi}_{k_n}(P_{t_0-1})) - L(\psi_0)\}(O(t_0), Z(t_0)) \\
\leq \;& \frac{1}{n}\sum_{t_0} \bar{P}_{0,z(t_0)}\{L(\hat{\Psi}_{k_n}(P_{t_0-1})) - L(\psi_0)\} \\
& -(1+\delta)\frac{1}{n}\sum_{t_0}\{L(\hat{\Psi}_{k_n}(P_{t_0-1})) - L(\psi_0)\}(O(t_0), Z(t_0)) \\
& +(1+\delta)\frac{1}{n}\sum_{t_0}\{L(\hat{\Psi}_{\tilde{k}_n}(P_{t_0-1})) - L(\psi_0)\}(O(t_0), Z(t_0)) \\
= \;& \frac{1}{n}\sum_{t_0} \bar{P}_{0,z(t_0)}\{L(\hat{\Psi}_{k_n}(P_{t_0-1})) - L(\psi_0)\} \\
& -(1+\delta)\frac{1}{n}\sum_{t_0}\{L(\hat{\Psi}_{k_n}(P_{t_0-1})) - L(\psi_0)\} \\
& +(1+\delta)\frac{1}{n}\sum_{t_0}\{L(\hat{\Psi}_{\tilde{k}_n}(P_{t_0-1})) - L(\psi_0)\} \\
& -(1+2\delta)\frac{1}{n}\sum_{t_0} \bar{P}_{0,z(t_0)}\{L(\hat{\Psi}_{\tilde{k}_n}(P_{t_0-1})) - L(\psi_0)\} \\
& +(1+2\delta)\frac{1}{n}\sum_{t_0} \bar{P}_{0,z(t_0)}\{L(\hat{\Psi}_{\tilde{k}_n}(P_{t_0-1})) - L(\psi_0)\}.
\end{aligned}$$

Denote the sum of the first two terms in the last expression by R_{n,k_n} and the sum of the third and fourth term by $T_{n,\tilde{k}_n}$; the last term is the benchmark $(1+2\delta)d_{0n}(\hat{\Psi}_{\tilde{k}_n}, \psi_0)$. Hence, we have

$$0 \leq d_{0n}(\hat{\Psi}_{k_n}, \psi_0) \leq (1+2\delta)d_{0n}(\hat{\Psi}_{\tilde{k}_n}, \psi_0) + R_{n,k_n} + T_{n,\tilde{k}_n} \tag{6}$$

Rewriting $R_{n,k}$ (and $T_{n,k}$) as a martingale: For notational convenience, we

introduce the following notation for the relevant random variables

$$\begin{aligned}\tilde{H}_k &= \frac{1}{n}\sum_{t_0} \bar{P}_{0,z(t_0)}\{L(\hat{\Psi}_k(P_{t_0-1})) - L(\psi_0)\} \\ \bar{H}_k &= \frac{1}{n}\sum_{t_0} \{L(\hat{\Psi}_k(P_{t_0-1})) - L(\psi_0)\}(O(t_0), Z(t_0)),\end{aligned}$$

where, by definition of ψ_0, $\tilde{H}_k \geq 0 \ \forall \ k$. Rewrite $R_{n,k}$ and $T_{n,k}$ as

$$R_{n,k} = (1+\delta)\left[\tilde{H}_k - \bar{H}_k\right] - \delta\tilde{H}_k$$

and

$$T_{n,k} = (1+\delta)\left[\bar{H}_k - \tilde{H}_k\right] - \delta\tilde{H}_k.$$

Approximating $R_{n,k}$ (and $T_{n,k}$) with a negatively deterministically shifted martingale sum, up to a negligible remainder: In order to exploit that a negatively shifted martingale sum has desirable exponential tail behavior, it is important that the random shift $\delta\tilde{H}_k \geq 0$ be replaced by a deterministic shift that is guaranteed to be larger than a constant we can control. We will now utilize assumption A3 for that purpose. For a K, we define

$$\tilde{R}_{n,k} = \frac{2K^2}{n}\sum_{t_0=1}^{n} \bar{P}_{0,z(t_0)}\phi\left(\frac{\mid L^0(\hat{\Psi}_k(P_{t_0-1})) \mid}{K}\right) ,$$

where $\phi(x) = \exp(x) - x - 1$. By Lemma 1, we have

$$\tilde{R}_{n,k} \leq M_2(1/2K^2 + 1/6M_1K\exp(M_1/K))\tilde{H}_k \ .$$

We denote this constant with $C_1(M_1, M_2, K)$ so that $\tilde{R}_{n,k} \leq C_1(M_1, M_2, K)\tilde{H}_k$. Define the event $E_{nk} = \{M_{3n}^{-1} < \tilde{H}_k/E_0\tilde{H}_k < M_{3n}\}$, and let $I_{E_{nk}}$ denote the indicator of this event. By assumption A3, we have $P_0^n(I_{E_{n,k_n}} = 1) \to 1$, and $P_0^n(I_{E_{n,\tilde{k}_n}} = 1) \to 1$, as $n \to \infty$. This also implies that $P_0^n(\tilde{R}_{n,k_n}/E_0\tilde{H}_{k_n} < C_1M_{3n}) \to 1$. For notational convenience, let M_{3n} be redefined by $\max(C_1, 1)M_{3n}$. We decompose $R_{n,k}$ as follows:

$$\begin{aligned}R_{n,k} &= (1+\delta)\left[\tilde{H}_k - \bar{H}_k\right] I_{E_{n,k}} + (1+\delta)\left[\tilde{H}_k - \bar{H}_k\right] I_{E_{n,k}^c} \\ &\quad -\delta\tilde{H}_k I(\tilde{H}_k > M_{3n}^{-1}E_0\tilde{H}_k) - \delta\tilde{H}_k I(\tilde{H}_k < M_{3n}^{-1}E_0\tilde{H}_k) \\ &= R_{n,k}^* + e_{n,k},\end{aligned}$$

where

$$\begin{aligned}R_{n,k}^* &= (1+\delta)\left[\tilde{H}_k - \bar{H}_k\right] I_{E_{n,k}} - \delta\tilde{H}_k I(\tilde{H}_k > M_{3n}^{-1}E_0\tilde{H}_k) \\ e_{n,k} &= (1+\delta)\left[\tilde{H}_k - \bar{H}_k\right] I_{E_{n,k}}^c - \delta\tilde{H}_k I(\tilde{H}_k < M_{3n}^{-1}E_0\tilde{H}_k).\end{aligned}$$

Thus, $R_{n,k_n} = R^*_{n,k_n} + e_{n,k_n}$. By assumption A3 we have $P_0^n(\mid e_{n,k_n} \mid = 0) \to 1$, as $n \to \infty$. Similarly,

$$T_{n,k} = T^*_{n,k} + f_{n,k},$$

where

$$\begin{aligned} T^*_{n,k} &= (1+\delta)\left[\bar{H}_k - \tilde{H}_k\right] I_{E_{n,k}} - \delta \tilde{H}_k I(\tilde{H}_k > M_{3n}^{-1} E_0 \tilde{H}_k) \\ f_{n,k} &= (1+\delta)\left[\bar{H}_k - \tilde{H}_k\right] I_{E_{n,k}} - \delta \tilde{H}_k I(\tilde{H}_k < M_{3n}^{-1} E_0 \tilde{H}_k). \end{aligned}$$

By the same argument as used for e_{n,k_n}, we have $P_0^n(\mid f_{n,\tilde{k}_n} \mid = 0) \to 1$ as $n \to \infty$. Thus, $T_{n,\tilde{k}_n} = T^*_{n,\tilde{k}_n} + f_{n,\tilde{k}_n}$ where $f_{n,\tilde{k}_n}$ equals zero with probability tending to 1.

Let $Z_{n2} = e_{n,k_n} + f_{n,\tilde{k}_n}$ and $Z_{n1} = R^*_{n,k_n} + T^*_{n,\tilde{k}_n}$. We have shown that $d_{0n}(\hat{\Psi}_{k_n}, \psi_0) \leq (1+2\delta) d_{0n}(\hat{\Psi}_{\tilde{k}_n}, \psi_0) + Z_{n1} + Z_{n2}$, where $P_0^n(Z_{n2} = 0) \to 1$ as $n \to \infty$. We now show that

$$EZ_{n1} = ER^*_{n,k_n} + ET^*_{n,\tilde{k}_n} \leq C(M_1, M_2, M_{3n}, \delta)(1 + \log(K(n)))/n$$

for some specified $C(M_1, M_2, M_{3n}, \delta) < \infty$, which completes the proof.

Bounding the tail probability of R^*_{n,k_n}:

Step I: a deterministic negative shift. We define the event $E_{n,k,1} = \{\tilde{H}_k > M_{3n}^{-1} E_0 \tilde{H}_k\}$. Let $s > 0$. We have

$$\begin{aligned} P_0^n(R^*_{n,k_n} > s) &= P_0^n\left(I_{E_{n,k_n}}\{\tilde{H}_{k_n} - \bar{H}_{k_n}\} > \frac{1}{1+\delta}\left\{s + \delta \tilde{H}_{k_n} I_{E_{n,k_n,1}}\right\}\right) \\ &\leq P_0^n\left(I_{E_{n,k_n}}\{\tilde{H}_{k_n} - \bar{H}_{k_n}\} > \frac{1}{1+\delta}\left\{s + \delta M_{3n}^{-1}\, E_0 \tilde{H}_k\Big|_{k=k_n} I_{E_{n,k_n,1}}\right\}\right), \end{aligned}$$

where we used that event $E_{n,k_n,1}$ implies $\tilde{H}_{k_n} \geq M_{3n}^{-1}\, E_0 \tilde{H}_k\Big|_{k=k_n}$, allowing us to replace the random $\tilde{H}_{k_n}$ by this bound that is only random through k_n. We denote the event in the last displayed probability by A_n so that the last displayed bound is denoted $P_0^n(A_n)$. We can write

$$P_0^n(A_n) = P_0^n(A_n \text{ and } I_{E_{n,k_n,1}} = 1\ ) + P_0^n(A_n \text{ and } I_{E_{n,k_n,1}} = 0\ ).$$

Note that if $I_{E_{n,k_n,1}} = 0$, then the right-hand side of the equality equals $\frac{1}{1+\delta} s > 0$, while the left-hand side of inequality in event A_n equals 0. This shows that $P_0^n(A_n \text{ and } I_{E_{n,k_n,1}} = 0) = 0$. This yields the following bound for $P_0^n(R^*_{n,k_n} > s)$:

$$\begin{aligned} &P_0^n(R^*_{n,k_n} > s) \\ &\leq P_0^n\left(I_{E_{n,k_n}}\{\tilde{H}_{k_n} - \bar{H}_{k_n}\} > \tfrac{1}{1+\delta}\left\{s + \delta M_{3n}^{-1}\, E_0 \tilde{H}_k\Big|_{k=k_n}\right\} \text{ and } E_{n,k_n,1} = 1\ \right) \\ &\leq P_0^n\left(I_{E_{n,k_n}}\{\tilde{H}_{k_n} - \bar{H}_{k_n}\} > \tfrac{1}{1+\delta}\left\{s + \delta M_{3n}^{-1}\, E_0 \tilde{H}_k\Big|_{k=k_n}\right\}\right) \\ &\leq K(n) \max_k P_0^n\left(I_{E_{n,k}}\{\tilde{H}_k - \bar{H}_k\} > \tfrac{1}{1+\delta}\left\{s + \delta M_{3n}^{-1} E_0 \tilde{H}_k\right\}\right). \end{aligned}$$

In the last inequality we used that for some collection of random variables $(X(k) : k)$ and constants $(c(k) : k)$ and random index k_n, we have

$$\begin{aligned} P_0^n(X(k_n) < c(k_n)) &\leq P_0^n(X(k) < c(k) \text{ for at least one } k) \\ &\leq \sum_{k=1}^{K(n)} P_0^n(X(k) < c(k)) \\ &\leq K(n) \max_k P_0^n(X(k) < c(k)). \end{aligned}$$

Similarly, for $T^*_{n,\tilde{k}_n}$, we obtain

$$P_0^n(T^*_{n,\tilde{k}_n} > s) \leq K(n) \max_k P_0^n \left(I_{E_{n,k}}\{\bar{H}_k - \tilde{H}_k\} > \frac{1}{1+\delta}\left\{ s + \delta M_{3n}^{-1} E_0 \tilde{H}_k \right\} \right).$$

Step 2: martingale exponential tail probability. We have that $\bar{H}_k - \tilde{H}_k$ equals a martingale sum $\frac{1}{n}\sum_{t_0} Z_{k,t_0} - E(Z_{k,t_0} \mid \bar{O}(t_0 - 1))$ where

$$Z_{k,t_0} = \{L(\hat{\Psi}_k(P_{t_0-1})) - L(\psi_0)\}(\bar{O}(t_0)).$$

By Assumption A1, the random variables Z_{k,t_0} are bounded: $\mid Z_{k,t_0} \mid \leq M_1$ a.s.

We are now ready to apply the Martingale inequality (5) of Theorem 3 to $\tilde{H}_k - \bar{H}_k$ with $R = R_k = M_{3n} E_0 \tilde{H}_k$, for each k separately. Due to this choice of R, we obtain a tail probability at $s > 0$ that behaves for s small as $\exp(-cM_{3n}ns)$ instead of the usual $\exp(-cns^2)$. This in turn proves that the expectation of the remainder terms R^*_{n,k_n} and $T^*_{n,\tilde{k}_n}$ converge at a rate $\log(K(n))/n$ instead of the usual $\log(K(n))/\sqrt{n}$.

For ease of reference, we state the martingale exponential inequality at a k explicitly:

Lemma 2 *Let K be set, and*

$$\tilde{R}_{n,k} = \frac{2K^2}{n} \sum_{i=1}^{n} \bar{P}_{0,z(i)} \phi \left(\frac{\mid L^0(\hat{\Psi}_k(P_{i-1})) \mid}{K} \right),$$

where $\phi(x) = \exp(x) - x - 1$. Let

$$M_{n,k} = \sum_{i=1}^{n} \{ L^0(\hat{\Psi}_k(P_{i-1}))(Z(i), O(i)) - E_0(L^0(\hat{\Psi}_k(P_{i-1})) \mid Z(i)) \} \ .$$

For any R_k, $c_0, c_1 \geq 0$ satisfying $c_0^2 \geq C^2(c_1 + 1)$ and $\alpha \in (c_0/\sqrt{n}\sqrt{R_k}, c_1 R_k/K)$, we have

$$P\left(I(\tilde{R}_{n,k} \leq R_k) \frac{M_{n,k}}{n} \geq \alpha \right) \leq 2\exp\left\{ -\frac{n\alpha^2}{C^2(K\alpha + R_k)} \right\}.$$

In order to apply this inequality to the above tail probability for $I(\tilde{R}_{n,k} < R_k)(\bar{H}_k - \tilde{H}_k)$ with $R_k = M_{3n}E_0\tilde{H}_k$ at a given $\alpha(s) = \frac{1}{1+\delta}(s + \delta M_{3n}^{-1}E_0\tilde{H}_k)$, we need to be able to select c_0, c_1 with $c_0^2 \geq C^2(c_1+1)$ so that

$$\frac{c_0\sqrt{R_k}}{\sqrt{n}} \leq \frac{1}{1+\delta}\left[s + \delta M_{3n}^{-1}E_0\tilde{H}_k\right] \leq c_1\frac{R_k}{K}. \tag{7}$$

Note $M_{3n}^{-1}E_0\tilde{H}_k = M_{3n}^{-2}R_k$, so that we need to apply the inequality at

$$\alpha(s) = 1/(1+\delta)(s + \delta M_{3n}^{-2}R_k).$$

So we now need to select c_0, c_1 so that this $\alpha(s) \in (c_0R_k^{0.5}/n^{0.5}, c_1R_k/K)$. We select $c_0^2 = c_0^2(c_1) = C^2(c_1+1)$. Since the martingale process $\bar{H}_k - \tilde{H}_k$ is bounded by $2M_1$, the upper bound is non-existent if $c_1R_k/K > 2M_1$. This implies the choice $c_1 = c_1(M_1) = 2M_1K/R_k$, thereby guaranteeing that there is no upper bound on $\alpha(s)$ for all s. Let $c_0(M_1) = c_0(c_1(M_1))$ be the corresponding choice for c_0. Thus, for any $\alpha \in (c_0(M_1)R_k^{0.5}n^{-0.5}, \infty)$, we have $P_0^n(I(\tilde{R}_{nk} < R_k)(\bar{H}_k - \tilde{H}_k) > \alpha) \leq 2\exp(-n\alpha^2/\{C^2(K\alpha + R_k)\})$.

The left-inequality $\alpha(s) > c_0(M_1)R_k^{0.5}n^{-0.5}$ is equivalent with

$$s > -\delta M_{3n}^{-2}R_k + C^2(c_1(M_1)+1)n^{-0.5}R_k^{0.5}(1+\delta). \tag{8}$$

The first term on the right-hand side is negative and converges to zero at rate $M_{3n}^{-2}R_k$, while the second term is positive and converges to zero at rate $R_k^{0.5}n^{-0.5}$. By assumption A4, we have

$$\max_k \frac{R_k^{0.5}n^{-0.5}}{M_{3n}^{-2}R_k} \to 0.$$

This implies that for n large enough, we have that the right-hand side of (8) is negative, proving that the inequality $\alpha(s) > c_0(M_1)R_k^{0.5}n^{-0.5}$ holds for all $s > 0$. Thus, there exists an n_1 so that for all $n > n_1$, we have for all $s > 0$,

$$\begin{aligned} P_0^n(I(\tilde{R}_{nk} < R_k)(\bar{H}_k - \tilde{H}_k) > \alpha(s)) &\leq 2\exp(-n\alpha(s)^2/\{C^2(K\alpha(s) + R_k)\}) \\ &= 2\exp\left(-C^{-2}\frac{n}{(1+\delta)^2}\frac{\left(s+\delta M_{3n}^{-2}R_k\right)^2}{R_k + \frac{K}{(1+\delta)}(s+\delta M_{3n}^{-2}R_k)}\right). \end{aligned}$$

Step 3: asymptotic behavior of tail probability. We note that

$$\begin{aligned} \frac{\left(s+\delta M_{3n}^{-2}R_k\right)^2}{R_k + \frac{K}{(1+\delta)}(s+\delta M_{3n}^{-2}R_k)} &= \frac{\left(s+\delta M_{3n}^{-2}R_k\right)}{\frac{R_k}{s+\delta M_{3n}^{-2}R_k} + \frac{K}{(1+\delta)}} \geq \frac{\left(s+\delta M_{3n}^{-2}R_k\right)}{\frac{M_{3n}^2}{\delta} + K} \\ &\geq \frac{s}{\frac{M_{3n}^2}{\delta} + K} \\ &= M_{3n}^{-2}\frac{s}{\delta^{-1} + M_{3n}^{-2}K} \\ &\geq M_{3n}^{-2}\frac{s}{\delta^{-1} + K} \end{aligned}$$

where we use that $M_{3n} > 1$ for all n so that $KM_{3n}^{-2} \leq K$. This shows that, for $s > 0$,

$$P_0^n(R^*_{n,k_n} > s) \leq 2K(n) \exp\left(-\frac{nM_{3n}^{-2}}{c(M_1, M_2, \delta)} s\right),$$

where $c(M_1, M_2, \delta) = 2C^2(1+\delta)^2\,(K + \delta^{-1})$.

Bounding the expectation of R^*_{n,k_n} based on tail probability bounds: Since $ER^*_{n,k_n} \leq \int_0^\infty P_0^n(R^*_{n,k_n} > s)ds$, for each $u > 0$, we have

$$ER^*_{n,k_n} \leq u + \int_u^\infty 2K(n) \exp\left(-\frac{M_{3n}^{-2} n}{c(M_1, M_2, \delta)} s\right) ds.$$

The minimum is attained at $u_n = c(M_1, M_2, \delta)\log(2K(n))/(nM_{3n}^{-2})$ and is given by $c(M_1, M_2, \delta)(\log(2K(n)) + 1)/(nM_{3n}^{-2})$. Thus,

$$ER^*_{n,k_n} \leq c(M_1, M_2, \delta)\frac{1 + \log(2K(n))}{nM_{3n}^{-2}}.$$

Similarly, we obtain his bound for $ET_{n,\tilde{k}_n}$. This proves the theorem under assumption **A4**.

If assumption A4 does not hold: Now consider the case that assumption A4 fails to hold. We have that the leading term $E_0 d_{0n}(\hat{\Psi}_{\tilde{k}_n}, \psi_0) = O(n^{-1}M_{3n}^3)$. First, consider the case that the right-hand side of (8) is negative. In that case, we have our desired inequality for $P_0^n(R^*_{n,k} > s)$ for all $s > 0$ provided above. Consider now the case that the right-hand side of (8) is positive. Then, we have that

$$R_k^{0.5} < (1+\delta)\delta^{-1}C^2(c_1(M_1) + 1)M_{3n}^2 n^{-0.5},$$

which implies that the right-hand side of (8) is bounded by

$$cM_{3n}^2 n^{-1} = (1+\delta)\delta^{-1}C^4(c_1(M_1) + 1)^2 M_{3n}^2 n^{-1}.$$

Thus, in this case, we have the desired exponential bound for $P_0^n(I(\tilde{R}_{nk} < R_k)(\bar{H}_k - \tilde{H}_k) > \alpha(s))$ for any $s > cM_{3n}^2 n^{-1}$ for this specified constant $c > 0$.

We proceed as follows: for any $u > cM_{3n}^2 n^{-1}$, we have

$$\begin{aligned}
E_0 R^*_{n,k_n} &= \int_0^u P_0^n(R^*_{n,k_n} > s)ds + \int_u^\infty P_0^n(R^*_{n,k_n} > s)ds \\
&\leq u + K(n)\max_k \int_u^\infty P_0^n(R^*_{n,k} > s)ds \\
&\leq u + 2K(n)\int_u^\infty \exp\left(-\frac{nM_{3n}^{-2}}{c(M_1, M_2, \delta)} s\right) ds \\
&= u + 2K(n)\frac{c(M_1, M_2, \delta)}{nM_{3n}^{-2}} \exp\left(-\frac{nM_{3n}^{-2}}{c(M_1, M_2, \delta)} u\right).
\end{aligned}$$

The optimal u is given by

$$u^* = \max(cM_{3n}^2 n^{-1}, c(M_1, M_2, \delta)M_{3n}^2 n^{-1} \log(2K(n))).$$

Suppose that $c > c(M_1, M_2, \delta)\{\log 2 + \log K(n)\}$, so that $u^* = cM_{3n}^2 n^{-1}$. Note also that this implies that $K(n) < \exp(cc(M_1, M_2, \delta)^{-1})$. Plugging this u^* in the final expression yields a first term equal to u^* plus a term

$$2K(n)c(M_1, M_2, \delta)^{-1} n^{-1} M_{3n}^2 \exp(-c/c(M_1, M_2, \delta)) \ .$$

Using the bound on $K(n)$ shows that the final expression is $O(M_{3n}^2 n^{-1})$. Suppose now that

$$c < c(M_1, M_2, \delta)\{\log 2 + \log K(n)\}$$

, so that $u^* = c(M_1, M_2, \delta)M_{3n}^2 n^{-1}(\log 2 + \log K(n))$. Plugging this u^* in the final expression now shows that the final expression is $O(M_{3n}^2 n^{-1}(1 + \log K(n))$. Thus, we have shown that in either case, we have that $E_0 R_{n,k_n}^* < C_1 M_{3n}^2 n^{-1}(1 + \log K(n))$ for some universal $C_1 = C_1(M_1, M_2, \delta) < \infty$. The same bounding applies to $E_0 T_{n,\tilde{k}_n}^*$. Thus we have shown that if assumption A4 does not hold, we have

$$d_{0n}(\hat{\Psi}_{k_n}, \psi_0) = o_P(n^{-1} M_{3n}^3) + o_P(n^{-1} M_{3n}^2 (1 + \log K(n))).$$

This completes the proof of Theorem 1.

E Proof of Theorem 2.

This proof is easier than the proof of Theorem 1 because we only have to obtain a rate for the expectation of the relevant martingale processes of $n^{-0.5}$ instead of n^{-1}. Let $\delta_{o(t_0)}$ be the conditional probability distribution of $O(t_0)$, given $Z(t_0)$, that puts

mass 1 on $o(t_0)$. We have

$$\begin{aligned}
0 &\leq d_{0n}(\hat{\Psi}_{k_n}, \psi_0) \\
&= \frac{1}{n}\sum_{t_0} \bar{P}_{0,z(t_0)}\{L(\hat{\Psi}_{k_n}(P_{t_0-1})) - L(\psi_0)\} \\
&= \frac{1}{n}\sum_{t_0} (\bar{P}_{0,z(t_0)} - \delta_{O(t_0)})\{L(\hat{\Psi}_{k_n}(P_{t_0-1})) - L(\psi_0)\} \\
&\quad + \frac{1}{n}\sum_{t_0} \{L(\hat{\Psi}_{k_n}(P_{t_0-1})) - L(\psi_0)\} \\
&\leq \frac{1}{n}\sum_{t_0} (\bar{P}_{0,z(t_0)} - \delta_{O(t_0)})\{L(\hat{\Psi}_{k_n}(P_{t_0-1})) - L(\psi_0)\} \\
&\quad + \frac{1}{n}\sum_{t_0} \{L(\hat{\Psi}_{\tilde{k}_n}(P_{t_0-1})) - L(\psi_0)\} \\
&= \frac{1}{n}\sum_{t_0} (\bar{P}_{0,z(t_0)} - \delta_{O(t_0)})\{L(\hat{\Psi}_{k_n}(P_{t_0-1})) - L(\psi_0)\} \\
&\quad + \frac{1}{n}\sum_{t_0} (\delta_{O(t_0)} - \bar{P}_{0,z(t_0)})\{L(\hat{\Psi}_{\tilde{k}_n}(P_{t_0-1})) - L(\psi_0)\} \\
&\quad + \frac{1}{n}\sum_{t_0} \bar{P}_{0,z(t_0)}\{L(\hat{\Psi}_{\tilde{k}_n}(P_{t_0-1})) - L(\psi_0)\}
\end{aligned}$$

.

Denote the first term with R_{n,k_n} and the second term with $T_{n,\tilde{k}_n}$. The third term is $d_{0n}(\hat{\Psi}_{\tilde{k}_n}, \psi_0)$. Thus, we have

$$0 \leq d_{0n}(\hat{\Psi}_{k_n}, \psi_0) \leq d_{0n}(\hat{\Psi}_{\tilde{k}_n}, \psi_0) + R_{n,k_n} + T_{n,\tilde{k}_n}.$$

As in the proof of Theorem 1, we have $P_0^n(R_{n,k_n} > s) \leq K(n) \max_k P_0^n(R_{n,k} > s)$, and similarly for T_{n,k_n}. We apply Lemma 2 to $R_{n,k}$.

Lemma 3 *Let K be set, and*

$$\tilde{R}_{n,k} = \frac{2K^2}{n}\sum_{i=1}^{n} \bar{P}_{0,z(i)}\phi\left(\frac{\mid L^0(\hat{\Psi}_k(P_{i-1})) \mid}{K}\right),$$

where $\phi(x) = \exp(x) - x - 1$. Let

$$M_{n,k} = \sum_{i=1}^{n}\{L^0(\hat{\Psi}_k(P_{i-1}))(Z(i), O(i)) - \bar{P}_{0,z(i)}L^0(\hat{\Psi}_k(P_{i-1}))\} \ .$$

For any R_k, $c_0, c_1 \geq 0$ satisfying $c_0^2 \geq C^2(c_1+1)$ and $\alpha \in (c_0/\sqrt{n}\sqrt{R_k}, c_1 R_k/K)$, we have

$$P_0^n\left(I(\tilde{R}_{n,k} \leq R_k)\frac{M_{n,k}}{n} \geq \alpha\right) \leq 2\exp\left\{-\frac{n\alpha^2}{C^2(K\alpha + R_k)}\right\}.$$

First, we note that $R_{n,k} \leq R = 2K^2\phi(M_1/K)$. Thus, we can apply Lemma 3 with R_k set equal to this R in which case $I(\tilde{R}_{n,k} \leq R) = 1$ with probability 1. This proves that for a certain C (e.g., $C = 100$) for any $c_0, c_1 \geq 0$, $c_0^2 \geq C^2(c_1 + 1)$ and $s \in (c_0/\sqrt{n}\sqrt{R}, c_1 R/K)$, we have

$$P_0^n(R_{n,k} > s) \leq 2\exp\left(-\frac{ns^2}{C^2(Ks + R)}\right).$$

We can replace s by $c_1 R/K$ to obtain the bound:

$$P_0^n(R_{n,k} > s) \leq 2\exp\left(-\frac{ns^2}{C^2 R(c_1 + 1)}\right).$$

We select $c_0^2 = C^2(c_1 + 1)$. Lets denote this choice with $c_0(c_1)$. We can still select c_1 as large as we want. Thus for any given c_1, we have that for $s \in (c_0 R^{0.5}/n^{0.5}, c_1 R/K)$, we have the above tail probability $2\exp(-ns^2/(C^2(c_1 + 1)R)$. Since the loss function is bounded by M_1, we have that $R_{n,k} < 2M_1$. So we only need our upper bound $c_1 R/K$ to be larger or equal than $2M_1$. Therefore, we select $c_1 R/K = 2M_1$, and thus $c_1 = c_1(M_1) = 2M_1 K/R$. This implies the choice $c_0(M_1) = C^2(c_1(M_1) + 1)$ for c_0. So we have shown that for all $s > c_0(M_1)R^{0.5}n^{-0.5}$, we have the desired

$$P_0^n(R_{n,k} > x) < 2\exp\left(-\frac{ns^2}{C^2 R(c_1(M_1) + 1)}\right).$$

Thus, we have shown that for all $x > c_0(M_1)R^{0.5}/n^{0.5}$

$$P_0^n(R_{n,k_n} > s) \leq 2K(n)\exp\left(-\frac{ns^2}{C^2 R(c_1(M_1) + 1)}\right).$$

For notational convenience, let $c_1 = c_0(M_1)R^{0.5}$ and $c_2 = C^{-2}R^{-1}(c_1(M_1) + 1)^{-1}$. We have for any $u > c_1/n^{0.5}$

$$\begin{aligned} ER_{n,k_n} &\leq \int_0^\infty P_0^n(R_{n,k_n} > s)ds \\ &= \int_0^{c_1/n^{0.5}} P_0^n(R_n > s)ds + \int_{c_1/n^{0.5}}^u P_0^n(R_n > s)ds + \int_u^\infty P_0^n(R_n > s)ds \\ &\leq u + \int_u^\infty 2K(n)\exp(-c_2 ns^2). \end{aligned}$$

The minimum over u is given by $u^* = \max(c_1 n^{-0.5}, c_2^{-0.5}n^{-0.5}(\log 2 + \log K(n)))$. The resulting value at u^* can be bounded by a constant times u^*, giving us the bound $C(M_1)(1 + \log K(n))/n^{0.5}$ for some $C(M_1)$. Similarly, we obtain $ET_{n,\tilde{k}_n} \leq C(M_1)(1 + \log K(n))/n^{0.5}$. This completes the proof of the theorem. □

F Online-cross-validation selector for independent identically distributed observations

In this section we consider the case that the observations are i.i.d. so that $\bar{P}_{0,z} = \bar{P}_0$ is a common probability distribution in time t that does not depend on summary measures of an observed past. In this case, we define a cross-validated risk that averages across different orderings, thereby potentially enhancing the precision of the corresponding cross-validation selector.

F.1 Online cross-validation selector

Consider an initial ordering $O(1), \ldots, O(n)$. A new ordering $O(\pi(1)),\ldots,O(\pi(n))$ is defined by a permutation $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$ that is 1-1 and onto. Consider V such permutations $\pi_1, \ldots, \pi_V$. Let $\hat{\Psi}_k$ be candidate estimators that can be applied to data sets $O(i)$ for i ranging over a subset of $\{1, \ldots, n\}$, $k = 1, \ldots, K(n)$. Let P_{v,t_0} be the empirical distribution based on $O(\pi_v(1)), \ldots, O(\pi_v(t_0))$. Given a candidate estimator $\hat{\Psi}_k$ we define its online cross-validated risk as follows:

$$R_{CV,n}(\hat{\Psi}_k) = \frac{1}{V} \sum_{v=1}^{V} \frac{1}{n - n_\ell + 1} \sum_{t_0=n_\ell+1}^{n} L(\hat{\Psi}_k(P_{v,t_0-1}))(O(\pi_v(t_0))).$$

The corresponding online cross-validation selector is defined as

$$k_n = \arg\min_k R_{CV,n}(\hat{\Psi}_k) \ .$$

The online super learner is defined as follows:

$$\hat{\Psi}(P_t) = \hat{\Psi}_{k_t}(P_t), \ t = n_\ell, \ldots, n \ \ .$$

Thus, at t observations, the online super learner again uses the estimator $\hat{\Psi}_k$ with index $k = k_t$, indicating the lowest cross-validated risk after t steps.

Suppose that we partition the n observations in V subgroups of observations and let the permutation π_v be defined by an ordering of the n observations for which the last n/V observations belong to the v-th subgroup, $v = 1, \ldots, V$. We could also define $n_\ell = n(1 - p)$ for $p = 1/V$. In this case, computing the online cross-validated risk involves training the estimators on the v-specific training samples of size at least $n(1 - p)$ and evaluating the performance of each observation in the corresponding v-specific validation sample. This is performed across each of the V orderings, which makes the online cross-validated risk similar to the cross-validated risk for the usual V-fold cross-validation.

The online cross-validated risk estimates the online cross-validated true risk,

$$\tilde{R}_{CV,n}(\hat{\Psi}_k) = \frac{1}{V} \sum_{v=1}^{V} \frac{1}{n - n_\ell + 1} \sum_{t_0=n_\ell+1}^{n} \bar{P}_0 L(\hat{\Psi}_k(P_{v,t_0-1})) \ ,$$

which is minimized by ψ_0. The difference between $\tilde{R}_{CV,n}(\hat{\Psi}_k)$ and $\tilde{R}_{CV,n}(\psi_0)$ defines a loss-based dissimilarity between a candidate estimator $\hat{\Psi}_k$ and the true target ψ_0:

$$d_{0n}(\hat{\Psi}_k, \psi_0) = \frac{1}{V}\sum_{v=1}^{V} \frac{1}{n - n_\ell + 1} \sum_{t_0=n_\ell+1}^{n} \bar{P}_0\{L(\hat{\Psi}_k(P_{v,t_0-1})) - L(\psi_0)\} \ .$$

The online oracle selector is defined as the minimizer of this loss-based dissimilarity,

$$\tilde{k}_n = \arg\min_k d_{0n}(\hat{\Psi}_k, \psi_0) \ .$$

Below is the precise statement of the oracle inequality in this setting.

Theorem 4 *Consider the above model $\mathcal{M}$ for the distribution $\bar{P}_0$ of O in which $O(i) \sim_{iid} \bar{P}_0$ and the target parameter $\Psi : \mathcal{M} \to \mathbf{\Psi}$ defined as the minimizer of a loss function $L(\psi)$, with $\psi_0 = \Psi(\bar{P}_0)$. Consider also the above defined online cross-validation selector k_n, online oracle selector $\tilde{k}_n$, and $d_{0n}(\hat{\Psi}_k, \psi_0)$ for i.i.d. data defined in terms of an average over V permutations of O^n.*

Assumptions

A6. *There exists an $M_1 < \infty$ so that*

$$\sup_{\psi \in \mathbf{\Psi}} \sup_{i, O(i)} \mid L(\psi)(O(i)) - L(\psi_0)(O(i)) \mid \leq M_1,$$

where the supremum over $O(i)$ is taken over a support of the distribution $\bar{P}_0$.

A7. *There exists an $M_2 < \infty$ so that with probability 1*

$$\sup_{\psi \in \mathbf{\Psi}} \frac{\bar{P}_0\{L(\psi) - L(\psi_0)\}^2}{\bar{P}_0\{L(\psi) - L(\psi_0)} \leq M_2 < \infty.$$

A8. *There exists a sequence $M_{3n} < \infty$ (e.g., $M_{3n} = \log n$) so that with probability tending to 1,*

$$\frac{1}{M_{3n}} < \frac{d_{0n}(\hat{\Psi}_k, \psi_0)}{E_0 d_{0n}(\hat{\Psi}_k, \psi_0)} < M_{3n} \text{ for all } k = 1, \ldots, K(n).$$

A9.

$$n M_{3n}^{-3} \min_k E_0 d_{0n}(\hat{\Psi}_k, \psi_0) \to \infty \text{ as } n \to \infty.$$

Under assumptions **A6-A9**, *for any $\delta > 0$, there exists a constant $C(\delta, M_1, M_2) < \infty$ universal in n and choice of candidate estimators such that*

$$d_{0n}(\hat{\Psi}_{k_n}, \psi_0) \ \leq \ (1 + 2\delta) d_{0n}(\hat{\Psi}_{\tilde{k}_n}, \psi_0) + Z_n,$$

where $Z_n = Z_{n1} + Z_{n2}$ *with* $P_0^n(Z_{n2} = 0) \to 1$ *as* $n \to \infty$, *and for* $n > n_1$ *for some* $n_1 < \infty$, *we have*

$$E_0 Z_{n1} \leq C(\delta, M_1, M_2) \frac{M_{3n}^2(1 + \log(K(n)))}{n}.$$

If Assumption **A9** *does not hold, then*

$$d_{0n}(\hat{\Psi}_{k_n}, \psi_0) = o_P(n^{-1} M_{3n}^3) + o_P(n^{-1} M_{3n}^2(1 + \log K(n))).$$

The proof of this theorem is analogous to the proof of Theorem 1 and is therefore omitted. The only new observation is that the terms $R_{n,k}, T_{n,k}$ now involve an average over V terms $R_{n,k,v}, T_{n,k,v}$, $v = 1, \ldots, V$ and we can apply the same proof to $R_{n,k,v}$ for each v. □