Challenges in statistical inference for complicated datasets structures

Thesis submitted for the degree of
"Doctor of Philosophy"
by

Jonathan Yefenof

Submitted to the Senate of the Hebrew University
October 2019

This work was carried out under the supervision of
Professor Ya'acov Ritov
and Professor Yair Goldberg

Acknowledgements

I would like to thank very much my advisors Ya'acov Ritov and Yair Goldberg.

I would like to thank Ya'acov Ritov for all the meetings that we had, the nice chats and to the tremendous knowledge he has in statistics. I'm very thankful for the nice and elegant solutions that he gave to the theoretical questions that I had.

I would like to thank Yair Goldberg for the endless help and support he gave me through this journey. It was always nice to meet him no matter where it was. He taught me how to love statistics. Yair probably don't remember this, but when I was a first year student I was curious to know how to calculate the variance of a geometric distribution, I went to a class room where teaching assistants help undergraduate students. When I told the people there what is the problem I want to solve, they introduced me to Yair and told me that he is the man I should ask these type of questions. And yes, Yair explained me so well and politely how to calculate the variance using a trick of looking at the term as a derivative of another term. I feel that this manner of explaining is remaining in him even if it's related to heavy theoretical issues.

I would also like to thank Avishai Mandelbaum for all his help. It is nice to keep in touch with him and discuss about common theoretical concepts.

I would also like to thank David Zucker and Pavel Chigansky for all their help and support.

I'm very happy that throughout my journey I visited seminars and conferences I feel that the community of statisticians in Israel is very warm and I'd like to thank all the colleagues that I met in these places. I'd like to thank my dear fellows who studied with me, Saar Gershon, Yoni Sidi, Daniel Nevo, Ariel Mantzura, Uri Isserles, Osama Eleimy, Sarit Agami, Danielle Fogel-Afterman, Naomi Kaplan-Damary, Bella Vakulenko-Lagun and Yael Travis-Lumer. I'd like to thank Clare Pagis, Orel Levy and Tali Alcalay for their great administrative help.

A big thanks is to dear colleagues at the QBI statistics unit, Daniel Rothenstein, Nir Sharon, Yuval Mathews and Eyal Shalom. I joined a great place where many interesting statistical questions have arisen.

Last thanks is to my dear family, my dear wife Efrat and our lovely children, Matan, Assaf and Ella. I'd like to thank my dear sister Einat and her family. A huge thanks is to my dear and wonderful parents.

Abstract

In recent years, many datasets which need to be analyzed are complicated. Complicated data structures may arise when the assumptions of the probabilistic structure of the observations are very general or when some of the observations are not complete. The goal is to develop statistical theory to overcome these difficulties and to develop a theory that will ensure reasonable inference.

A common model which assumes a very general probabilistic structure is statistical classification. In statistical classification, the explaining variable is multi-dimensional and the dependent variable is binary. The goal is to construct a classifier, which is a function that classifies a new explanatory point to one of the two values of the dependent variable. The success of a classifier is commonly measured by its test error which is the probability of misclassification. Hence, the test error is an important quantity that measures the success of the classifier. However, inference for the test error is difficult due to the generality of the model; therefore, even the rate of convergence is unclear. In chapter 2 we propose inference for the test error which is based on confidence intervals (CIs). We construct two types of CIs—one is called a naive CI. The naive CI has a relatively simple structure. The other type is called an adaptive CI. The construction of the adaptive CI is more complicated and requires gentle arguments in the theory of empirical processes. Our proposed CIs are based on two common methods—normal approximation and empirical bootstrap.

Another form of complicated data structure is when the data is not complete. Such a structure is common in survival analysis. In Chapter 3, we consider survival analysis data. This data arises from screening a medical condition and considers patients who wait in an emergency department (ED) queue. The observations are distinguished by three categories. Some patients lose their patience and decide to quit the queue. If a patient notifies the system that he is leaving, then an exact record of his patience time is observed; if a patient enters the emergency department, then only a lower bound of his patience time is observed; if a patient decides the leave the queue but is not notifying the system, then eventually he will be called to the ED and hence only an upper bound for his

patience time will be observed. In this scenario, the exact patience time is recorded only among patients who leave the queue and report it. Despite the fact that for a large portion of the patients, the patience time is not recorded, in Chapter 3, we propose parametric and nonparametric estimation methods for the patience time distribution.

Contents

1. Introduction

1.1 The Motivation

The goal of statistics is to analyze data. In recent years, many datasets which need to be analyzed are complicated. Complicated data structures may arise when handling big data, i.e., when the number of explaining variables in the dataset is very large, or when the number of observations is very large. Complicated data structures can also arise in settings in which some of the observations are not complete. Specifically, exact information may not be recorded due to missing data, censored data, or when only an indicator whether an event happened or not at some point in time is given, where the last type of data is referred to a current status data. Complicated data structures can be found in biostatistics, bioinformatics, queuing systems, clinical trails, and many more. These datasets need to be analyzed carefully, which results in the need to develop advanced statistical tools to address the analysis challenges. Models that are common in analyzing complicated data structures include machine-learning and nonparametric approaches.

In his seminal paper, Breiman (2001) describes two common cultures in statistical modeling—the data-modeling culture and the algorithmic-modeling culture. These two cultures are widely used by analysts of complicated datasets. One culture assumes that the data are generated by a given stochastic data model; this culture is common in regression and survival analysis. The other culture uses algorithmic models and treats the data mechanism as unknown; this culture is common in classification. In the data-modeling culture, the analysis starts with assuming a stochastic data model, with parameters of unknown value. The values of the parameters are estimated from the data, and the model is then used for explanation and prediction. In the algorithmic-modeling culture, the analyst considers the data-generating mechanism as a black box and treats it as unknown. Instead, the analyst looks for an algorithm, described by a function $f(X)$, of the explanatory variable X, which can predict well the response Y.

The algorithmic modeling culture is common when addressing classification problems. Here, a classification problem is when there is a vector of explanatory variables X which

can be continuous, and a response variable Y which is categorical. When the response can take only two values, the problem is referred to as a binary classification problem. In this problem, the goal of the analyst is to use the observed data in order to construct a function that will classify a new vector of explanatory variables X to one of the two possible values of the response variable Y. The obtained function is called a classifier. A well-known family of classifiers are the kernel-machines algorithms. Kernel-machine algorithms were introduced by Vapnik (1998), who referred to these algorithms as support vector machines. More general kernel machines are discussed in Steinwart and Christmann (2008) and Hofmann et al. (2008). Kernel-machine classifiers can be implemented easily, have strong generalization abilities, and converge, under some mild conditions, to the optimal classifier (Steinwart and Christmann, 2008). Here, the optimal classifier, also known as the Bayes rule, is the classifier that minimizes some asymptotic optimization problems.

The test error of a given classifier is defined as the probability of missclassification, i.e., the probability that the classifier fails to classify correctly. For example, consider a dataset of patients that consists of explaining measurements such as weight, height, blood pressure, and age, and a binary response variable that indicates whether the patient is diagnosed with diabetes or not. Given such a dataset, the goal is to construct a classifier that for any given profile, i.e., a set of the explaining variables, will result in a prediction of the diagnosis of having diabetes or not. A missclassification occurs when the classifier classifies a profile as diagnosed where the patient does not have diabetes, or when the classifier classifies a profile as not diagnosed where the patient has diabetes. Since the test error is defined as the probability of missclassification, a small test error indicated that the classifier classifies well. Therefore, the test error is considered as an important quantity in classification problems, and there is a strong interest in estimating correctly the test error. A discussion of the problem appears in Chapter 2. Chapter 2 provides a comprehensive analysis of how to construct confidence intervals for the test error in kernel-machine algorithms.

So far, a classification problem was discussed that is typically handled using the algorithmic-culture tools. We now move to discuss a survival analysis problem that is typically handled using the data-modeling culture tools. Specifically, consider a setting in which there is an interest of estimating the distribution of a time-to-event variable, but this variable is not completely observed. For example, some of the observations might be right censored. Here, right censoring means that some of the time-to-event observations have only a lower bound on the event time. It is common to refer to this event as a failure event and to its time as the failure time. In the standard right-censored setting,

for each patient, one observes either the failure time or the censoring time. Another kind of survival data format is left-censored data. For this type of data format, for some of the observations, the failure time is not given, and instead only an upper bound is given.

We consider survival data that combine three types of observations: uncensored, right-censored, and left-censored data. Our motivation example comes from the need to estimate the patience of patients who arrive at an emergency department and wait for treatment. Three categories of patients are observed: those who leave the system and announce it, and thus their patience time is observed; those who get service and thus their patience time is right-censored by the waiting time; and those who leave the system without announcing it. For this third category, the patients' absence is revealed only when they are called to service, which is after they have already left; formally, their patience time is left-censored. Chapter 3 proposes both a semiparametric and a nonparametric novel estimators for the patience-time distribution, as well as theoretical and numerical analysis of these estimators. This chapter also presents the analysis of our motivating example using the two proposed estimators.

1.2 Confidence intervals for the test error

In Chapter 2 we study the properties of the test error for kernel machines in the setting of binary classification. In a binary classification setting, one observes a dataset that consists of multi-dimensional explanatory variables and a binary response. A classifier is a function that labels a new vector of explanatory variables to one of the two response values. For a given classifier, the test error is the probability of missclassification. Therefore, the test error measures the classifier successes, and hence it is important to estimate its value. Due to its structure, it is difficult to provide a point estimator for the test error; instead it is common to construct a confidence interval for this quantity**?**.

More specifically, consider the following statistical classification problem where the observed data consist of $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ such that the explanatory variable X is multi-dimensional and typically continuous, while the response variable Y gets the values -1 and 1. A classifier $\widehat{f}$ is a function which for every explanatory variable X returns a real value $\widehat{f}(X)$, classified according to its sign. It follows that the event of missclassification is $\{Y\widehat{f}(X) \leq 0\}$. By definition, a kernel-machine classifier is the minimizer of a regularized empirical risk over a reproducing kernel Hilbert space (RKHS) of functions. The size of the RKHS affects whether the method is parametric or nonparametric. Finite-dimensional functional spaces results in a parametric setting, whereas infinite-dimensional spaces results in a nonparametric setting. Here, we allow these kernel machines to be defined with

respect to an infinite-dimensional RKHS.

The test error of a classifier $\widehat{f}$ is the probability that the function missclassifies. Formally, it is defined as the expectation of the indicator $1\{Y\widehat{f}(X) \leq 0\}$ given the sample $\{(X_1, Y_1), ..., (X_n, Y_n)\}$. Estimation methods of the test error are usually based on resampling techniques, such as leave-one-out or k-fold cross-validation. Previous works have shown that test error estimators are plagued by bias and high variance across training sets and consequently the test error is accepted to be a difficult quantity to estimate (Isaksson et al., 2008; Zhang, 1995; Hastie et al., 2009). The reason for this problematic behavior is that the test error is the expectation of the non-smooth function $1\{Y\widehat{f}(X) \leq 0\}$.

In Chapter 2 we propose two methods for constructing confidence intervals for the test error in general infinite-dimensional RKHS settings. The idea is to replace the indicator function $1\{t \leq 0\}$ by smooth functions. For the first method we choose functions that approximate $1\{t \leq 0\}$ from above and below, such that their empirical mean over $Y\widehat{f}(X)$ converges with a root-n rate to Gaussian variables. This allows us to construct conservative confidence intervals using both Gaussian approximation and empirical bootstrap. While this construction is simple, and results in asymptotically conservative confidence intervals, the length of these intervals may not converge to zero with sample size.

To overcome this problem, we propose a second method for constructing confidence intervals. In this method, we replace the fixed functions that approximate $1\{t \leq 0\}$ by two sequences of functions that converge to the indicator function $1\{t \leq 0\}$ from below and from above. Based on the setting proposed by Hable (2012), and using empirical process theory (Kosorok, 2008; van der Vaart, 2000), we are able to construct conservative confidence intervals with length converging to zero, using both Gaussian approximation and empirical bootstrap.

1.3 Self-reporting and screening

In Chapter 3 we study the estimation of failure time distribution where failure times can be either observed directly, right-censored, or left-censored. The motivating example is estimating the patience of patients who wait for treatment in an emergency department (ED). Each patient has its (virtual) waiting time for service and patience time. Here, the waiting time is the time passed until the patient is called to enter for treatment, while the patience time is the time passed in the queue until the patient loses patience and decides to quit the queue. Estimating the patience-time distribution is of importance, as the decision of patients to leave the system before being served may have a strong effect on their physical well-being.

The scenario described above leads to three categories of patients. The first category consists of patients who get service. For a patient in this category, the waiting time is given while only a lower bound on patience time is given, namely the waiting time. Therefore, in this category, the patience time is right-censored by the waiting time while the waiting time is fully observed. The second category comprises those who leave the system and announce it. For patients in this category, their patience time is given, while their waiting time is known to exceed the patience time. Thus, their patience time is observed while the waiting time is right-censored. The third category consists of patients who leave the system without announcing it; their absence is hence revealed only when they are called to service, which is after they have already left. In this case their waiting time is observed, while only an upper bound on their patience time is given, i.e., it is left-censored.

Summarizing, in the ED setting, the patience time is either exactly observed for the second category; or right-censored for the first category; or left-censored for the third category. Since the patience time may follow three types of observations, estimating it is challenging. On the other hand, the waiting time is either exactly observed or right-censored; hence, estimating it can be done using classical survival analysis methodology.

In Chapter 3 we propose novel parametric and nonparametric estimators for the distribution function of the patience time using this three-category survival data. We then study their rates of convergence. The parametric estimator is based on both full and partial likelihoods. We provide conditions under which the parametric estimator is a linear asymptotic normal (LAN) estimator and converges to a normal distribution in a root-n rate. The nonparametric estimator is based on nonparametric kernel estimators for density functions and on a novel estimator of the cumulative probability function that has some similarities to the Nelson–Aalen estimator (e.g., Klein and Moeschberger, 2013, Chapter 4). We show, under some regularity conditions, that the nonparametric estimator point-wise converges to the normal distribution. We study, using simulation, the properties of both the parametric and nonparametric estimators. We then carry out a case study that is based on data of patients waiting for treatment in an ED, in the U.S. in 2008. We analyzed separately different severity levels. We conclude with a comparison of the parametric and nonparametric estimators for the three different severity levels of this dataset.

References

Leo Breiman. Statistical modeling: The two cultures). *Statistical science*, 16(3):199–231, 2001.

Robert Hable. Asymptotic normality of support vector machine variants and other regularized kernel methods. *Journal of Multivariate Analysis*, 106:92–117, 2012.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Overview of supervised learning. In *The elements of statistical learning*, pages 9–41. Springer, 2009.

Thomas Hofmann, Bernhard Schölkopf, and Alexander J Smola. Kernel methods in machine learning. *The annals of statistics*, pages 1171–1220, 2008.

Anders Isaksson, Mikael Wallman, Hanna Göransson, and Mats G Gustafsson. Cross-validation and bootstrapping are unreliable in small sample classification. *Pattern Recognition Letters*, 29(14):1960–1965, 2008.

J. P. Klein and M. L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, New York, NY u.a., October 2013.

M. R. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Springer, New York, 2008.

Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000.

V. N. Vapnik. *Statistical learning theory*, volume 1. Wiley New York, 1998.

Ping Zhang. Assessing prediction error in non-parametric regression. *Scandinavian journal of statistics*, pages 83–94, 1995.

2. Confidence intervals for the test error

Status: This paper was submitted to the Journal of Machine Learning Research

CONFIDENCE INTERVALS FOR THE TEST ERROR IN A GENERAL KERNEL MACHINE CLASSIFICATION

BY JONATHAN YEFENOF[1] YAIR GOLDBERG[2] AND YA'ACOV RITOV[1,3]

[1]*The Hebrew University of Jerusalem*

[2]*Technion - Israel Institute of Technology*

[3]*University of Michigan*

In statistical learning, the successes of classifiers is commonly measured by the test error which is the misclassification probability. Therefore, it is of an interest to construct a high quality estimator for the test error. In this paper we consider the test error of general kernel-machine classifiers. Inference for kernel-machine classifiers, and more specifically, for the test error of these classifiers, is difficult since even the rate of convergence is unclear. We propose confidence intervals which are asymptotically correct. The proposed confidence intervals are constructed by two different approaches. The first approach is based on converging to a normal distribution approximation and the second is based on empirical bootstrap.

1. Introduction. Consider the following statistical classification problem. We observe $\{(X_1, Y_1), ..., (X_n, Y_n)\}$ such that the explanatory variable X is multidimensional and typically continuous, while the response variable Y gets the values -1 and 1. A classifier $\widehat{f}$ is a function wherein every explanatory variable X returns a real value $\widehat{f}(X)$ which is classified according to its sign. It follows that the event of misclassification is $\{Y\widehat{f}(X) \leq 0\}$. In this work we consider classifiers from the kernel-machine family which are a family of classification algorithms (Vapnik, 1998; Steinwart and Christmann, 2008). Kernel-machine classifiers can be implemented easily, have a strong generalization ability, and converge, under some conditions, to the optimal classifier (Steinwart and Christmann, 2008). By definition, a kernel-machine classifier is the minimizer of a regularized empirical risk over a reproducing kernel Hilbert space (RKHS) of functions. The size of the RKHS affects whether the method is parametric or nonparametric. Finite-dimensional functional spaces results in a parametric setting whereas infinite-dimensional spaces results in a

Keywords and phrases: Classification, Kernel-machine classifier, Reproducing kernel Hilbert space, Empirical processes, Confidence intervals, Gaussian process, Bootstrap

nonparametric setting.

The test error of a classifier $\widehat{f}$ is the probability that the function misclassifies. Formally, it is defined as the expectation of the indicator $\mathbf{1}\left\{Y\widehat{f}(X) \leq 0\right\}$ given the sample $\{(X_1, Y_1), ..., (X_n, Y_n)\}$. Estimation methods of the test error are usually based on resampling techniques such as leave-one-out or k-fold cross-validation. Efron (1983) suggested bootstrap estimates of the test error and later refinements were given by Efron and Tibshirani (1995, 1997). There have been a number of simulation studies comparing these approaches; some references include Efron (1983); Chernick et al. (1985); Kohavi (1995); Krzanowski and Hand (1997). A nice survey of estimators is given by Schiavo and Hand (2000). Previous works have shown that test error estimators are plagued by bias and high variance across training sets and consequently the test error is accepted as a difficult quantity to estimate (Isaksson et al., 2008; Zhang, 1995; Hastie et al., 2009). The reason for this problematic behavior is that the test error is the expectation of the non-smooth function $\mathbf{1}\left\{Y\widehat{f}(X) \leq 0\right\}$.

Instead of point estimation, one can consider a confidence interval. Previous works on confidence intervals for the test error considered only parametric settings. Laber and Murphy (2012) propose confidence intervals for linear classifiers. Jiang et al. (2008) propose confidence intervals for finite-dimensional RKHSs. Both use the fact that in their settings the derived classifiers have the property that the empirical mean of $\mathbf{1}\left\{Y\widehat{f}(X) \leq 0\right\}$ converges with a root-n rate. In a general infinite-dimensional RKHS, the empirical mean of $\mathbf{1}\left\{Y\widehat{f}(X) \leq 0\right\}$ may not converge with a root-n rate and therefore similar techniques to those used by Laber and Murphy (2012) and Jiang et al. (2008) cannot be applied.

In this work we propose two methods to construct confidence intervals for the test error for general infinite-dimensional RKHS settings. The idea is to replace the indicator function $\mathbf{1}\{t \leq 0\}$ by smooth functions h^- and h^+ such that $h^-(t) \leq \mathbf{1}\{t \leq 0\} \leq h^+(t)$. For the first method, we choose h^- and h^+ such that the empirical mean of $h^-(Y\widehat{f}(X))$ and $h^+(Y\widehat{f}(X))$ converge with a root-n rate to Gaussian variables. This allows us to construct conservative confidence intervals using both Gaussian approximation and empirical bootstrap. While this construction is simple, and results in asymptotically conservative confidence intervals, the length of these confidence intervals may not converge to zero when the sample size grows.

To overcome this problem, we propose a second method for constructing confidence intervals. In this method, we replace the fixed functions h^- and h^+ with sequences of functions $\{h_n^-\}$ and

$\{h_n^+\}$ such that $\{h_n^-\}$ is a pointwise increasing sequence that converges to the indicator function $1_{t\leq 0}$ from below, and $\{h_n^+\}$ is a pointwise decreasing sequence defined similarly. However, since the classifier $\widehat{f}$ and the sequences $\{h_n^-\}$ and $\{h_n^+\}$ are changing with n, it is challenging to find conditions ensuring that the empirical means of $h_n^-(Y\widehat{f}(X))$ and $h_n^+(Y\widehat{f}(X))$ converge with a root-n rate to Gaussian variables. Based on the setting proposed by Hable (2012), and using empirical process theory (Kosorok, 2008; van der Vaart, 2000), we develop conditions which ensure such a convergence. These conditions enable the construction of conservative confidence intervals with a length converging to zero, using both Gaussian approximation and empirical bootstrap.

The paper is organized as follows. In Section 2 we present kernel-machines classifiers. Confidence intervals based on the first method are presented in Section 3. Confidence intervals based on the second method are presented in Section 4. Simulation studies are given in Section 5. Concluding remarks are given in Section 6.

2. Preliminaries. Assume that the observed data $D := \{(X_1, Y_1), ..., (X_n, Y_n)\}$, are i.i.d. following a common distribution P where the explaining variable $X \in \mathcal{X}$, $\mathcal{X} \subset \mathbb{R}^d$ is compact, the response Y gets the values $\{-1, 1\}$. The data set D is used in order to construct a classifier which is a function $\widehat{f} : \mathcal{X} \longmapsto \mathbb{R}$, such that $\widehat{f}(x)$ classifies a new explanatory value according to $\operatorname{sign}\widehat{f}(x)$ where $\operatorname{sign}\widehat{f}(x) = 1$ if $\widehat{f}(x) \geq 0$. In this case x is classified to 1 and $\operatorname{sign}\widehat{f}(x) = -1$ if $\widehat{f}(x) < 0$; in this case x is classified to -1.

The test error of a classifier $\widehat{f}$ is defined by $\tau(\widehat{f}) \equiv P\mathbf{1}\Big\{\operatorname{sign}\widehat{f}(X) \neq Y\Big\}$, where, for a general measurable function $g : \mathcal{X} \times \{-1, 1\} \longmapsto \mathbb{R}$, $Pg(X, Y)$ is the expectation of the function g, i.e., $Pg(X, Y) \equiv \int g(x, y)dP(x, y)$. Hence, the test error is the probability of misclassification and therefore a small test error indicates a good classifier. We assume that $Pr\big(\widehat{f}(X) = 0\big) = 0$; therefore the event of misclassification is $\big\{Y\widehat{f}(X) \leq 0\big\}$ and the test error is $\tau(\widehat{f}) = P\mathbf{1}\Big\{Y\widehat{f}(X) \leq 0\Big\}$.

The kernel-machine classifier belongs to an RKHS. The RKHS $\mathcal{H}$ is a space of functions from $\mathcal{X}$ to $\mathbb{R}$ and is defined by a kernel function $k : \mathcal{X} \times \mathcal{X} \longmapsto \mathbb{R}$ (Steinwart and Christmann, 2008, Chapter 4). Under some conditions, an RKHS is dense in the space $C(\mathcal{X})$ which is the space of all the continuous functions from $\mathcal{X}$ to $\mathbb{R}$. A loss function is a function $L : \{-1, 1\} \times \mathbb{R} \to [0, \infty)$. The risk of a given function $f \in \mathcal{H}$ with respect to a loss function L is defined by $\mathcal{R}(f) \equiv PL(Y, f(X))$. The empirical risk is $\mathbb{P}_n L(Y, f(X))$, where $\mathbb{P}_n$ is the empirical measure, i.e., for a

measurable function $g : \mathcal{X} \times \{-1, 1\} \longmapsto \mathbb{R}$,

$$\mathbb{P}_n g(X, Y) \equiv \frac{1}{n} \sum_{i=1}^{n} g(X_i, Y_i).$$

The kernel-machine classifier is the minimizer of the regularized empirical risk

$$\widehat{f}_{D,\lambda_n} \equiv \underset{f \in \mathcal{H}}{\operatorname{argmin}} \left\{ \mathbb{P}_n L(Y, f(X)) + \lambda_n \|f\|^2_{\mathcal{H}} \right\}. \tag{1}$$

The first term in (1) is the empirical risk and the second term penalizes the complexity of f in order to avoid overfitting. The regularization parameter λ_n is a positive real number usually chosen by cross validation.

3. Naive Confidence Intervals.

3.1. *Normal approximation.* For a general measurable function $g : \mathcal{X} \times \{-1, 1\} \longmapsto \mathbb{R}$, denote the following

$$\mathbb{G}_n g\,(X, Y) \equiv \sqrt{n}\,(\mathbb{P}_n - P)\,g(X, Y).$$

The central limit theorem (CLT) ensures that $\mathbb{G}_n g(X, Y) \rightsquigarrow N(0, V_g)$, where $\rightsquigarrow$ denotes convergence in distribution and V_g is the variance of $g(X, Y)$. This convergence allows a construction of a CI for the unknown deterministic term $Pg(X, Y)$.

More precisely,

$$\left[\mathbb{P}_n g(X, Y) - \sqrt{\frac{\widehat{V}_g}{n}} Z_{1-\frac{\alpha}{2}}, \quad \mathbb{P}_n g(X, Y) + \sqrt{\frac{\widehat{V}_g}{n}} Z_{1-\frac{\alpha}{2}} \right]$$

is a confidence interval for $Pg(X, Y)$ with an asymptotic confidence level of $1 - \alpha$, where $\widehat{V}_g \equiv \mathbb{P}_n g^2(X, Y) - \mathbb{P}^2_n g(X, Y)$ and $Z_{1-\frac{\alpha}{2}}$ is the $1 - \frac{\alpha}{2}$ quantile of the standard normal distribution.

When $\widehat{g}_n : \mathcal{X} \times \{-1, 1\} \longmapsto \mathbb{R}$ is a function that depends on the sampled data, the CLT does not hold. Hence, the asymptotic convergence of $\mathbb{G}_n \widehat{g}_n(X, Y)$ to a normal distribution may not hold. Here in the definition of $\mathbb{G}_n \widehat{g}_n(X, Y)$ the term $P\widehat{g}_n(X, Y)$ is defined as $P\widehat{g}_n(X, Y) \equiv \int \widehat{g}_n(x, y) dP(x, y)$. Our interest is to construct a CI for the test error $\tau(\widehat{f}_{D,\lambda_n}) = P\mathbf{1}\left\{ Y \widehat{f}_{D,\lambda_n}(X) \leq 0 \right\}$. Note that $\mathbf{1}\left\{ Y \widehat{f}_{D,\lambda_n}(X) \leq 0 \right\}$ depends on the sampled data, so according to the last comment, the standard CLT may not hold.

The following theorem is from Bolthausen et al. (2002, Theorem 6.15, Chapter 6.4) and gives conditions in which a convergence of $\mathbb{G}_n \widehat{g}_n(X, Y)$ to a normal distribution holds where $\widehat{g}_n$ is a given function that depends on the sampled data.

LEMMA 1. *Let $\widehat{g}_n : \mathcal{X} \longmapsto \mathbb{R}$ be a function that depends on the sampled data such that*

i) $Pr(\widehat{g}_n \in \mathcal{F}) \to 1$ *for a* P-Donsker *class* $\mathcal{F}$

ii) $P(\widehat{g}_n - g_0)^2 \to 0$ *in probability for some* $g_0 \in L_2(P)$

Then $\mathbb{G}_n(\widehat{g}_n - g_0)(X, Y) \rightsquigarrow 0$.

The proof is in Bolthausen et al. (2002, Theorem 6.15, Chapter 6.4)

This lemma gives conditions in which $\mathbb{G}_n(\widehat{g}_n - g_0)(X, Y) \rightsquigarrow 0$. Since

$$\mathbb{G}_n\widehat{g}_n(X, Y) = \mathbb{G}_n(\widehat{g}_n - g_0)(X, Y) + \mathbb{G}_n g_0(X, Y)$$

and since by the CLT $\mathbb{G}_n g_0(X, Y) \rightsquigarrow N(0, V_{g_0})$, by the Slutsky theorem (Kosorok, 2008, Theorem 7.15) it follows that $\mathbb{G}_n\widehat{g}_n \rightsquigarrow N(0, V_{g_0})$.

Our interest is in constructing a CI to the quantity $P\mathbf{1}\left\{Y\widehat{f}_{D,\lambda_n}(X) \leq 0\right\}$. Since the functional $1_{t\leq 0}$ is not continuous the first condition in Lemma 1 does not hold. In order to apply Lemma 1 to our setting, we need to replace the function $1_{t\leq 0}$ by a smooth function.

The following theorem is based on Lemma 1.

THEOREM 1. *Consider a setting in which $\lambda_n \to \lambda_0 > 0$; then for any given function $h : \mathbb{R} \mapsto \mathbb{R}$ which is a Lipschitz continuous and bounded function,*

$$\mathbb{G}_n h(\widehat{f}_{D,\lambda_n}(X)Y) \rightsquigarrow N(0, V_{h(Yf^*(X))}), \tag{2}$$

where

$$f^* = \underset{f\in\mathcal{H}}{\operatorname{argmin}}\, PL(Y, f(X)) + \lambda_0\|f\|^2_{\mathcal{H}}. \tag{3}$$

The proof of the theorem is based on the following lemmas

LEMMA 2. *Let $B_1(\mathcal{H})$ be the open unit ball of the RKHS $\mathcal{H}$. Then, there exists a constant $c > 0$ such that the classifier $\widehat{f}$ belongs to $B_c(\mathcal{H}) \equiv c \cdot B_1(\mathcal{H})$ with probability* 1.

PROOF. The SVM classifier is defined as

$$\widehat{f}_{D,\lambda_n} := \underset{f\in\mathcal{H}}{\operatorname{argmin}}\, \{\mathbb{P}_n L(Y, f(X)) + \lambda_n\|f\|^2_{\mathcal{H}}\}.$$

Therefore,

$$\mathbb{P}_n L(Y, \widehat{f}_{D,\lambda_n}(X)) + \lambda_n\|\widehat{f}_{D,\lambda_n}\|^2_{\mathcal{H}} \leq \mathbb{P}_n L(Y, 0) + \lambda_n\|0\|^2_{\mathcal{H}}. \tag{4}$$

From the inequality in (4) we have

$$\lambda_n \|\widehat{f}_{D,\lambda_n}\|_{\mathcal{H}}^2 \leq \mathbb{P}_n L(Y, \widehat{f}_{D,\lambda_n}(X)) + \lambda_n \|\widehat{f}_{D,\lambda_n}\|_{\mathcal{H}}^2 \leq \mathbb{P}_n L(Y,0) + \lambda_n \|0\|_{\mathcal{H}}^2 = \mathbb{P}_n L(Y,0) \leq M$$

where $M = \min\{L(-1,0), L(1,0)\}$. It follows from the inequalities in (4) that $\|\widehat{f}_{D,\lambda_n}\|_{\mathcal{H}} \leq \frac{\sqrt{M}}{\sqrt{\lambda_n}}$. Therefore, for $m = \inf\{\lambda_0, \lambda_1, \lambda_2, ...\}$ then $\|\widehat{f}_{D,\lambda_n}\| \leq \sqrt{\frac{M}{m}}$. Hence, for any $c > \sqrt{\frac{M}{m}}$, we obtain $\widehat{f}_{D,\lambda_n} \in B_c(\mathcal{H})$ as desired. □

LEMMA 3. *The function $\widehat{f}_{D,\lambda_n}$ converges weakly to f^* where $\widehat{f}_{D,\lambda_n}$ and f^* are defined in 1 and 3 respectively.*

PROOF. Let $\mathcal{F} \subseteq \mathcal{H}$ be a closed ball. Let $M_n : \mathcal{F} \longmapsto \mathbb{R}$ be defined by

$$M_n(f) := -(\mathbb{P}_n L(Y, f(X)) + \lambda_n \|f\|_{\mathcal{H}}^2).$$

By the definition of the SVM classifier in 1 we have that $M_n(\widehat{f}_{D,\lambda_n}) \geq \sup_{f \in \mathcal{F}} M_n(f)$.

For $M : \mathcal{F} \longmapsto \mathbb{R}$ defined as $M(f) := -(PL(Y, f(X)) + \lambda_0 \|f\|_{\mathcal{H}}^2)$ it follows that f^* is the unique maximizer of M. Since $\mathcal{F}$ is a closed ball, the sequence $\widehat{f}_{D,\lambda_n}$ together with f^* are uniformly tight. Note that

$$M_n(f) - M(f) = (P - \mathbb{P}_n)L(Y, f(X)) + (\lambda_0 - \lambda_n)\|f\|_{\mathcal{H}}^2.$$

Since $\mathcal{F}$ is a closed ball and $\lambda_n \to \lambda_0$, $\sup_{f \in \mathcal{F}}(\lambda_0 - \lambda_n)\|f\|_{\mathcal{H}}^2 \to 0$.

Since $\mathcal{F}$ is a *P-Donsker* class (Kosorok, 2008) and L is a convex function and thus a continuous function, $L \circ F = \{L \circ f | f \in \mathcal{F}\}$ is a Glivenko-Cantelli class (Kosorok, 2008, Corollary 9.27). Therefore $\sup_{f \in \mathcal{F}} \mid (P - \mathbb{P}_n)L(Y, f(X)) \mid \to 0$. It follows that $M_n \rightsquigarrow M$,from the Argmax theorem (Kosorok, 2008, Theorem 14.1), $\widehat{f}_{D,\lambda_n} \rightsquigarrow f^*$ as desired.

□

LEMMA 4. *If h is* Lipschitz *continuous and bounded, then $P^2\big(h(\widehat{f}_{D,\lambda_n}) - h(f^*)\big) \to 0$.*

PROOF. Because h is continuous, then due to Lemma 3 and the Slutsky theorem, we have that $h(\widehat{f}_{D,\lambda_n}) - h(f^*) \to 0$ in probability; since h is bounded, then $P^2\big(h(\widehat{f}_{D,\lambda_n}) - h(f^*)\big) \to 0$, by the bounded convergence theorem, as desired □

LEMMA 5. *If h is* Lipschitz *continuous and bounded, then $h(\widehat{f}_{D,\lambda_n}) - h(f^*)$ belongs to a P-*Donsker *class.*

PROOF. According to Steinwart and Christmann (2008, Corollary 4.36) and Kosorok (2008, Theorem 9.19) $B_1(\mathcal{H})$, a closed ball in an RKHS is a *P-Donsker* class and since a composition of a *Lipschitz* continuous and bounded function with a *P-Donsker* is a *P-Donsker* class (Kosorok, 2008, Corollary 9.32), then $h(\widehat{f}_{D,\lambda_n})$ belongs to a *P-Donsker* class. Since $h(f^*)$ is deterministic, then $h(\widehat{f}_{D,\lambda_n}) - h(f^*)$ belongs to a *P-Donsker* class, as desired □

Lemmas 4 and 5 show that the two conditions in Lemma 1 hold. Therefore, we conclude that

$$\mathbb{G}_n[h(\widehat{f}_{D,\lambda_n}(X)) - h(f^*(X))] \rightsquigarrow 0.$$

Similar arguments show that

$$\mathbb{G}_n[h(-\widehat{f}_{D,\lambda_n}(X)) - h(-f^*(X))] \rightsquigarrow 0.$$

From this we conclude that

$$\mathbb{G}_n[h(Y\widehat{f}_{D,\lambda_n}(X)) - h(Yf^*(X))] \rightsquigarrow 0$$

which completes the proof of Theorem 1.

From Theorem 1 we have the following conclusion

CONCLUSION 2. *Let $\widehat{f}_{D,\lambda_n}$ be a kernel classifier over an RKHS and let $h : \mathbb{R} \longmapsto \mathbb{R}$ be a* Lipschitz continuous *and bounded function. Then*

$$Pr\left(Ph\big(\widehat{f}_{D,\lambda_n}(X)Y\big) < \mathbb{P}_n h\big(\widehat{f}_{D,\lambda_n}(X)Y\big) - \frac{Z_{1-\frac{\alpha}{2}}}{\sqrt{n}}\sqrt{\widehat{V}_h}\right) = \frac{\alpha}{2} + o(1) \tag{5}$$

$$Pr\left(Ph\big(\widehat{f}_{D,\lambda_n}(X)Y\big) > \mathbb{P}_n h\big(\widehat{f}_{D,\lambda_n}(X)Y\big) + \frac{Z_{1-\frac{\alpha}{2}}}{\sqrt{n}}\sqrt{\widehat{V}_h}\right) = \frac{\alpha}{2} + o(1) \tag{6}$$

where $Z_{1-\frac{\alpha}{2}}$ is the $1-\frac{\alpha}{2}$ quantile of the standard normal distribution.

Our interest is a CI to the quantity $P1_{\{Y\widehat{f}_{D,\lambda_n}(X)\leq 0\}}$. Since the functional $1_{t\leq 0}$ is not continuous, a construction for $P1_{\{Y\widehat{f}_{D,\lambda_n}(X)\leq 0\}}$, as stated in Conclusion 2, does not hold. Instead, take any two functions $h^+, h^- : \mathbb{R} \mapsto \mathbb{R}$ which are Lipschitz continuous and bounded with the property that

$$h^-(t) \leq 1_{\{t\leq 0\}} \leq h^+(t).$$

Then, by definition,

$$Ph^-\left(\widehat{f}_{D,\lambda_n}(X)Y\right) \leq P1_{\{Y\widehat{f}_{D,\lambda_n}(X)\leq 0\}} \leq Ph^+\left(\widehat{f}_{D,\lambda_n}(X)Y\right). \tag{7}$$

By Conclusion 2

$$\left[\mathbb{P}_n h^-\big(\widehat{f}_{D,\lambda_n}(X)Y\big) - \sqrt{\frac{\widehat{V}_{h^-}}{n}} Z_{1-\frac{\alpha}{2}},\quad \mathbb{P}_n h^+\big(\widehat{f}_{D,\lambda_n}(X)Y\big) + \sqrt{\frac{\widehat{V}_{h^+}}{n}} Z_{1-\frac{\alpha}{2}}\right] \tag{8}$$

is a conservative CI for the test error with an asymptotic confidence level of $1-\alpha$.

The length of the CI given in (15) is

$$\mathbb{P}_n\left[h^+\big(\widehat{f}_{D,\lambda_n}(X)Y\big) - h^-\big(\widehat{f}_{D,\lambda_n}(X)Y\big)\right] + \frac{Z_{1-\frac{\alpha}{2}}}{\sqrt{n}}\left(\sqrt{\widehat{V}_{h^+}} - \sqrt{\widehat{V}_{h^-}}\right)$$

which converges to $P\left[h^+\big(f^*(X)Y\big) - h^-\big(f^*(X)Y\big)\right]$. Therefore, the length of the CI given in (15) does not necessarily converge to zero.

3.2. *Empirical Bootstrap.* Another common method to construct a CI for a deterministic quantity Pg is by using the empirical bootstrap process. The empirical bootstrap process, indexed by g, is defined as follows

$$\mathbb{G}_n^{(b)} g(X,Y) = \sqrt{n}\big(\mathbb{P}_n^{(b)} - \mathbb{P}_n\big) g(X,Y),$$

where, for a measurable function $g : \mathcal{X} \times \{-1,1\} \longmapsto \mathbb{R}$, the empirical bootstrap mean is defined as

$$\mathbb{P}_n^{(b)} g(Y,X) = \frac{1}{n}\sum_{i=1}^{n} W_{ni} g(X_i, Y_i)$$

where

$$\big(W_{n1}, W_{n2}, ..., W_{nn}\big) \sim \text{Multinomial}\left(n; \frac{1}{n}, \frac{1}{n}, ..., \frac{1}{n}\right).$$

According to (Kosorok, 2008, theorem 10.4), $\mathbb{G}_n^{(b)} g(X,Y)$ and $\mathbb{G}_n g(X,Y)$ converge to the same limit. Therefore, let $\widehat{p}_{g,q}$ be a consistent estimator to the q percentile of $\mathbb{G}_n^{(b)} g(X,Y)$. Then

$$\left[\mathbb{P}_n g(Y,X) - \frac{\widehat{p}_{g,1-\frac{\alpha}{2}}}{\sqrt{n}},\quad \mathbb{P}_n g(Y,X) - \frac{\widehat{p}_{g,\frac{\alpha}{2}}}{\sqrt{n}}\right]. \tag{9}$$

is a CI to the quantity Pg with an asymptotic confidence level of $1-\alpha$. As seen in Chapter 3.1, since the test error is a random quantity, the CI given in (9) does not necessarily hold when plugging in a random function $\widehat{g}_n$. The next lemma gives a condition in which a plugging in a random function $\widehat{g}$ to (9) results with a CI to $P\widehat{g}$ in a confidence level of $1-\alpha$.

LEMMA 6. *Let $\widehat{g}_n : \mathcal{X} \longmapsto \mathbb{R}$ be a function that depends on the sampled data such that*

i) $Pr(\widehat{g}_n \in \mathcal{F}) \to 1$ *for a* P-Donsker *class* $\mathcal{F}$

ii) $P(\widehat{g}_n - g_0)^2 \to 0$ *in probability for some* $g_0 \in L_2(P)$

Then $\mathbb{G}_n^b(\widehat{g}_n - g_0)(X,Y) \rightsquigarrow 0$.

The proof of this lemma is similar to that of Lemma 1.

By Lemma 6 we have that

$$\left[\mathbb{P}_n h^-\big(\widehat{f}_{D,\lambda_n}(X)Y\big) - \frac{\widehat{p}_{h^-,1-\frac{\alpha}{2}}}{\sqrt{n}}, \quad \mathbb{P}_n h^+\big(\widehat{f}_{D,\lambda_n}(X)Y\big) + \frac{\widehat{p}_{h^+,\frac{\alpha}{2}}}{\sqrt{n}}\right] \tag{10}$$

is a conservative CI for the test error where $\widehat{p}_{h^-,1-\frac{\alpha}{2}}$ is the $1-\frac{\alpha}{2}$ quantile of $\mathbb{G}_n^{(b)} h^-(Y\widehat{f}_{D,\lambda_n}(X))$ and $\widehat{p}_{h^+,\frac{\alpha}{2}}$ is the $\frac{\alpha}{2}$ quantile of $\mathbb{G}_n^{(b)} h^+(Y\widehat{f}_{D,\lambda_n}(X))$.

4. Adaptive Confidence Intervals. The aim of this section is to construct confidence intervals for the kernel classifier's test error with asymptotic $(1-\alpha)100\%$ confidence level. However, such a construction is challenging since the test error is an expectation of the non-continuous function $\mathbf{1}\{t \le 0\}$ where $t = Y\widehat{f}_{D,\lambda_n}(X)$. The idea is to replace the construction of the naive method, given in Section 3, by taking sequences of smooth functions that converge to the function $\mathbf{1}\{t \le 0\}$ as the sample size grows. In order to keep a P-Donsker structure, delicate arguments are given, discussing the assumptions on the smoothness and the rate of convergence of the functions in the sequence. Let $\{h_m^-\}_{m=1}^\infty$ and $\{h_m^+\}_{m=1}^\infty$ be sequences of functions from $\mathbb{R}$ to $\mathbb{R}$ with the following properties:

i) For all $t \in \mathbb{R}$, $h_m^-(t) \le \mathbf{1}\{t \le 0\} \le h_m^+(t)$.

ii) For all $t \in \mathbb{R}$, both $h_m^-(t)$ and $h_m^+(t)$ converge pointwise to the function $\mathbf{1}\{t \le 0\}$ as $m \to \infty$.

Let $\{h_m^\circ\}_{m=1}^\infty$ be a general sequence of functions which satisfies $h_m^\circ(t) \to \mathbf{1}\{t \le 0\}$ as $m \to \infty$, specifically h° can be either h^- or h^+. By equation (5), for any fixed m,

$$\mathbb{G}_n h_m^\circ\left(Y\widehat{f}_{D,\lambda_n}(X)\right) \underset{n\to\infty}{\rightsquigarrow} N\left(0, V_{h_m^\circ}(Yf^*(X))\right). \tag{11}$$

By the dominant convergence theorem,

$$N\left(0, V_{h_m^\circ}\left(Yf^*(X)\right)\right) \underset{m\to\infty}{\rightsquigarrow} N\left(0, P\mathbf{1}\{Yf^*(X) \le 0\}(1 - P\mathbf{1}\{Yf^*(X)\} \le 0)\right). \tag{12}$$

Therefore, we would like to find a sequence $\{m_n\}_{n=1}^\infty$ with $m_n \to \infty$, such that

$$\mathbb{G}_n h_{m_n}^\circ\left(Y\widehat{f}_{D,\lambda_n}(X)\right) \underset{n\to\infty}{\rightsquigarrow} N\left(0, P\mathbf{1}\{Yf^*(X) \le 0\}(1 - P\mathbf{1}\{Yf^*(X)\} \le 0)\right). \tag{13}$$

We assume the following

(A1) The kernel k has $\frac{d}{2}+1$ derivatives.

(A2) The loss function is a convex locally-Lipschitz continuous function satisfying $L(y,0)\leq 1$.

(A3) $\mathcal{X}\subseteq\mathbb{R}^d$ is bounded and convex with nonempty interior.

(A4) $\lambda_n\to\lambda_0>0$.

In the following, we present conditions on the sequence $\{m_n\}_{n=1}^{\infty}$ which ensures that both $\{h^-_{m_n}\}_{n=1}^{\infty}$ and $\{h^+_{m_n}\}_{n=1}^{\infty}$ satisfy (13). Using this result we will be able to prove the following theorem.

THEOREM 3. *Let $m_n=o(n^{\varepsilon})$, where $0<\varepsilon<\frac{1}{d+3}$, $s(d)=\lceil\frac{d+3}{4}\rceil+1$, and let Assumptions (A1)-(A4) hold. Define the following functions from $\mathbb{R}$ to $\mathbb{R}$,*

$$g(t)\equiv\frac{1}{2}\exp\left(-\frac{t^{2s(d)}}{1-t^{2s(d)}}\right)\mathbf{1}\{|t|<1\},$$

$$h(t)\equiv\begin{cases}g(t) & \text{if } t\geq 0,\\ 1-g(t) & \text{if } t<0,\end{cases}\tag{14}$$

$h^-_{m_n}(t)\equiv h(m_nt+1)$ and $h^+_{m_n}(t)\equiv h(m_nt-1)$. Then, the normal-approximation-based CI

$$\left[\mathbb{P}_nh^-_{m_n}\big(\widehat{f}_{D,\lambda_n}(X)Y\big)-\sqrt{\frac{\widehat{V}_{h^-_{m_n}}}{n}}Z_{1-\frac{\alpha}{2}},\quad \mathbb{P}_nh^+_{m_n}\big(\widehat{f}_{D,\lambda_n}(X)Y\big)+\sqrt{\frac{\widehat{V}_{h^+_{m_n}}}{n}}Z_{1-\frac{\alpha}{2}}\right]\tag{15}$$

and the empirical-bootstrap-based CI

$$\left[\mathbb{P}_nh^-_{m_n}\big(\widehat{f}_{D,\lambda_n}(X)Y\big)-\frac{\widehat{p}_{h^-_{m_n},1-\frac{\alpha}{2}}}{\sqrt{n}},\quad \mathbb{P}_nh^+_{m_n}\big(\widehat{f}_{D,\lambda_n}(X)Y\big)+\frac{\widehat{p}_{h^+_{m_n},\frac{\alpha}{2}}}{\sqrt{n}}\right]\tag{16}$$

are both confidence intervals for the test error with an asymptotic confidence level of $(1-\alpha)100\%$.

The proof of the theorem is based on the following lemmas. First we prove the theorem for the confidence interval in (15).

LEMMA 7. *Let assumptions (A1)-(A4) hold. Let $B_1(\mathcal{H})$ be the open unit ball of the RKHS $\mathcal{H}$. Then, for any $0<\varepsilon<\frac{1}{2}$*

$$Pr\left(n^{\frac{1}{2}-\varepsilon}(\widehat{f}_{D,\lambda}-f^*)\in B_1(\mathcal{H})\right)\underset{n\to\infty}{\longrightarrow}1.\tag{17}$$

PROOF. By the proof of Theorem 6.24 in Steinwart and Christmann (2008), for any $\tau > 0$,

$$Pr\left(\|n^{\frac{1}{2}-\varepsilon}(\widehat{f}_{D,\lambda} - f^*)\|_{\mathcal{H}} \geq \lambda^{-1}|L|_{\lambda^{-\frac{1}{2}},1} n^{\frac{1}{2}-\varepsilon}\left(\sqrt{\frac{2\tau}{n}} + \sqrt{\frac{1}{n}} + \frac{4\tau}{3n}\right)\right) \leq e^{-\tau}. \tag{18}$$

Set $\tau = \tau_n = n^{\varepsilon}$; then the inequality given in (18) is

$$Pr\left(\|n^{\frac{1}{2}-\varepsilon}(\widehat{f}_{D,\lambda} - f^*)\|_{\mathcal{H}} \geq a_n\right) \leq b_n, \tag{19}$$

where $a_n = \lambda^{-1}|L|_{\lambda^{-\frac{1}{2}},1}(\sqrt{2}n^{-\frac{\varepsilon}{2}} + n^{-\varepsilon} + \frac{4}{3}n^{-\frac{1}{2}})$ and $b_n = e^{-n^{\varepsilon}}$. Note that $a_n \to 0$ and $b_n \to 0$; therefore by (19), $Pr\left(\|n^{\frac{1}{2}-\varepsilon}(\widehat{f}_{D,\lambda} - f^*)\|_{\mathcal{H}} \geq 1\right) \longrightarrow 0$. Hence,
$Pr\left(n^{\frac{1}{2}-\varepsilon}(\widehat{f}_{D,\lambda} - f) \in B_1(\mathcal{H})\right) \longrightarrow 1$, as desired. □

We shall define a class of bounded smooth functions on $\mathcal{X} \subseteq \mathbb{R}^d$ as follows. For any vector $k = (k_1, ..., k_d)$ of non-negative integers, define the differential operator $D^k \equiv \frac{\partial^{|k|}}{(\partial x_1^{k_1}, ..., \partial x_d^{k_d})}$, where $|k| \equiv k_1 + ... + k_d$. For any $x \in \mathbb{R}$, let $\lfloor x \rfloor$ be the largest integer $j \leq x$. Let $f : \mathcal{X} \mapsto \mathbb{R}$ and $\alpha > 0$; define the norm

$$\|f\|_{\alpha} := \max_{k:|k| \leq \lfloor \alpha \rfloor} \sup_{x \in \mathcal{X}} |D^k f(x)| + \max_{k:|k| = \lfloor \alpha \rfloor} \sup_{x, x' \in \mathcal{X}, x \neq x'} \frac{|D^k f(x) - D^k f(x')|}{\|x - x'\|^{\alpha - \lfloor \alpha \rfloor}}. \tag{20}$$

When $k = 0$ we set $D^k f = f$. Now, let $\mathcal{C}_M^{\alpha}(\mathcal{X})$ be the set of all continuous functions $f : \mathcal{X} \mapsto \mathbb{R}$ with $\|f\|_{\alpha} \leq M$. By Kosorok (2008, Theorem 9.19), if $\mathcal{X} \subseteq \mathbb{R}^d$ is bounded and convex with a nonempty interior and $\alpha > \frac{d}{2}$, then the class of functions $\mathcal{C}_M^{\alpha}(\mathcal{X})$ is a P-Donsker class.

Let $C(\mathcal{X})$ be the class of all the continuous functions from $\mathcal{X}$ to $\mathbb{R}$. For a class of functions $\mathcal{F} \subseteq C(\mathcal{X})$, and differential function $h : \mathbb{R} \mapsto \mathbb{R}$, and given constants $m > 1$, $s > 1$, and $c \in \mathbb{R}$, define $H_{m,s,c,h} : \mathcal{F} \times \mathcal{F} \mapsto C(\mathcal{X})$ as

$$H_{m,s,c,h}(g, f) := \begin{cases} \frac{h(mf+c) - h(mg+c)}{m^{s+1}(f-g)} & f \neq g, \\ \frac{h^{(1)}(mf+c)}{m^s} & f = g. \end{cases}$$

LEMMA 8. *Let $h : \mathbb{R} \mapsto \mathbb{R}$ be such that h has $s+1$ derivatives that are all uniformly bounded by a constant M, and let $\mathcal{F} \subseteq C_M^s(\mathcal{X})$. Then, there is a constant $\widetilde{M}$, independent of m, such that the class of functions*

$$\mathcal{H}_{\mathcal{F},s,c,h} \equiv \{H_{m,s,c,h}(g, f) | m > 1, f, g \in \mathcal{F}\}$$

is a sub-class of $C_{\widetilde{M}}^s(\mathcal{X})$.

PROOF. Clearly, since a change in scale does not affect the boundaries of a function, it is sufficient to prove the lemma where $c = 0$, i.e., for $\mathcal{H}_{\mathcal{F},s,0,h}$. We will prove the case in which $\mathcal{F}$

is a class of functions from $\mathcal{X} \subseteq \mathbb{R}$ to $\mathbb{R}$, similar arguments hold when $\mathcal{X} \subseteq \mathbb{R}^d$. For any integer $j \geq 0$, we use the notation $h^{(j)}(y) = \frac{d^j}{dy^j}h(y)$ and $f^{(j)}(x) = \frac{d^j}{dx^j}f(x), g^{(j)}(x) = \frac{d^j}{dx^j}g(x)$. Note that

$$m\int_0^1 h^{(1)}(tmf + (1-t)mg)dt = \int_0^1 \frac{d}{dt}\frac{h(tmf + (1-t)mg)}{f-g}dt = \frac{h(mf) - h(mg)}{f-g}.$$

Hence, the l-th derivative of $\frac{h(mf)-h(mg)}{f-g}$ is by the Faa di Bruno formula

$$\begin{aligned}
&\frac{d^l}{dx^l}\frac{h(mf) - h(mg)}{f-g} \\
&= m\int_0^1 \frac{d^l}{dx^l}h^{(1)}(tmf + (1-t)mg)dt \\
&= m\int_0^1 l!\sum \big(\prod_{j=1}^{l} n_j!(j!)^{n_j}\big)^{-1} h^{(1+n_1+\cdots n_l)}\big(tmf + (1-t)mg\big)\prod_{j=1}^{l}[tmf^{(j)} + (1-t)mg^{(j)}]^{n_j}dt \\
&= m\int_0^1 l!\sum \big(\prod_{j=1}^{l} n_j!(j!)^{n_j}\big)^{-1} h^{(1+n_1+\cdots n_l)}\big(tmf + (1-t)mg\big)\prod_{j=1}^{l}[tf^{(j)} + (1-t)g^{(j)}]^{n_j}m^{n_j}dt \\
&= m\int_0^1 l!\sum \big(\prod_{j=1}^{l} n_j!(j!)^{n_j}\big)^{-1} m^{n_1+\cdots n_l} h^{(1+n_1+\cdots n_l)}\big(tmf + (1-t)mg\big)\prod_{j=1}^{l}[tf^{(j)} + (1-t)g^{(j)}]^{n_j}dt \\
&= \int_0^1 l!\sum \big(\prod_{j=1}^{l} n_j!(j!)^{n_j}\big)^{-1} m^{1+n_1+\cdots n_l} h^{(1+n_1+\cdots n_l)}\big(tmf + (1-t)mg\big)\prod_{j=1}^{l}[tf^{(j)} + (1-t)g^{(j)}]^{n_j}dt
\end{aligned}$$

where the sum is taken over all l-tuples of non-negative integers $(n_1, n_2, ..., n_l)$ satisfying the constraint $1 \cdot n_1 + 2 \cdot n_2 + \cdots + l \cdot n_l = l$ under this constraint $n_1 + n_2 + \cdots n_l \leq l$. Therefore, $\left|\frac{d^l}{dx^l}\frac{h(mf)-h(mg)}{f-g}\right| \leq m^{l+1}M^{l+1}l!\sum(\prod_{j=1}^{l} n_j!(j!)^{n_j})^{-1}$. Define

$$\widetilde{M} = \max_{1\leq l\leq s}\left\{M^{l+1}l!\sum\big(\prod_{j=1}^{l} n_j!(j!)^{n_j}\big)^{-1}\right\}.$$

Then, for any $1 \leq l \leq s$,

$$\left|\frac{d^l}{dx^l}\frac{h(mf) - h(mg)}{m^{s+1}(f-g)}\right| \leq \widetilde{M}.$$

Hence,

$$\mathcal{H}_{\mathcal{F},s,0,h} \equiv \{H_{m,s,0,h}(g,f) | m > 1, f, g \in \mathcal{F}\}$$

is a sub-class of $C^s_{\widetilde{M}}(\mathcal{X})$. Therefore, for any $c \in \mathbb{R}$,

$$\mathcal{H}_{\mathcal{F},s,c,h} \equiv \{H_{m,s,c,h}(g,f) | m > 1, f, g \in \mathcal{F}\}$$

is a sub-class of $C^s_{\widetilde{M}}(\mathcal{X})$ as desired $\square$

LEMMA 9. *For any $x \in \mathbb{R}$, let $\lceil x \rceil$ be the smallest integer $j \geq x$. Let Assumptions (A1)-(A3) hold, and assume that h° has $\lceil \frac{d+3}{2} \rceil$ derivatives all bounded by constant M. Let $m_n = n^{\frac{1}{d+3}-\varepsilon}$, where $0 < \varepsilon < \frac{1}{d+3}$. Then, for any constant c, the random function $h^\circ(m_n \widehat{f} + c) - h^\circ(m_n f^* + c)$ belongs to a P-Donsker class with probability converging to 1.*

PROOF. Write

(21)
$$h^\circ(m_n \widehat{f}) - h^\circ(m_n f^*) = \frac{h^\circ(m_n \widehat{f} + c) - h^\circ(m_n f^* + c)}{m_n^{s+1}(\widehat{f} - f^*)} \cdot m_n^{s+1}(\widehat{f} - f^*) \equiv D_{n,m_n} \cdot E_{n,m_n},$$

where $s = \frac{d+1}{2}$. By definition and by Lemma 7, the term $D_{n,m_n} \equiv \frac{h^\circ(m_n \widehat{f}+c) - h^\circ(m_n f^*+c)}{m_n^{s+1}(\widehat{f}-f^*)}$ belongs to functions class $\mathcal{H}_{\mathcal{F},s,h}$ where $\mathcal{F} = B_c(\mathcal{H})$. According to Steinwart and Christmann (2008, Corollary 4.36), there exists a bound M such $B_c(\mathcal{H})$ belongs to $C_M^s(\mathcal{X})$. Hence, by Lemma 8, there exists a bound $\widetilde{M}$ such that the function class $\mathcal{H}_{B_c(\mathcal{H}),s,h}$ is contained in $C_{\widetilde{M}}^s(\mathcal{X})$. Since that $s = \frac{d+1}{2} > \frac{d}{2}$, by Kosorok (2008, Theorem 9.19), we have that $C_{\widetilde{M}}^s(\mathcal{X})$ is a P-Donsker class. The term E_{n,m_n} equals $m_n^{s+1}(\widehat{f} - f^*)$, since

$$m_n^{s+1} = m_n^{\frac{d+1}{2}+1} = m_n^{\frac{d+3}{2}} = (n^{\frac{1}{d+3}-\varepsilon})^{\frac{d+3}{2}} = n^{\frac{1}{2}-\varepsilon'}$$

where $\varepsilon' \equiv \frac{d+3}{2}\varepsilon$. Since $0 < \varepsilon < \frac{1}{d+3}$ we have $0 < \varepsilon' < \frac{1}{2}$. Hence, according to Lemma 7, E_{n,m_n} belongs to $B_1(\mathcal{H})$ with probability converging to 1 which is a P-Donsker class as discussed before. By Kosorok (2008, Corollary 9.32) a product of two uniformly bounded P-Donsker classes is a P-Donsker class, and hence the random function

$$h^\circ(m_n \widehat{f} + c) - h^\circ(m_n f^* + c)$$

belongs to a P-Donsker class with probability converging to 1 as desired. □

LEMMA 10. *Let $\{h_{m_n}^\circ\}_{n=1}^\infty$ be a sequence of functions which satisfies $h_{m_n}^\circ(t) \to \mathbf{1}\{t \leq 0\}$ as $n \to \infty$. Then the convergence in (13) occurs if and only if*

(22)
$$A_{n,m_n} \equiv \mathbb{G}_n\big(h_{m_n}^\circ(Y \widehat{f}_{D,\lambda_n}(X)) - h_{m_n}^\circ(Y f^*(X))\big) \rightsquigarrow 0.$$

PROOF. Note that

$$\begin{aligned}\mathbb{G}_n\big(h^\circ_{m_n}(Y\widehat{f}_{D,\lambda_n}(X)\big) =& \mathbb{G}_n\big(h^\circ_{m_n}(Y\widehat{f}_{D,\lambda_n}(X) - h^\circ_{m_n}(Yf^*(X))\big)\\ &+ \mathbb{G}_n\big(h^\circ_{m_n}(Yf^*(X) - \mathbf{1}\{(Yf^*(X)) < 0\}\big)\\ &+ \mathbb{G}_n\mathbf{1}\{(Yf^*(X)) < 0\}\\ \equiv& A_{n,m_n} + B_{n,m_n} + C_n\,.\end{aligned}$$

Clearly,

$$C_n \rightsquigarrow N\big(0, P\mathbf{1}\{Yf^*(X) \le 0\}(1 - P\mathbf{1}\{Yf^*(X)\} \le 0)\big). \tag{23}$$

Since f^* is a deterministic function, then by the dominant convergence theorem, $B_{n,m_n} \rightsquigarrow 0$. Therefore (13) holds if and only if $A_{n,m_n} \underset{n\to\infty}{\rightsquigarrow} 0$. □

LEMMA 11. *Let Assumptions (A1)-(A3) hold, and let* $m_n = n^{\frac{1}{d+3}-\varepsilon}$ *where* $0 < \varepsilon < \frac{1}{d+3}$. *Let* $h^\circ_{m_n}(t) = h^\circ(m_n t + c)$, *where* c *is some constant, such that*

$$\lim_{n\to\infty} h^\circ(m_n t) \equiv \begin{cases} 1 & \text{if } t > 0,\\ -1 & \text{if } t < 0\end{cases}$$

and h° *have* $\lceil \frac{d+3}{2} \rceil$ *derivatives, all bounded by a constant* M. *Then,*

$$\mathbb{G}_n\big(h^\circ(m_n Y\widehat{f}_{D,\lambda_n}(X)) + c\big) \rightsquigarrow N\left(0, P\mathbf{1}\{Yf^*(X) \le 0\}(1 - P\mathbf{1}\{Yf^*(X)\} \le 0)\right). \tag{24}$$

PROOF. We will prove that

$$\mathbb{G}_n\big(h^\circ(m_n\widehat{f}_{D,\lambda_n}(X) + c) \rightsquigarrow N\left(0, P\mathbf{1}\{f^*(X) \le 0\}(1 - P\mathbf{1}\{f^*(X)\} \le 0)\right). \tag{25}$$

Similar arguments hold when $Y = -1$. By Lemma 10, it is sufficient to prove that

$$\mathbb{G}_n\left(h^\circ(m_n\widehat{f}_{D,\lambda_n}(X) + c) - h^\circ(m_n f^*(X) + c)\right) \rightsquigarrow 0. \tag{26}$$

To prove the convergence in 26 it is sufficient to show that the conditions in Lemma 1 hold. Clearly, since h° is bounded and $h^\circ(m_n\widehat{f}_{D,\lambda_n}(X) + c) - h^\circ(m_n f^*(X) + c) \rightsquigarrow 0$ then the second condition of Lemma 1 holds. By Lemma 9, also the first condition of Lemma 1 holds. Consequently,

$$\mathbb{G}_n\big(h^\circ(m_n Y\widehat{f}_{D,\lambda_n}(X) + c) \rightsquigarrow N\left(0, P\mathbf{1}\{Yf^*(X) \le 0\}(1 - P\mathbf{1}\{Yf^*(X)\} \le 0)\right) \tag{27}$$

as desired. □

LEMMA 12. *Let* $f(t) = \exp\left(\frac{-t^{2s}}{1-t^{2s}}\right)\mathbf{1}\{|t|<1\}$. *Then, for any* $1 \le l \le 2s-1$

$$\lim_{t\to -1} f^{(l)}(t) = \lim_{t\to 1} f^{(l)}(t) = f^{(l)}(0) = 0.$$

PROOF. For any function of the form $\exp(r(t))$, we have by the Faa di Bruno formula:

$$\frac{d^l}{dt^l}\exp(r(t)) = \exp(r(t))\sum \frac{l!}{\prod_{j=1}^{l} n_j! j!^{n_j}} \prod_{j=1}^{l} (r^{(j)}(t))^{n_j} \tag{28}$$

where the sum is taken over all l-tuples of non-negative integers $(n_1, n_2, ..., n_l)$ satisfying the constraint $1\cdot n_1 + 2\cdot n_2 + \cdots + l\cdot n_l = l$.

By its definition $f(t) = \frac{1}{2}\exp(r(t))\mathbf{1}\{|t|<1\}$ where $r(t) = \frac{-t^{2s}}{1-t^{2s}}$. Therefore, by (28)

$$\frac{d^l}{dt^l} g(t) = \frac{d^l}{dt^l}\frac{1}{2}\exp(r(t)) = \frac{1}{2}\exp(r(t))\sum \frac{l!}{\prod_{j=1}^{l} n_j! j!^{n_j}} \prod_{j=1}^{l} (r^{(j)}(t))^{n_j}\mathbf{1}\{|t|<1\} \tag{29}$$

where the sum is taken over all l-tuples of non-negative integers $(n_1, n_2, ..., n_l)$ satisfying the constraint $1\cdot n_1 + 2\cdot n_2 + \cdots + l\cdot n_l = l$. Recall that $r(t) = \frac{-t^{2s}}{1-t^{2s}}$ and note that for any $1 \le l \le 2s-1$, we have $(t^{2s})^{(l)} = t^{2s-l}\prod_{j=0}^{l-1}(2s-j)$. Then, by elementary derivative rules it follows that for any $1 \le l \le 2s-1$, one can write $r^{(l)}(t) = \frac{p_l(t)}{(1-t^{2s})^l}$ where p_l is a polynomial satisfying $p_l(0) = 0$, and thus $g^{(l)}(0) = 0$. On the other hand, $r^{(l)}(t) = \frac{p_l(t)}{(1-t^{2s})^l} = p_l(t)[\frac{-r(t)}{t^{2s}}]^l$; clearly $\lim_{t\to 1^-}\exp r(t) = 0$ and $\lim_{t\to -1^+}\exp r(t) = 0$. Therefore, $\lim_{t\to 1^-} r^l(t)\exp r(t) = 0$ and $\lim_{t\to -1^+} r^l(t)\exp r(t) = 0$ which yields that $\lim_{t\to -1} f^{(l)}(t) = \lim_{t\to 1} f^{(l)}(t) = 0$, as desired. □

LEMMA 13. *The function* h *defined in* (14) *satisfies the following properties*

(i)

$$\lim_{n\to\infty} h(m_n t) \equiv \begin{cases} 1 & \text{if } t > 0, \\ -1 & \text{if } t < 0. \end{cases}$$

(ii) h *has* $\lceil \frac{d+3}{2} \rceil$ *derivatives all bounded by a constant* M.

PROOF. (i) By definition

$$g(t) \equiv \frac{1}{2}\exp\left(-\frac{t^{2s(d)}}{1-t^{2s(d)}}\right)\mathbf{1}\{|t|<1\},$$

$$h(t) \equiv \begin{cases} g(t) & \text{if } t \ge 0, \\ 1-g(t) & \text{if } t < 0. \end{cases}$$

For any $\{m_n\}_{n=1}^{\infty}$ with $m_n \to \infty$, we have that for all $t \neq 0$, $\lim_{n\to\infty} g(m_n t) = 0$. Therefore, for any $t > 0$, $\lim_{n\to\infty} h(m_n t) = \lim_{n\to\infty} g(m_n t) = 0$, and for any $t < 0$, $\lim_{n\to\infty} h(m_n t) = 1 - \lim_{n\to\infty} g(m_n t) = 1$.

(ii) Clearly, by (28), all the derivatives of the function g are bounded. Using again (28), we have that the limit at zero for all the $\lceil \frac{d+3}{2} \rceil$ is the same for both g and $1-g$, which concludes the proof.

□

The proof of Theorem 3 for the confidence interval in (15) is now followed by defining $h^-(m_n t) = h(m_n t + 1)$ and $h^+(m_n t) = h(m_n t - 1)$ and applying Lemmas 11 and 13. The proof for the empirical bootstrap CI which is presented in (16) follows similarly using Lemmas 9 and 6.

5. Empirical study. In this section we check the performance of the naive and adaptive confidence intervals discussed in Sections 3 and 4, respectively. We checked three different classification settings. For each setting, sample sizes of 100, 200, 400, and 800 were considered. Both normal approximation and empirical bootstrap CIs with a confidence level of 95% were calculated. We also approximated the true test error of a given classifier by constructing a new independent data of $10,000$ observations. We then calculated the empirical mean of misclassification. For each setting and sample size, we repeated the simulation 400 times. Each empirical bootstrap CI was calculated with 800 bootstrap resamples. In our simulated settings we used the square loss function $L(y,t) = (y-t)^2$ and a Gaussian kernel.

The data generating mechanism for the three settings includes a vector of covariates $X = (X_1, X_2, ..., X_d)$ where $X_j \sim \text{Uniform}(0,5)$ are i.i.d, and noise $\varepsilon \sim N(0,1)$ which is independent of the vector X. The response vector in the first setting is similar to that of Laber and Murphy (2012) and is given by $Y = \text{sign}(X_2 - \frac{4}{25}X_1^2 - 1 + \frac{1}{2}\varepsilon)$, The distribution of the response in the second setting is

$$Y = \text{sign}(X_1 - 2.3X_2^2 + 1.25X_3^3 + X_4X_5 - 1.5X_6^3 + X_7/X_8 + 0.3\varepsilon)\,.$$

Finally, the distribution of the response in the third setting is

$$Y = \text{sign}(X_1 - 2X_2^2 + 0.8X_3^3 + 2X_4X_5 - X_6^3 + 0.6\varepsilon)\,.$$

Table 1 shows the resulting coverage of this study. The results are for using the naive CIs and adaptive CIs with a confidence level of 95%; each method was checked on the naive normal and

TABLE 1
The coverage of the four methods for Settings 1-3.

Setting 1

N	Naive Normal	Naive Bootstrap	Adaptive Normal	Adaptive Bootstrap
100	0.98	0.9775	0.97	0.965
200	0.9925	0.9875	0.9675	0.965
400	0.985	0.985	0.9625	0.965
800	0.98	0.975	0.96	0.9575

Setting 2

N	Naive Normal	Naive Bootstrap	Adaptive Normal	Adaptive Bootstrap
100	0.995	0.995	0.975	0.9725
200	0.9975	0.995	0.955	0.9475
400	1	1	0.965	0.9625
800	1	1	0.955	0.955

Setting 3

N	Naive Normal	Naive Bootstrap	Adaptive Normal	Adaptive Bootstrap
100	0.99	0.99	0.955	0.95
200	0.9975	0.9975	0.955	0.9475
400	1	0.9975	0.96	0.955
800	1	1	0.9525	0.9475

naive bootstrap methods along with the adaptive normal and adaptive bootstrap methods. Figure 1 shows the length of CIs using all four methods.

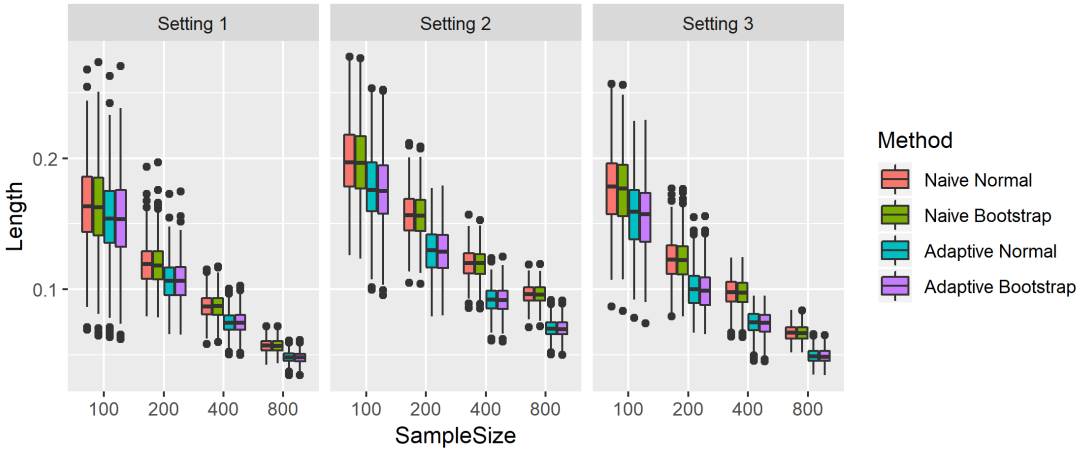

FIG 1. *Boxplots of the length of the confidence intervals of Settings 1-3.*

Motivated by the empirical study given in Laber and Murphy (2012), we evaluate the proposed methodology on four datasets taken from the UCI machine learning repository[1]. For each of the datasets the CIs were constructed for three different sample sizes 30, 100, and 250. For each dataset, and each sample size, a random sample from the original dataset was drawn. A classifier was constructed based on the drawn sampled data. Normal approximation and empirical bootstrap CIs with a confidence level of 95% were calculated using both the naive and adaptive approaches.

[1] www.ics.uci.edu/ mlearn/MLRepository.html

TABLE 2
The coverage of the four methods for the datasets Heart, Ion, Liver, and Spam.

Heart

N	Naive Normal	Naive Bootstrap	Adaptive Normal	Adaptive Bootstrap
30	0.975	0.9675	0.945	0.945
100	0.9625	0.955	0.9425	0.94
250	0.975	0.965	0.975	0.965

Ion

N	Naive Normal	Naive Bootstrap	Adaptive Normal	Adaptive Bootstrap
30	0.9825	0.96	0.9525	0.9425
100	0.9675	0.9625	0.9475	0.9425
250	1	0.9975	0.9625	0.93

Liver

N	Naive Normal	Naive Bootstrap	Adaptive Normal	Adaptive Bootstrap
30	0.985	0.98	0.9525	0.9575
100	0.9725	0.9725	0.96	0.96
250	0.97	0.97	0.9525	0.9325

Spam

N	Naive Normal	Naive Bootstrap	Adaptive Normal	Adaptive Bootstrap
30	0.9575	0.9575	0.9575	0.9575
100	0.975	0.97	0.9675	0.96
250	0.975	0.975	0.955	0.95

The true test error was approximated using the observations from the original dataset, as the empirical misclassification rate. The process was repeated 400 times. Table 2 and Figure 2 show the coverage and CIs' length, respectively for the four datasets. The heart data set consists 296 observations, the ion data set consists 351 observations, the liver data set consists 345 observations and the spam data set consists 4601 observations.

As can be seen, from both the simulation study and the empirical study, the coverage of the adaptive CIs is very close to the confidence level (of 95%). The naive CIs resulted in more conservative coverage than the adaptive CI and with longer confidence intervals. There is no major difference between the CIs based on normal approximation and CIs based on empirical bootstrap.

6. Concluding remarks. In this work we addressed the challenge of estimating the test error of a classifier which belongs to an infinite-dimensional RKHS. The test error was not estimated directly and instead we constructed CIs for the test error. We proposed two approaches for estimating the CIs, namely, the naive approach and the adaptive approach. For each approach, estimation of the CI was done using both normal approximation and empirical bootstrap methods. The simulation comparing the naive and the adaptive methods showed that the adaptive method achieves a good coverage level while the naive method results in more conservative CIs. The challenge of constructing these CIs has led to a theoretical understanding in various fields about the conver-

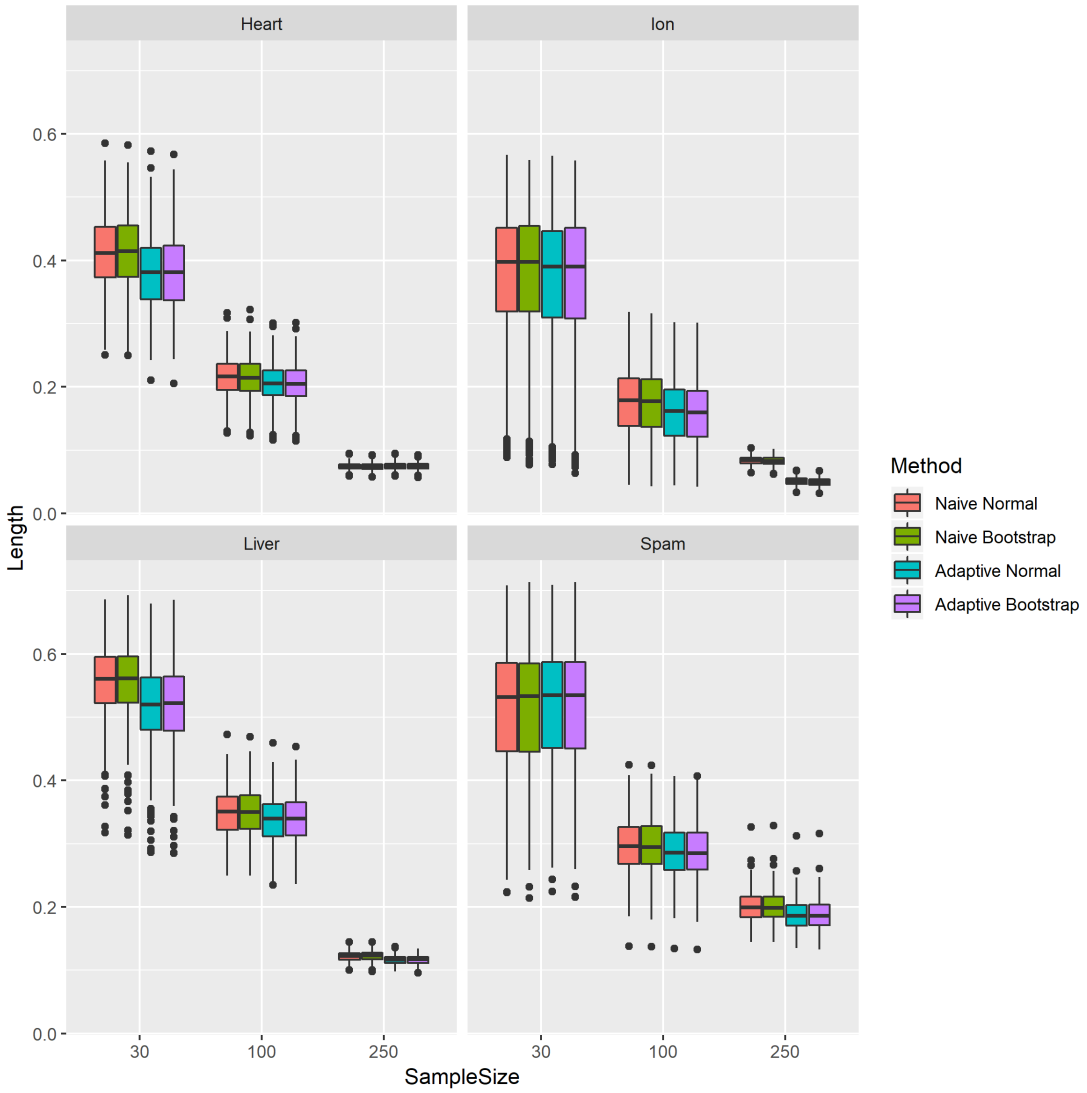

FIG 2. *Boxplots of the length of the confidence intervals for datasets Heart, Ion, Liver, and Spam.*

gence of empirical processes that are indexed by a class of functions belonging to RKHS which can be used in future research.

In this paper we estimated the test error which is the probability of the classifier's misclassification over all the sample space. Future research includes estimating the probability of misclassification over a subset $S \subset \mathcal{X}$. When the the subset S is a specific point, this quantity is interesting, since the obtained confidence interval for the test error is actually a confidence interval for point estimation. Other subsets of interest are subsets in which the classifier is likely to classify well or to be unreliable.

References.

E. Bolthausen, E. Perkins, and A. W. van der Vaart. *Lectures on Probability Theory and Statistics: Ecole D'Eté de Probabilités de Saint-Flour XXIX-1999*. Springer Science & Business Media, 2002.

M. R. Chernick, V. K. Murthy, and C. D. Nealy. Application of bootstrap and other resampling techniques: Evaluation of classifier performance. *Pattern Recognition Letters*, 3(3):167–178, 1985.

B. Efron. Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association*, 78(382):316–331, 1983.

B. Efron and R. Tibshirani. Improvements on cross-validation: The 632+ bootstrap method. *Journal of the American Statistical Association*, 92(438):548–560, 1997.

B. Efron and R. J. Tibshirani. *Cross-validation and the bootstrap: Estimating the error rate of a prediction rule*. Division of Biostatistics, Stanford University, 1995.

R. Hable. Asymptotic normality of support vector machine variants and other regularized kernel methods. *Journal of Multivariate Analysis*, 106:92–117, 2012.

T. Hastie, R. Tibshirani, and J. Friedman. Overview of supervised learning. In *The Elements of Statistical Learning*, pages 9–41. Springer, 2009.

A. Isaksson, M. Wallman, H. Göransson, and M. G. Gustafsson. Cross-validation and bootstrapping are unreliable in small sample classification. *Pattern Recognition Letters*, 29(14):1960–1965, 2008.

B. Jiang, X. Zhang, and T. Cai. Estimating the confidence interval for prediction errors of support vector machine classifiers. *Journal of Machine Learning Research*, 9(Mar):521–540, 2008.

R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Artificial Intelligence*, volume 14, pages 1137–1145. Stanford, CA, 1995.

M. R. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Springer, New York, 2008.

W. J. Krzanowski and D. J. Hand. Assessing error rate estimators: The leave-one-out method reconsidered. *Australian & New Zealand Journal of Statistics*, 39:35–46, 1997.

E. B. Laber and S. A. Murphy. Adaptive confidence intervals for the test error in classification. *Journal of the American Statistical Association*, 2012.

R. A. Schiavo and D. J. Hand. Ten more years of error rate research. *International Statistical Review*, 68(3):295–310, 2000.

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, 2008.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000.

V. N. Vapnik. *Statistical Learning Theory*, volume 1. Wiley, 1998.

P. Zhang. Assessing prediction error in non-parametric regression. *Scandinavian Journal of Statistics*, 1995.

3. Self-reporting and screening

Status: This paper was submitted to Biometrics

SELF-REPORTING AND SCREENING: DATA WITH CURRENT-STATUS AND CENSORED OBSERVATIONS

By Jonathan Yefenof[1] Yair Goldberg[2] Jennifer Wiler[3] Avishai Mandelbaum[2] and Ya'acov Ritov[1,4]

[1] *The Hebrew University of Jerusalem*

[2] *Technion - Israel Institute of Technology*

[3] *University of Colorado*

[4] *University of Michigan*

We consider survival data that combine three types of observations: uncensored, right-censored, and left-censored. Such data arises from screening a medical condition, in situations where self-detection arises naturally. Our goal is to estimate the failure-time distribution, based on these three observation types. We propose a novel methodology for distribution estimation using both parametric and nonparametric techniques. We then evaluate the performance of these estimators via simulated data. Finally, as a case study, we estimate the patience of patients who arrive at an emergency department and wait for treatment. Three categories of patients are observed: those who leave the system and announce it, and thus their patience time is observed; those who get service and thus their patience time is right-censored by the waiting time; and those who leave the system without announcing it. For this third category, the patients' absence is revealed only when they are called to service, which is after they have already left; formally, their patience time is left-censored. Other applications of our proposed methodology are discussed.

1. Introduction. We study the estimation of failure time distribution where the failure times can be either observed directly, or be right-censored or left-censored. This type of survival data arises, for example, in estimation of time to the appearance of a medical condition where characteristic symptoms may or may not appear when the condition exists.

Keywords and phrases: Current status data, Left and Right-censoring, Nonparametric estimation, Survival analysis, Left without being seen

Specific medical settings include relapse in childhood brain tumors, which may be observed due to clinical symptoms, or right-censored due to periodic screening with negative result (no tumor), or left-censored due to periodic screening with a positive result (Minn et al., 2001). Another medical setting is melanoma cancer, which is observed if self-detected, or is right censored due to a negative screening (no melanoma), or left-censored if it goes undetected until screening. Additional examples can be found in Whitehead (1989).

The motivating example for this work comes from estimating customer patience in service system which, as discussed by Mandelbaum and Zeltyn (2007), is a challenging problem. In our study, we focus on patients who wait for treatment in an emergency department (ED). Three categories of patients are observed. The first category consists of patients who get service and thus their patience time is right-censored by the waiting time. The second category comprises those who leave the system and announce it, and thus their patience time is observed while the waiting time is right-censored. The third category consists of patients who leave the system without announcing it; their absence is hence revealed only when they are called to service, which is after they have already left; formally, their patience time is left-censored.

Estimating the patience time is of importance as the decision of patients to leave the system before getting served might have a strong effect on their physical well-being. There has been considerable research on the reasons why patients leave an ED before being served; see Baker et al. (1991), Hunt et al. (2006), Bolandifar et al. (2014), and Batt and Terwiesch (2015). However, these and other authors have not proposed a model by which ED patience time - namely the duration that a potential patient is willing to wait for ED service - can be estimated, and this is our goal here.

We propose novel parametric and nonparametric estimators of the unknown survival function for this 3-type survival data. We then study their rates of convergence. The parametric estimator is based on both full and partial likelihoods. We provide condition under which the parametric estimator is a linear asymptotic normal (LAN) estimator and converges to a normal distribution in a root-n rate. The nonparametric estimator is based on nonparametric kernel estimators for density functions and on a novel estimator of the cumulative probability function that has some similarities to the Nelson–Aalen estimator (e.g., Klein and Moeschberger, 2013, Chapter 4). We show that, under some regularity

conditions, the nonparametric estimator point-wise converges to the normal distribution.

We perform a simulation study and compare the proposed parametric and nonparametric estimators. For the parametric model, we study both correct and misspecified models and show the different corresponding results. We show how the accuracy changes with sample size. We then carry out a case study that is based on data of patients waiting for treatment in an ED, in the U.S. in 2008. We analyzed separately different severity levels (15106 observations in the emergency group, 43600 in the urgent group, and 26541 in the semi-urgent group). We conclude with a comparison of the parametric and nonparametric estimators for the three different severity levels of this dataset.

2. Brief Literature Review. Developing screening methods for medical conditions, such as breast and melanoma cancers, has a long history (Wilson et al., 1968; Zelen and Feinleib, 1969). In the classical setting, the medical condition either already exists at the time of screening and is thus left-censored, or does not exist, and is thus right-censored. The setting in which self-detection is possible, and thus the condition time is observed, has been surprisingly mostly ignored in the literature. For example, Minn et al. (2001) treat both self-detection times and screening times as event times, ignoring the censoring. The closest model to the one that we present here appears in Whitehead (1989). It is assumed there that the condition can be detected at screening or before screening due to symptoms. In both cases, the condition already exists at the time of detection. It is also assumed that screenings take place at a sequence of fixed time points. Whitehead (1989) recommends to ignore the extra knowledge gained due to self-reporting and to replace these times with the time of the next screening. The survival function is then estimated only at the discrete fixed screening times using standard techniques (Prentice and Gloeckler, 1978).

There has been considerable research effort, dedicated to modeling and analysis of customer (im)patience while waiting for service. Here we describe several papers that, together with references therein, provide what is required for a historical background and state-of-art perspective. First, we recommend the literature review (Section 3) in the recent Batt and Terwiesch (2015), accompanied by Gans et al. (2003): these survey patience-research from an operational/queueing view point (mainly Section 6.3.3 in the latter), while connecting it to the medical literature on patients who are left without being seen (LWBS) (mainly Section 3 in the former); see also Aksin et al. (2007) who, relative to

Gans et al. (2003), expand on managerial challenges. Next we mention Mandelbaum and Zeltyn (2013), which is an Explanatory Data Analysis of (im)patience in telephone call centers (that appears in a special issue that is devoted to models of queues abandonment). Finally, and the most related to the present study, are the following two studies. Brown et al. (2005) applies, in Section 5, the Kaplan–Meier estimator (Kaplan and Meier, 1958) to estimate the survival functions and consequently hazard rates, of both virtual waiting time and impatience; the data is that of a call center, in which times of abandonment are all recorded hence the data is right-censored. Then Wiler et al. (2013), which is also the source of our present ED data case study, estimate LWBS rates as a function of ED patient arrival rates, treatment times, and ED boarding times. There was no attempt in that work to estimate the patience-time distribution.

We conclude this brief survey with the observation that the estimation of customer (im)patience is relevant beyond screening, call centers, and EDs. For example, Nah (2004) studies tolerance of Web users (during information retrieval). Yom-Tov et al. (2018) analyzes chat services, in which customers abandon at any phase during chat-exchanges with a service center: one expects that such services give rise to the same options as in EDs: some customers receive service, others abandon without letting anyone know, and the rest announce their abandonment time.

3. The Model. In the standard setting of right-censored data one observes, for each patient, either the failure time or the censoring time. In terms of our motivating example, failure time is patience time while censoring time is the waiting time. Patience time is observed when patients leave the ED while informing the system of their departure; waiting time is observed when a patient is called for service. However, unlike in standard right-censored data and like in current status data, there are also patients who leave without informing; in this case their absence is observed only when they are called for service, and this latter time provides an upper bound for their patience time. In other words, the (virtual) waiting time is observed, and the only information on patience time is that it is less than this observed waiting time. Hence, in this case, the patience time is left-censored.

More formally, let T be the patient's failure time, i.e., the time until the patient loses patience. Let W be the censoring time, i.e., the waiting time until the patient gets (or could have gotten) service. We assume that T has a cumulative distribution function (cdf)

F and a probability density function (pdf) f, and that W has cdf G and pdf g. Let Δ be the indicator $\Delta \equiv 1\{T < W\}$; i.e., $\Delta = 1$ if the patient loses patience before being called to service, and $\Delta = 0$ otherwise.

Let Y be the indicator that is 1 for a patient who leaves and informs when leaving, and 0 otherwise. Denote by $q(t)$ the conditional probability that a patient reports leaving given that the waiting time equals to t. In other words, $q(t) = pr(Y = 1 \mid T = t)$. We assume that the waiting time W and the patience time T are independent. This assumption, which is common in the right-censored data literature (see, Klein and Moeschberger, 2013, Chapter 3, pages 65-66), seems appropriate in our case study, as we stratify by acuity levels. We also assume that announcement indicator Y is independent of the waiting time W, as it seems reasonable that the decision of a patient to report when leaving does not depend on the waiting time. Summarizing, we assume that the pair (Y, T) is independent of the waiting time W. When this assumption does not hold, different theoretical tools are needed for a valid estimation.

Let U be the recorded time: $U \equiv YT + (1 - Y)W$. The observed data consist of the triplets (U_i, Y_i, Δ_i), $i = 1, \ldots, n$, and there are three categories of patients:

$\mathcal{C} = 1$**:** The patient gets service, hence the waiting time is observed, which serves as a lower bound on the patience time; thus the patience time is right censored. Formally, $\Delta = 0$, $Y = 0$, and $U = W$.

$\mathcal{C} = 2$**:** The patient leaves without being treated and reports departure. The patience time is thus revealed: $Y = 1$, $\Delta = 1$, and $U = T$.

$\mathcal{C} = 3$**:** The patient leaves without reporting, hence virtual waiting time (the time that the patient would have waited had he stayed in the ED) is observed, which provides an upper bound for the patience time, thus the patience time is left-censored. Formally, $Y = 0$, $\Delta = 1$, and $U = W$.

LEMMA 1. *The following equalities hold:*

i) $pr(U \le t, \mathcal{C} = 1) = \int_0^t g(w)\overline{F}(w)dw.$

ii) $pr(U \le t, \mathcal{C} = 2) = \int_0^t q(w)f(w)\overline{G}(w)dw.$

iii) $pr(U \le t, \mathcal{C} = 3) = \int_0^t g(w) \int_0^w \{1 - q(x)\} f(x)dxdw.$

Here, $\overline{F}(t) = 1 - F(t)$ and $\overline{G}(t) = 1 - G(t)$ are the survival functions of the patience

time and the waiting time, respectively.

See the proof in A.1.

For $i = 1, 2, 3$, we introduce the following sub-stochastic density functions

$$h_i(t) := \frac{d}{dt} pr(U \leq t, \mathcal{C} = i). \tag{1}$$

From Lemma 1 above, we deduce that

$$h_1(t) = g(t)\overline{F}(t), \quad h_2(t) = q(t)f(t)\overline{G}(t), \quad h_3(t) = g(t)\int_0^t \{1 - q(x)\} f(x)dx.$$

Define

$$r_1(t) \equiv \frac{h_1(t)}{pr(W \leq T)}, \quad r_2(t) \equiv \frac{h_2(t)}{pr(Y = 1, W > T)}, \quad r_3(t) \equiv \frac{h_3(t)}{pr(Y = 0, W > T)}. \tag{2}$$

Then r_i is the density function of the observed time U given $\mathcal{C} = i$. Our model assumes that all denominators are positive.

To summarise what is known and what is to be estimated, there are two unknown distributions in our setting, G and F, and we aim to estimate them using both parametric and nonparametric techniques. For each patient, the waiting time is either observed or right censored. If the patient reports and then leaves, the waiting time is longer than the observed patience time. Hence, the waiting time is right-censored. Therefore, parametric and nonparametric estimation for the distribution of waiting time W can be done by standard techniques for right-censored data. However, estimation of the distribution of patience time T, is more complicated and is discussed in Sections 4 and 5.

4. Parametric estimation. Assume now that the distributions of both the patience time and the waiting time belong to some parametric families. More formally, let $\mathcal{F} = \{f(\cdot;\theta), \theta \in \Theta\}$ where $\Theta \subseteq \mathbb{R}^d$, $\mathcal{G} = \{g(\cdot;\gamma), \gamma \in \Gamma\}$ where $\Gamma \subseteq \mathbb{R}^p$. We assume that the density of the patience time can be written as $f(t;\theta_0) \in \mathcal{F}$. We also assume that the density of the waiting time can be written as $g(t;\gamma_0) \in \mathcal{G}$. Write $h_1(t;\theta,\gamma) \equiv g(t;\gamma)\overline{F}(t;\theta)$, and similarly $h_2(t;\theta,\gamma) \equiv q(t)f(t;\theta)\overline{G}(t;\gamma)$ and $h_3(t;\theta,\gamma) \equiv g(t;\gamma)\int_0^t \{1 - q(x)\} f(x;\theta)dx$.

The likelihood of the observed data $D = \{(U_i, Y_i, \Delta_i), i = 1, \ldots, n\}$ can be written in terms of the functions h_1, h_2, and h_3, as follows:

$$L(D;\theta,\gamma) = \prod_{i=1}^{n} \{h_1(U_i;\theta,\gamma)\}^{1-\Delta_i} \{h_2(U_i;\theta,\gamma)\}^{\Delta_i Y_i} \{h_3(U_i;\theta,\gamma)\}^{\Delta_i(1-Y_i)}.$$

Using the explicit representations of h_1, h_2, h_3, we obtain that $L(D;\theta,\gamma)$ is given by

$$\prod_{i=1}^{n}\Big(\left\{g(U_i;\gamma)\overline{F}(U_i;\theta)\right\}^{1-\Delta_i} \left\{q(U_i)f(U_i;\theta)\overline{G}(U_i;\gamma)\right\}^{\Delta_i Y_i} \times \left[g(U_i;\gamma)\int_0^{U_i}\left\{1-q(s)\right\}f(s;\theta)ds\right]^{\Delta_i(1-Y_i)}\Big).$$

The value of γ that maximizes this likelihood is independent of θ. Therefore, a maximum likelihood estimator (MLE) $\widehat{\gamma}_n$ to γ_0 can be constructed from this likelihood. However, maximizing the likelihood with respect to θ is difficult. Even if γ is given or estimated, the maximizer of θ depends on the unknown function $q(t)$. Therefore, we consider the partial likelihood $L_{partial}(D;\theta;\gamma)$ of category $\mathcal{C}=1$,

$$\prod_{i=1}^{n}\left\{\frac{g(U_i;\gamma)\overline{F}(U_i;\theta)}{\int_0^{\infty}g(s;\gamma)\overline{F}(s;\theta)ds}\right\}^{1-\Delta_i}.$$

The value of θ that maximizes this partial likelihood depends on γ. We plug the MLE $\widehat{\gamma}_n$ into this partial likelihood. In Theorem 1 below we show that, under standard regularity conditions, the maximizer $\widehat{\theta}_n$ of $L_{partial}(D;\theta;\widehat{\gamma}_n)$ is a consistent and asymptotically normal estimator for θ_0.

We need the following assumptions:

(A1) The derivative $\frac{\partial}{\partial\theta}f(t;\theta)$ is continuous in t for each $\theta\in\Theta$, $\frac{\partial}{\partial\gamma}g(t;\gamma)$ is continuous in t for each $\gamma\in\Gamma$.

(A2) For all $\theta\in\Theta$, $\arg\max_{\gamma\in\Gamma}L(D;\theta,\gamma)$ is unique, hence denote $\widehat{\gamma}(\theta)\equiv\arg\max_{\gamma\in\Gamma}L(D;\theta,\gamma)$. It is assumed as well that for each $\theta\in\Theta$, $\frac{\partial}{\partial\gamma}L\left\{D;\theta,\widehat{\gamma}(\theta)\right\}=0$.

(A3) For all $\gamma\in\Gamma$, $\arg\max_{\theta\in\Theta}L_{partial}(D;\theta,\gamma)$ is unique, hence denote $\widehat{\theta}(\gamma)\equiv\arg\max_{\theta\in\Theta}L_{partial}(D;\theta,\gamma)$. It is assumed as well that for each $\gamma\in\Gamma$, $\frac{\partial}{\partial\theta}L_{partial}\left\{D;\widehat{\theta}(\gamma),\gamma\right\}=0$.

THEOREM 1. *Let $\widehat{\gamma}_n$ be the maximizer of $L(D;\theta;\gamma)$ and let $\widehat{\theta}_n$ be the maximizer of $L_{partial}(D;\theta;\widehat{\gamma}_n)$. Then, as $n\to\infty$,*

i) $\widehat{\gamma}_n\to\gamma_0$ in probability.

ii) $\sqrt{n}\left(\widehat{\gamma}_n-\gamma_0\right)\to N\left(0,V_{\gamma_0}\right)$ in distribution.

iii) $\widehat{\theta}_n\to\theta_0$ in probability.

iv) $\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) \to N\left(0, S_{\theta_0,\gamma_0}\right)$ *in distribution.*

Here V_{γ_0}, S_{θ_0}, γ_0 *are covariance matrices as defined in Appendix A.1.*

The proof appears in Appendix A.1.

EXAMPLE 1. *Assume that* T *follows an exponential distribution with rate* θ *and* W *follows an exponential distribution with rate* γ. *Then*

$$\widehat{\gamma}_n = \frac{n - \Sigma_{i=1}^n \Delta_i Y_i}{\Sigma_{i=1}^n U_i}$$
$$\widehat{\theta}_n = \frac{\Sigma_{i=1}^n (1-\Delta_i)}{\Sigma_{i=1}^n U_i (1-\Delta_i)} - \widehat{\gamma}_n = \frac{\Sigma_{i=1}^n (1-\Delta_i)}{\Sigma_{i=1}^n U_i (1-\Delta_i)} - \frac{n - \Sigma_{i=1}^n \Delta_i Y_i}{\Sigma_{i=1}^n U_i}..$$

The details of the computation appears in Appendix A.5

5. Nonparametric estimation. In this section we propose nonparametric estimators for the survival function of the patience time $\overline{F}$ and study its theoretical properties. For simplicity, we restrict the estimation to an interval $[0, \tau]$ for some $\tau > 0$, such that the probability of W and T being larger than τ is positive. This is a standard condition in survival estimation (see Kosorok, 2008, Chapter 4.2). Note that for observations of Categories 1 and 3, the waiting-time is observed. For Category 2, only a lower bound of the waiting time is observed. Hence, the waiting time is either observed or right-censored. Therefore, estimating the waiting time distribution can be done by using standard survival analysis estimators such as the Kaplan–Meyer estimator (see Klein and Moeschberger, 2013, Chapter 4). On the other hand, estimating the distribution of the patience time is more challenging since we cannot distinguish between the density function f and the unknown function q. Our goal is thus to estimate the distribution of the patience time F.

Assume that over all positive numbers, the waiting time density function g is strictly positive. Recall that $h_1(t) = g(t)\overline{F}(t)$, $h_3(t) = g(t)\int_0^t \{1 - q(s)\} f(s)ds$, where the functions h_1, h_3 are defined as in (1). Therefore,

$$\frac{h_1(t)}{h_3(t)} = \frac{\overline{F}(t)}{F(t) - \int_0^t q(s)f(s)ds}. \tag{3}$$

which is well defined as $g(t) > 0$. Reordering the terms in (3), we get that

$$\left\{F(t) - \int_0^t q(s)f(s)ds\right\} \frac{h_1(t)}{h_3(t)} = 1 - F(t).$$

Hence,

$$F(t) = \frac{h_3(t) + h_1(t)\int_0^t q(s)f(s)ds}{h_3(t) + h_1(t)}.$$

From the definitions in (2), it follows that

$$F(t) = \frac{pr(Y=0, T<W)r_3(t) + pr(W \leq T)r_1(t)\int_0^t q(s)f(s)ds}{pr(Y=0, T<W)r_3(t) + pr(W \leq T)r_1(t)}. \tag{4}$$

Therefore, we propose to estimate $F(t)$ by estimating the following terms:

(i) $pr(W \leq T)$ and $pr(Y=0, T<W)$,

(ii) $r_1(t)$ and $r_3(t)$,

(iii) $A(t) \equiv \int_0^t q(s)f(s)ds$.

Estimating the expression in (i) can be done by the empirical estimators: $\widehat{pr}(T \leq W) = n^{-1}\Sigma_{i=1}^n(1-\Delta_i)$, $\widehat{pr}(Y=0, W<T) = n^{-1}\Sigma_{i=1}^n\Delta_i(1-Y_i)$. These estimators converge, by the central limit theorem (CLT), to $pr(W \leq T)$ and $pr(Y=0, T<W)$, respectively, at the rate of $n^{1/2}$.

Since r_1 and r_3 are density functions, they can be estimated using a kernel estimator (Tsybakov, 2008, Chapter 1.2). Let $\widehat{r}_1$ and $\widehat{r}_3$ be kernel estimators of r_1 and r_3, respectively. Assume that both r_1 and r_3 belong to a Sobolev function class of order β. Then for each $t > 0$, both $\widehat{r}_1(t)$ and $\widehat{r}_3(t)$ converge at a rate of $n^{\beta/(2\beta+1)}$ (see Tsybakov, 2008, Chapter 1.7, for both the definition of a Sobolev class and the proof).

We now turn to estimate the term $A(t) = \int_0^t q(s)f(s)ds$. A nonparametric estimator that we created for this term is defined and proven to be consistent in the following lemma.

LEMMA 2. *Let*

$$\widehat{N}_n(t) \equiv \frac{1}{n}\sum_{i=1}^n Y_i\Delta_i 1\{U_i \leq t\}, \quad \widehat{Y}_n(t) \equiv \frac{1}{n}\sum_{i=1}^n 1\{U_i \geq t\}.$$

Define $\widehat{D}_n(t) \equiv \int_0^t \frac{d\widehat{N}_n(s)}{\widehat{Y}_n(s)}$. *Then* $\widehat{A}(t) \equiv 1 - \exp\left\{-\widehat{D}_n(t)\right\}$ *converges pointwise to* $A(t)$, *at a rate of* $n^{1/2}$, *for every* $t \in [0, \tau]$.

The proof is given in Appendix A.3.

By plugging in the estimators

$$\widehat{pr}(Y=0, W<T),\ \widehat{pr}(T \leq W),\ \widehat{r}_3(t),\ \widehat{r}_1(t),\ \widehat{A}(t),$$

to the equation in (4), we get that

$$\widehat{F}_n(t) = \frac{\widehat{pr}(Y=0, W<T)\widehat{r}_3(t) + \widehat{pr}(T \le W)\widehat{r}_1(t)\widehat{A}(t)}{\widehat{pr}(Y=0, W<T)\widehat{r}_3(t) + \widehat{pr}(T \le W)\widehat{r}_1(t)}, \tag{5}$$

is an estimator of $F(t)$.

THEOREM 2. *The estimator $\hat{F}_n(t)$ converges pointwise to $F(t)$ at a rate of $n^{\beta/(2\beta+1)}$, for every $t \in [0, \tau]$.*

The proof appears in Appendix A.4. Since that $\widehat{F}$ is based on density estimation, it is not necessarily monotonic, we therefore replace it with a monotonic approximation. The monotonic approximation is by taking the cumulative sup.

6. Simulations. We study the performance of both the parametric and nonparametric estimators that were proposed in Sections 4 and 5, respectively. Based on the setting of the case study discussed in Section 7, we consider two simulation settings. In the case study, both the exponential and Weibull distributions seem to fit well the waiting time and patience time distributions, respectively.

The case study data also indicated that the mean of the waiting time W is smaller than the patience time T. Thus, the two simulation settings consist of samples from exponential and Weibull distributions in which the waiting time has a smaller mean then the patience time mean. In the first setting, a sample was taken from the model in which the patience time T follows an exponential distribution with rate 4 (or scale $\frac{1}{4}$), and the waiting time W follows an exponential distribution with rate 10 (or scale $\frac{1}{10}$). In the second setting a sample was taken from a model in which the patience time T follows a Weibull distribution with scale $\frac{1}{4}$ and shape 2, and the waiting time W follows an exponential distribution with rate 10. In both settings, the unknown probability of announcement is $q(t) = \exp(-12t)$. Taking the probability of announcement to be the increasing function $q(t) = 1-\exp(-12t)$ or the constant function $q(t) = 0.5$ yields similar results which are omitted. Moreover, we experimented with additional numerical values. The behavior and conclusions, as reported here, remain consistent across these experiments.

In each setting, we calculated the parametric estimator for the scale of T for five different sample sizes ($N = 100, 200, 500, 1000, 2000$). For each sample size, we repeated the simulation 100 times. When using the parametric method, it was assumed that both T

and W follow an exponential distribution with unknown parameters. Note that this assumption holds for the first setting but does not hold for the second one. In other words, the second setting is carried out under a misspecified model. The results are shown in Figure 1.

We compare $\widehat{\overline{F}}_n$, the estimator of the survival function of T, to the true survival function $\overline{F}_0$. For the parametric estimation, $\widehat{\overline{F}}_n(t) = \exp(-\hat{\theta}t)$, while for the nonparametric estimator $\widehat{\overline{F}}_n(t)$ is given by (5). The comparison is done using mean square error (MSE), which is defined by

$$MSE(\widehat{\overline{F}}_n, \overline{F}_0) \equiv \int_{-\infty}^{\infty} \left\{\widehat{\overline{F}}_n(t) - \overline{F}_0(t)\right\}^2 f_0(t)dt\,,$$

where f_0 is the density of T. The parametric and nonparametric survival function estimators are demonstrated in Figures 2 and 3. Figure 2 represents the results of the first setting in which T follows an exponential distribution with rate 4 and W follows an exponential distribution with rate 10. Figure 3 represents the results of the second setting in which T follows a Weibull distribution with scale $\frac{1}{4}$ and shape 2, and W follows an exponential distribution with rate 10. Summaries of the MSE are given in Table 1. Not surprisingly, for Setting 1, since the parametric model is correct, the MSE is smaller for the parametric estimator. Similarly, since in Setting 2 the parametric model is incorrect, the MSE is smaller for the nonparametric estimator.

7. Case study. Retrospective data were collected from all patient presentations to triage at an urban, academic, adult-only emergency department (ED) with visits in calendar year 2008. This data was used for the analysis in Wiler et al. (2013). The data consist of the waiting time of patients arriving at emergency rooms. One of the categories defined in this data is acuity. Since our model assumes that all patients follow the same distribution, we calculated the estimators for each level of acuity separately. We focused on the following three levels of acuity: emergency, urgent, and semi-urgent. The emergency level consist of 15106 patients, the urgent level consist of 43600 patients, and the semi-urgent level consist of 26541 patients.

The data consists of the triple variables (U_i, Δ_i, Y_i) described in Section 3. At each acuity level, an observation is categorized to one of the three possible categories.

Parametric and nonparametric estimators for the survival of each acuity level were

calculated. The results of these estimators are given in Figures 4 and 5. As can be seen in Figure 4, the nonparametric estimators of the patience time are stochastically ordered by levels of acuity. In other words, patients at the severe acuity level are less probable to lose patience than patients at the urgent level, who in turn are less prone to lose patience than patients at the semi-urgent level. The results for the parametric estimator seem unreasonable since one would expect that patients with more severe acuity level are more likely to lose patience, and lose it faster.

8. Discussion. In this paper, we consider survival data that combine observed, right-censored, and left-censored data. The setting we analyzed was that of patients who wait for treatment in an emergency department, where some patients may leave without being seen. We proposed both parametric and nonparametric estimators for the distribution of the patience time. Using simulation, we showed that when the parametric model holds, the parametric estimator estimates the patience time well. However, when the model is misspecified, the nonparametric estimator behaved better. In our case study, we also observed that the nonparametric estimator performed better.

So far, no baseline covariates were given. Novel parametric and nonparametric estimators are needed for addressing settings that include baseline covariates.

9. Acknowledgement. Y. Ritov was partially supported by the Israeli Science Foundation (grant No. 1770/15).

Y. Goldberg was partially supported by the Israeli Science Foundation (grant No. 849/17).

APPENDIX A: PROOFS

A.1. Proof of Lemma 1.

$$\begin{aligned} pr(U \le t, \mathcal{C} = 1) =& pr(U \le t, W \le T) \\ =& pr(W \le t, W \le T) \\ =& \int_0^t pr(W \le T \mid W = s) g(s) ds \\ =& \int_0^t pr(s \le T) g(s) ds \\ =& \int_0^t g(s) \overline{F}(s) ds, \end{aligned}$$

where in the fourth equality we use the independence between W and (Y,T).

This establishes i). For ii), we have

$$\begin{aligned} pr(U \le t, \mathcal{C}=2) =& pr(U \le t, Y=1, T<W) \\ =& pr(T \le t, Y=1, T<W) \\ =& \int_0^t pr(s<W \mid Y=1, T=s) q(s) f(s) ds \\ =& \int_0^t pr(s<W) q(s) f(s) ds \\ =& \int_0^t q(s) f(s) \overline{G}(s) ds, \end{aligned}$$

where in the fourth equality we use the independence between W and (Y,T).

Finally, for iii),

$$\begin{aligned} P(U \le t, \mathcal{C}=3) =& pr(U \le t, Y=0, T<W) \\ =& pr(W \le t, Y=0, T<W) \\ =& \int_0^t pr(Y=0, T<s \mid W=s) g(s) ds \\ =& \int_0^t g(s) pr(Y=0, T<s) ds \\ =& \int_0^t g(s) \int_0^s pr(Y=0 \mid T=x) f(x) dx ds \\ =& \int_0^t g(s) \int_0^s \{1-q(x)\} f(x) dx ds, \end{aligned}$$

where in the fourth equality we use the independence between W and (Y,T).

A.2. Proof of Theorem 1. The log of the full likelihood is

$$\begin{aligned} l(D;\theta,\gamma) = \sum_{i=1}^n \Bigg[& (1-\Delta_i) \left(\log g(U_i;\gamma) + \log \overline{F}(U_i;\theta) \right) \\ & + \Delta_i Y_i \left(\log q(U_i) + \log f(U_i;\theta) + \log \overline{G}(U_i;\gamma) \right) \\ & + \Delta_i (1-Y_i) \left\{ \log g(U_i;\gamma) + \log \int_0^{U_i} (1-q(s)) f(s;\theta) ds \right\} \Bigg]. \end{aligned}$$

Given the data D,

(6)

$$\frac{1}{n} l(D;\theta,\gamma) = \mathbb{P}_n \left\{ m_\gamma(U,\Delta,Y) + c(U,\Delta,Y;\theta) \right\} \equiv \frac{1}{n} \sum_{i=1}^n \left\{ m_\gamma(U_i,\Delta_i,Y_i) + c(U_i,\Delta_i,Y_i;\theta) \right\}$$

where $m_\gamma : \mathbb{R}^+ \times \{0,1\}^2 \to \mathbb{R}$ is defined by

$$m_\gamma(u,\delta,y) \equiv (1-\delta)\log g(u;\gamma) + \Delta y \log \overline{G}(u;\gamma) + \delta(1-y)\log g(u;\gamma).$$

and

$$\begin{aligned} c(u,\delta,y;\theta) \equiv & (1-\delta)\log \overline{F}(u;\theta) \\ & + \delta y\left(\log q(u) + \log f(u;\theta)\right) + \delta(1-y)\log \int_0^u \{1-q(s)\} f(s;\theta) ds. \end{aligned}$$

From assumption A1 we obtain that, for each $\theta \in \Theta$, $\arg\max_{\gamma\in\Gamma} l(D;\gamma,\theta) = \arg\max_{\gamma\in\Gamma} \mathbb{P}_n(m_\gamma)$. The γ that maximizes $L(D;\theta,\gamma)$ does not depend on the value of θ or the function p. Define $M_n(\gamma) \equiv \mathbb{P}_n m_\gamma$ and $M(\gamma) \equiv P m_\gamma$.If, for a general function h, $Ph \equiv \int h(x)dP(x)$ and $\mathbb{P}_n h \equiv n^{-1}\sum_{i=1}^n h(X_i)$ then by Assumptions A1–A3, Theorem 5.7 in van der Vaart (2000) can be applied. Therefore $\widehat{\gamma}_n \to \gamma_0$, in probability, which concludes the proof of i).

Given the data D, the term $\frac{\partial l(D;\theta,\gamma)}{\partial\gamma}$ is a function of γ and does not depend on the unknown function p. We also have

$$\frac{1}{n}\frac{\partial l(D;\theta,\gamma)}{\partial\gamma} = \mathbb{P}_n\psi_\gamma(U,\Delta,Y) \equiv \frac{1}{n}\sum_{i=1}^n \psi_\gamma(U_i,\Delta_i,Y_i),$$

where $\psi_\gamma : \mathbb{R}^+ \times \{0,1\}^2 \to \mathbb{R}$ is defined as

$$\psi_\gamma(u,\delta,y) \equiv (1-\delta)\frac{\frac{\partial}{\partial\gamma}g(u;\gamma)}{g(u;\gamma)} - \delta y\frac{\frac{\partial}{\partial\gamma}\overline{G}(u;\gamma)}{\overline{G}(u;\gamma)} + \delta(1-y)\frac{\frac{\partial}{\partial\gamma}g(u;\gamma)}{g(u;\gamma)}.$$

By Assumptions (A1)–(A3), ψ_γ satisfies the conditions of Theorem 5.41 in van der Vaart (2000) and, therefore,

$$\sqrt{n}(\hat{\gamma}_n - \gamma_0) = -(P\dot{\psi}_{\gamma_0})^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \psi_{\gamma_0}(U_i,\Delta_i,Y_i) + o_p(1),$$

where $\dot{\psi}_\gamma(x) = \frac{\partial}{\partial\gamma}\psi_\gamma(x)$. Hence $\hat{\gamma}_n$ is a linear asymptotically normal (LAN) estimator with influence function $\varphi \equiv -(P\dot{\psi}_{\gamma_0})^{-1}\psi_{\gamma_0}$. From all of the above we get that ii) is proved with $V_{\gamma_0} = (P\dot{\psi}_{\gamma_0})^{-1}P\psi_{\gamma_0}\psi_{\gamma_0}^t(P\dot{\psi}_{\gamma_0})^{-1}$.

To prove iii), note that due to the term $\log\int_0^{U_i}(1-q(s))f(s;\theta)ds$ that appears in $l(D;\theta,\gamma)$, the term $\frac{\partial l(D;\theta,\gamma)}{\partial\theta}$ depends on the unknown function p. We therefore consider a partial likelihood function such that its derivate with respect to θ does not depend on p. The partial likelihood that satisfies this request is the partial likelihood of $\mathcal{C} = 1$:

$$\prod_{i=1}^n \left\{\frac{g(U_i;\gamma)\overline{F}(U_i;\theta)}{\int_0^\infty g(s;\gamma)\overline{F}(s;\theta)ds}\right\}^{1-\Delta_i}$$

The log of the partial likelihood is

$$l_{partial}(D;\theta,\gamma) = \sum_{i=1}^{n}(1-\Delta_i)\left\{\log g(U_i;\gamma) + \log \overline{F}(U_i;\theta) - \log \int_0^\infty g(s;\gamma)\overline{F}(s;\theta)ds\right\}.$$

Given the data D, the term $l_{partial}(D;\theta,\gamma)$ is a function only of the parameters θ and γ. We also have

$$\frac{1}{n} l_{parital}(D;\theta,\gamma) = \mathbb{P}_n r_{\theta,\gamma}(U,\Delta,Y) \equiv \frac{1}{n}\sum_{i=1}^{n} r_{\theta,\gamma}(U_i,\Delta_i,Y_i),$$

where $r_{\theta,\gamma} : \mathbb{R}^+ \times \{0,1\}^2 \to \mathbb{R}$ is given by

$$r_{\theta,\gamma}(U,\Delta,Y) \equiv (1-\Delta)\left\{\log g(U;\gamma) + \log \overline{F}(U;\theta) - \log \int_0^\infty g(s;\gamma)\overline{F}(s;\theta)ds\right\}.$$

Define $M_n(\theta,\gamma) \equiv \mathbb{P}_n r_{\theta,\gamma}$, and $M(\theta,\gamma) \equiv P r_{\theta,\gamma}$. Then, Theorem 5.7 in van der Vaart (2000) can be applied. Therefore $\left(\widehat{\theta}_n, \widehat{\gamma}_n\right) \to (\theta_0,\gamma_0)$ in probability, and in particular $\widehat{\theta}_n \to \theta_0$ in probability, and iii) is proven.

In order to prove iv), note that

$$\frac{1}{n}\frac{\partial l_{partial}(D;\theta,\gamma)}{\partial \gamma} = \mathbb{P}_n \phi_{\theta,\gamma}(U,\Delta,Y) \equiv \frac{1}{n}\sum_{i=1}^{n}\phi_{\theta,\gamma}(U_i,\Delta_i,Y_i),$$

where $\phi_{\theta,\gamma} : \mathbb{R}^+ \times \{0,1\}^2 \to \mathbb{R}$ is defined as

$$\phi_{\theta,\gamma}(u,\delta,y) \equiv (1-\delta)\left\{\frac{\frac{\partial}{\partial\theta}\overline{F}(u;\theta)}{\overline{F}(u;\theta)} - \frac{\frac{\partial}{\partial\theta}\int_0^\infty g(s;\gamma)\overline{F}(s;\theta)ds}{\int_0^\infty g(s;\gamma)\overline{F}(s;\theta)ds}\right\}.$$

Using Assumptions A1–A3, together from Theorem 5.41 in van der Vaart (2000), we obtain that

$$\sqrt{n}(\hat{\gamma}_n - \gamma_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{-(P\dot{\psi}_{\gamma_0})^{-1}\psi_{\gamma_0}(U_i,\Delta_i,Y_i)\right\} + o_p(1).$$

Define $\Phi_n(\theta,\gamma) \equiv \mathbb{P}_n\phi_{\theta,\gamma}$ and note that $\Phi(\theta_0,\gamma_0) \equiv P\phi_{\theta_0,\gamma_0} = 0$ (since under the true parameters $P\phi_{\theta_0,\gamma_0} = \frac{\partial}{\partial\theta}\int dh_0 = \frac{\partial}{\partial\theta}1 = 0$, where dh_0 is the true distribution of category 1).

By Talyor's theorem,

$$0 = \Phi_n(\widehat{\theta}_n, \widehat{\gamma}_n) = \Phi_n(\theta_0, \gamma_0) + \left\{ \frac{\partial}{\partial \theta} \Phi_n(\theta_0, \gamma_0) \right\}^T (\widehat{\theta}_n - \theta_0) + \left\{ \frac{\partial}{\partial \phi} \Phi_n(\theta_0, \gamma_0) \right\}^T (\widehat{\gamma}_n - \gamma_0) + o_p\left(n^{-1/2}\right)$$

$$\Rightarrow 0 = \sqrt{n}\Phi_n(\theta_0, \gamma_0) + \left\{ \frac{\partial}{\partial \theta} \Phi_n(\theta_0, \gamma_0) \right\}^T \sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) + \left\{ \frac{\partial}{\partial \gamma} \Phi_n(\theta_0, \gamma_0) \right\}^T \sqrt{n}\left\{\widehat{\gamma}_n - \gamma_0\right\} + o_p(1).$$

$$= \sqrt{n}\Phi_n(\theta_0, \gamma_0) + \left\{ \frac{\partial}{\partial \theta} \Phi_n(\theta_0, \gamma_0) \right\}^T \sqrt{n}\left(\widehat{\theta}_n - \theta_0\right)$$

$$- \left\{ \frac{\partial}{\partial \gamma} \Phi_n(\theta_0, \gamma_0) \right\}^T (P\dot{\psi}_{\gamma_0})^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_{\gamma_0}(U_i, \Delta_i, Y_i) + o_p(1)$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left[\phi_{\theta_0,\gamma_0}(U_i, \Delta_i, Y_i) - \left\{ \frac{\partial}{\partial \gamma} \Phi_n(\theta_0, \gamma_0) \right\}^T (P\dot{\psi}_{\gamma_0})^{-1} \psi_{\gamma_0}(U_i, \Delta_i, Y_i) \right]$$

$$+ \left\{ \frac{\partial}{\partial \theta} \Phi_n(\theta_0, \gamma_0) \right\}^T \sqrt{n}(\widehat{\theta}_n - \theta_0) + o_p(1).$$

Elementary arithmetic leads to

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) = -\left(EE^T\right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ \phi_{\theta_0,\gamma_0}(U_i, \Delta_i, Y_i) - B^T (P\dot{\psi}_{\gamma_0})^{-1} \psi_{\gamma_0}(U_i, \Delta_i, Y_i) \right\} + o_p(1),$$

where $B \equiv \frac{\partial}{\partial \gamma}\Phi(\theta_0, \gamma_0)$, and $E \equiv \frac{\partial}{\partial \theta}\Phi(\theta_0, \gamma_0)$. Hence, $\widehat{\theta}_n$ is a LAN estimator with the influence function

$$\varphi = -\left(EE^T\right)^{-1} \frac{1}{\sqrt{n}} \left\{ \phi_{\theta_0,\gamma_0} - B^T (P\dot{\psi}_{\gamma_0})^{-1} \psi_{\gamma_0} \right\}.$$

Summarizing,

$\sqrt{n}(\widehat{\theta}_n - \theta_0) \to N(0, S_{\theta_0,\gamma_0})$ in distribution, where $S_{\theta_0,\gamma_0} = P\varphi\varphi^T$, hence, iv) is proven.

A.3. Proof of Lemma 2.

Proof. We use similar arguments to those in the proof of the convergence of the Nelson–Aalen estimator to a cumulative hazard function (see Kosorok, 2008, page 240). Hence we have that

$$\sqrt{n} \begin{Bmatrix} \widehat{N}_n(t) - N(t) \\ \widehat{Y}_n(t) - Y(t) \end{Bmatrix} = O_p(1),$$

where $N(t) = pr(Y = 1, T \le W, T \le t)$ and $Y(t) = pr(U \ge t)$.

Since, by Section 3,

$$\{U > t\} = \{Y = 0, W > t\} \cup \{Y = 1, W > t, T > t\}.$$

Hence,

$$\begin{aligned} pr(U > t) &= pr(Y = 0, W > t) + pr(Y = 1, W > t, T > t) \\ &= pr(W > t)\left\{pr(Y = 0) + pr(Y = 1, T > t)\right\} \\ &= pr(W > t)\big(pr(Y = 0) + pr(Y = 1) - pr(Y = 1, T \leq t)\big) \\ &= pr(W > t)\left\{1 - pr(Y = 1, T \leq t)\right\} \\ &= \overline{G}(t)\left\{1 - \int_0^t q(s)f(s)ds\right\}, \end{aligned}$$

where in the second equality we use the independence between W and (Y, T).

By Lemma 1

$$pr(Y = 1, T \leq W, T \leq t) = \int_0^t q(s)f(s)\overline{G}(s)ds.$$

Using the continuity of the derivative operator and of the integral operator, we get that

$$\widehat{D}_n(t) \to \int_0^t \frac{q(s)f(s)\overline{G}(s)}{\overline{G}(s)\left\{1 - \int_0^s q(x)f(x)ds\right\}}ds$$

in probability.

Note that

$$\begin{aligned} &\int_0^t \frac{\overline{G}(s)q(s)f(s)}{\overline{G}(s)\left\{1 - \int_0^s q(x)f(x)dx\right\}}ds = \int_0^t \frac{q(s)f(s)}{\left\{1 - \int_0^s q(x)f(x)dx\right\}}ds \\ &= -\int_0^t \frac{\partial}{\partial s}\log\left\{1 - \int_0^s q(x)f(x)dx\right\}ds = -\log\left\{1 - \int_0^t q(s)f(s)ds\right\}. \end{aligned}$$

Hence, by the delta method (see Kosorok, 2008, Chapter 12.2.2.2), we get that

$$\widehat{D}_n(t) \to -\log\left\{1 - \int_0^t q(s)f(s)ds\right\}$$

in probability, with convergence at rate $n^{1/2}$.

Since $y = -\log(1 - x) \Leftrightarrow x = 1 - \exp(-y)$ and by the continuous mapping theorem (see Kosorok, 2008, Theorem 7.7), we get that $\widehat{A}(t) = 1 - \exp(-\widehat{D}_n(t))$ is an estimator of $\int_0^t q(s)f(s)ds$, at the rate of $n^{1/2}$ as desired. □

A.4. Proof of Theorem 2. For the proof of Theorem 2, we need the following lemma, which is elementary hence stated without proof.

LEMMA 3. *Let* $(a_n)_{n=1}^{\infty}, (b_{n]})_{n=1}^{\infty}$ *be positive sequences. If* $X_n - X = O_p(a_n)$ *and* $Y_n - Y = O_p(b_n)$, *as well as* $P(|X| > l) = 1$ *for some* $l > 0$. *Then we have:*

i) $X_n + Y_n - (X + Y) = O_p(a_n \vee b_n)$,

ii) $X_n Y_n - XY = O_p(a_n \vee b_n)$,

iii) $\frac{1}{X_n} - \frac{1}{X} = O_p(a_n)$.

PROOF OF THEOREM 2. Recall that

$$\widehat{F}_n(t) = \frac{\widehat{pr}(Y = 0, W < T)\widehat{r}_3(t) + \widehat{pr}(T \leq W)\widehat{r}_1(t)\widehat{A}(t)}{\widehat{pr}(Y = 0, W < T)\widehat{r}_3(t) + \widehat{pr}(T \leq W)\widehat{r}_1(t)}$$

is an estimator of $F(t)$.

For all $t > 0$,

$$\widehat{pr}(Y = 0, W < T) - pr(Y = 0, W < T) = O_p(n^{-1/2}),$$

and

$$\widehat{pr}(T \leq W) - pr(T \leq W) = O_p(n^{-1/2}),$$

as both are empirical distribution estimators. By Chapter 1.7 of Tsybakov (2008), for all $t > 0$, $\widehat{r}_j(t) - r_j(t) = O_p(n^{-\beta/(2\beta+1)})$, for $j = 1, 3$. By Lemma 2, $\widehat{A}(t) - A(t) = O_p(n^{-1/2})$.

By Lemma 3.ii)

$$\widehat{pr}(Y = 0, W < T)\widehat{r}_3(t) - pr(Y = 0, W < T)\widehat{r}_3(t) = O_p(n^{-\beta/(2\beta+1)}),$$

and

$$\widehat{pr}(T \leq W)\widehat{r}_1(t)\widehat{A}(t) - pr(T \leq W)\widehat{r}_1(t)A(t) = O_p(n^{-\beta/(2\beta+1)}).$$

Therefore, by Lemma 3.i),

$$\widehat{pr}(Y = 0, W < T)\widehat{r}_3(t) + \widehat{pr}(T \leq W)\widehat{r}_1(t)\widehat{A}(t) - pr(Y = 0, W < T)r_3(t) - pr(T \leq W)r_1(t)A(t)$$
$$= O_p(n^{-\beta/(2\beta+1)}).$$

Similarly,

$$\widehat{pr}(Y=0,W<T)\widehat{r}_3(t)+\widehat{pr}(T\le W)\widehat{r}_1(t)-pr(Y=0,W<T)r_3(t)-pr(T\le W)r_1(t)$$
$$=O_p(n^{-\beta/(2\beta+1)}).$$

By Lemma 3.iii),

$$\frac{1}{\widehat{pr}(Y=0,W<T)\widehat{r}_3(t)+\widehat{pr}(T\le W)\widehat{r}_1(t)}-\frac{1}{pr(Y=0,W<T)r_3(t)+pr(T\le W)r_1(t)}$$
$$=O_p(n^{-\beta/(2\beta+1)}).$$

By Lemma 3.i) again,

$$\frac{\widehat{pr}(Y=0,W<T)\widehat{r}_3(t)+\widehat{pr}(T\le W)\widehat{r}_1(t)\widehat{A}(t)}{\widehat{pr}(Y=0,W<T)\widehat{r}_3(t)+\widehat{pr}(T\le W)\widehat{r}_1(t)}-\frac{pr(Y=0,W<T)r_3(t)+pr(T\le W)r_1(t)A(t)}{pr(Y=0,W<T)r_3(t)+pr(T\le W)r_1(t)}$$
$$=O_p(n^{-\beta/(2\beta+1)}).$$

In other words, $\widehat{F}_n(t)-F(t)=O_p(n^{-\beta/(2\beta+1)})$, which complete the proof of Theorem 2. □

A.5. Details on Example 1. Assuming that T follows an exponential distribution with rate θ and W follows an exponential distribution with rate *gamma* then the likelihood $L(D;\theta,\gamma)$ is

$$\prod_{i=1}^{n}\Big(\{\gamma\exp(-\gamma U_i)\exp(-\theta U_i)\}^{1-\Delta_i}\{q(U_i)\theta\exp(-\theta U_i)\exp(-\gamma U_i)\}^{\Delta_i Y_i}$$
$$\times\Big[\gamma\exp(-\gamma U_i)\int_0^{U_i}\{1-q(s)\}\theta\exp(-\theta s)\Big]^{\Delta_i(1-Y_i)}\Big).$$

Hence,

$$\log L(D;\theta,\gamma)$$
$$=\sum_{i=1}^{n}(1-\Delta_i)(\log\gamma-\gamma U_i)+\sum_{i=1}^{n}\Delta_i Y_i(-\gamma U_i)+\sum_{i=1}^{n}\Delta_i(1-Y_i)(\log\gamma-\gamma U_i)+C(U_1,U_2,...,U_n;\theta)$$
$$=(n-\sum_{i=1}^{n}\Delta_i Y_i)\log\gamma-\gamma\sum_{i=1}^{n}U_i+C(U_1,U_2,...,U_n;\theta).$$

Therefore,

$$\frac{\partial\log \text{Ł}(D;\theta,\gamma)}{\partial\gamma}=\frac{n-\sum_{i=1}^{n}\Delta_i Y_i}{\gamma}-\sum_{i=1}^{n}U_i,$$

and hence, the value of γ which maximize $L(D;\theta,\gamma)$ is $\widehat{\gamma}_n = \frac{n-\sum_{i=1}^n \Delta_i Y_i}{\sum_{i=1}^n U_i}$.

The partial likelihood $L_{partial}(D;\theta;\gamma)$ of category $\mathcal{C}=1$,

$$\prod_{i=1}^n \left\{ \frac{g(U_i;\gamma)\overline{F}(U_i;\theta)}{\int_0^\infty g(s;\gamma)\overline{F}(s;\theta)ds} \right\}^{1-\Delta_i}$$
$$= \prod_{i=1}^n \left\{ \frac{\gamma \exp(-\gamma U_i)\exp(-\theta U_i)}{\int_0^\infty \gamma \exp(-\gamma s)\exp(-\theta s)ds} \right\}^{1-\Delta_i} = \prod_{i=1}^n \left\{ \frac{\gamma \exp(-U_i(\gamma+\theta))}{\int_0^\infty \gamma \exp(-s(\gamma+\theta))ds} \right\}^{1-\Delta_i}$$
$$= \prod_{i=1}^n \left\{ \frac{\gamma \exp(-U_i(\gamma+\theta))}{\frac{\gamma}{\gamma+\theta}} \right\}^{1-\Delta_i} = \prod_{i=1}^n [(\gamma+\theta)\exp(-U_i(\gamma+\theta))]^{1-\Delta_i}.$$

Hence,

$$\log L_{partial}(D;\theta;\gamma) = \sum_{i=1}^n (1-\Delta_i)\left(\log(\gamma+\theta) - U_i(\gamma+\theta)\right).$$

Therefore,

$$\frac{\partial \log Ł_{partial}(D;\theta,\gamma)}{\partial\theta} = \sum_{i=1}^n (1-\Delta_i)(\frac{1}{\gamma+\theta} - U_i).$$

The parametric estimator for θ is the maximizer of $L_{partial}(D;\theta,\widehat{\gamma}_n)$ by θ which is

$$\widehat{\theta}_n = \frac{\sum_{i=1}^n (1-\Delta_i)}{\sum_{i=1}^n U_i(1-\Delta_i)} - \widehat{\gamma}_n = \frac{\sum_{i=1}^n (1-\Delta_i)}{\sum_{i=1}^n U_i(1-\Delta_i)} - \frac{n-\sum_{i=1}^n \Delta_i Y_i}{\sum_{i=1}^n U_i}.$$

REFERENCES

Z. Aksin, M. Armony, and V. Mehrotra. The modern call center: A multi-disciplinary perspective on operations management research. *Production and operations management*, 16(6):665–688, 2007.

D. W. Baker, C. D. Stevens, and R. H. Brook. Patients who leave a public hospital emergency department without being seen by a physician. Causes and consequences. *JAMA*, 266(8):1085–1090, 1991.

R. J. Batt and C. Terwiesch. Waiting patiently: An empirical study of queue abandonment in an emergency department. *Management Science*, 61(1):39–59, 2015.

E. Bolandifar, N. DeHoratius, T. Olsen, and J. L. Wiler. Modeling the behavior of patients who leave the ED without being seen. *Chicago Booth Research Paper*, (12–14), 2014.

L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American statistical association*, 100(469):36–50, 2005.

N. Gans, G. Koole, and A. Mandelbaum. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, 5(2):79–141, 2003.

K. A. Hunt, E. J. Weber, J. A. Showstack, D. C. Colby, and M. L. Callaham. Characteristics of frequent users of emergency departments. *Annals of Emergency Medicine*, 48(1):1–8, 2006.

Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

J. P. Klein and M. L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, New York, 2013.

M. R. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Springer, New York, 2008.

A. Mandelbaum and S. Zeltyn. Service engineering in action: The Palm/Erlang-A queue, with applications to call centers. *In Advances in Services Innovations, D. Spath and K. P. Fahnrich (eds.). Springer, New York,*, pages 17–45, 2007.

A. Mandelbaum and S. Zeltyn. Data-stories about (im) patient customers in tele-queues. *Queueing Systems*, 75(2-4):115–146, 2013.

A Yuriko Minn, Brad H Pollock, Linda Garzarella, Gary V Dahl, Larry E Kun, Jonathan M Ducore, Atsuko Shibata, James Kepner, and Paul G Fisher. Surveillance neuroimaging to detect relapse in childhood brain tumors: a pediatric oncology group study. *Journal of clinical oncology*, 19(21):4135–4140, 2001.

F. Nah. A study on tolerable waiting time: how long are web users willing to wait? *Behaviour & Information Technology*, 23(3):153–163, 2004.

Ross L Prentice and Lynn A Gloeckler. Regression analysis of grouped survival data with application to breast cancer data. *Biometrics*, pages 57–67, 1978.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Science & Business Media, New York, 2008.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, 2000.

John Whitehead. The analysis of relapse clinical trials, with application to a comparison of two ulcer treatments. *Statistics in medicine*, 8(12):1439–1454, 1989.

J. L. Wiler, E. Bolandifar, R. T. Griffey, R. F. Poirier, and T. Olsen. An emergency department patient flow model based on queueing theory principles. *Academic Emergency Medicine*, 20(9):939–946, 2013.

James Maxwell Glover Wilson, Gunnar Jungner, World Health Organization, et al. Principles and practice of screening for disease. 1968.

G. B. Yom-Tov, A. Rafaeli, S. Ashtar, D. Altman, M. Westphal, M. Natapov, and N. Barkay. Customer emotion in chat services: Automatic identification and new insights. *Under review*, 2018.

Marvin Zelen and Manning Feinleib. On the theory of screening for chronic diseases. *Biometrika*, 56(3): 601–614, 1969.

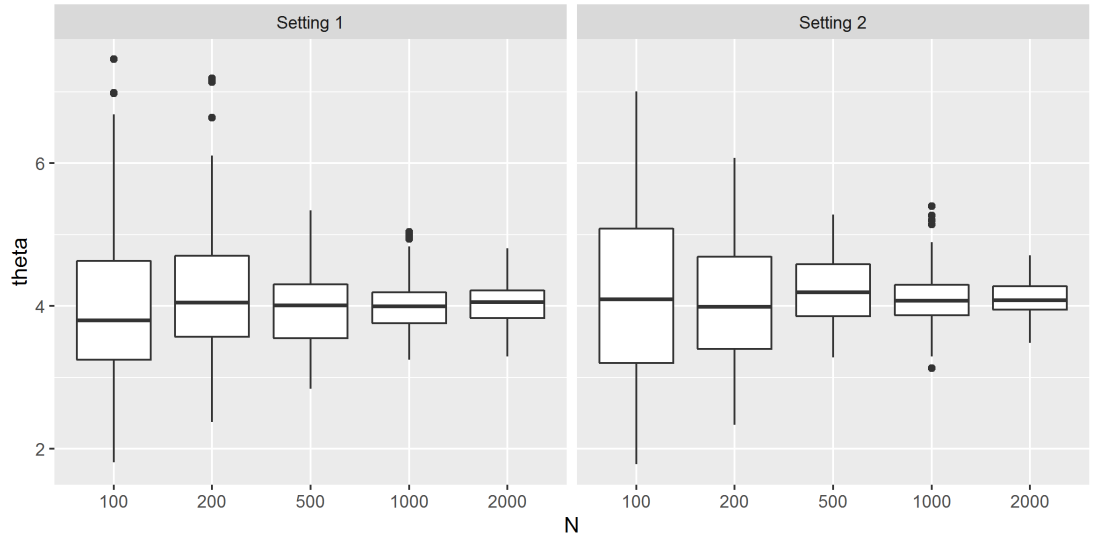

FIG 1. *Parametric estimation. Setting 1:The patience time T follows an exponential distribution with rate* 4 *(scale* $\frac{1}{4}$*) and the waiting time W follows and exponential distribution with rate* 10*, the parametric estimator converges to the true scale* $\frac{1}{4}$*. Setting 2:The patience time T follows a Weibull distribution with scale* $\frac{1}{4}$ *and shape* 2 *while the waiting time W follows an exponential distribution with rate* 10*, the parametric estimator does not converge to the true scale* $\frac{1}{4}$*.*

	Setting 1						Setting 2					
	Parametric			Nonparametric			Parametric			Nonparametric		
N	mean	median	sd	mean	median	sd	mean	median	sd	mean	median	sd
100	6.23	2.79	8.2	12.4	11.02	6.4	23.24	18.44	11.61	13.86	11.59	9.66
200	3.29	1.32	4.94	8.07	7.04	4.4	19.07	15.67	7.87	8.69	7.17	6.43
500	1.31	0.77	1.57	4.5	4.07	1.97	16.53	15.01	4.09	3.72	3.16	2.15
1000	0.72	0.25	0.99	3.24	3.05	1.49	15.38	14.16	3.68	2.14	1.88	1.18
2000	0.36	0.16	0.53	2.29	2.09	1.07	14.8	14.18	1.98	1.55	1.43	0.83

TABLE 1. *MSE for Settings 1 and 2. The table summarizes the MSE that was calculated (100 times) for each of the sample sizes. For Setting 1, the patience time T follows an exponential distribution with rate* 4 *and the waiting time W follows an exponential distribution with rate* 4*. In Setting 2 the patience time T follows a Weibull distribution with scale* $\frac{1}{4}$ *and shape* 2*, while the waiting time W follows an exponential distribution with rate* 10*. As can be seen the nonparametric estimator responded with a lower MSE.*

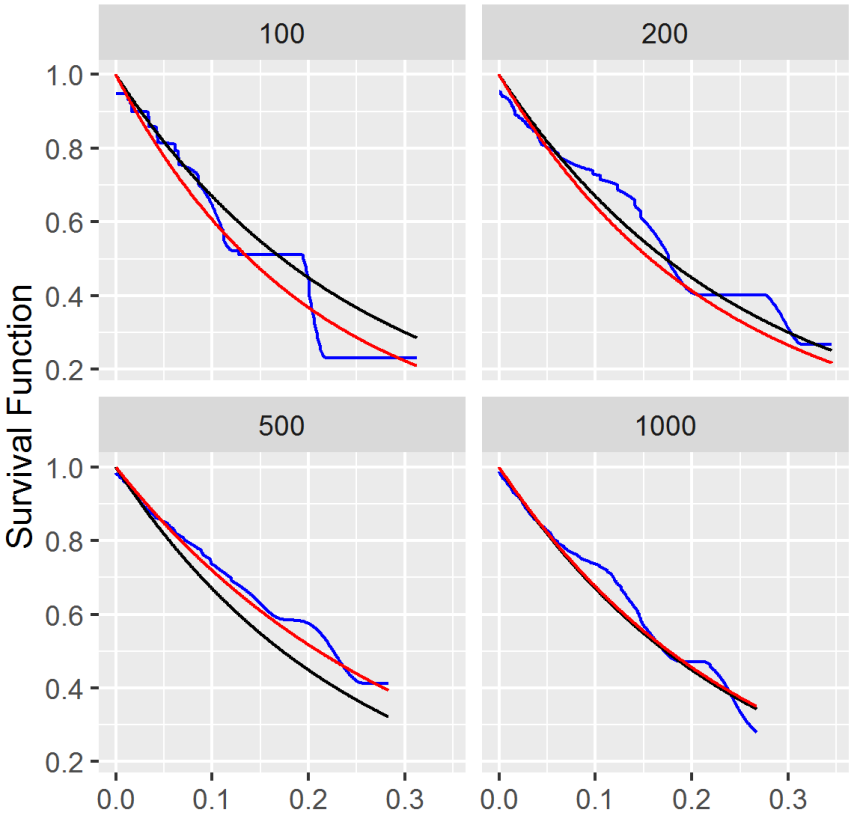

FIG 2. *Setting 1. The blue, red, and black curves represent the nonparametric, parametric, and true survival functions, respectively, for* $N = 100, 200, 500$ *and* 1000.

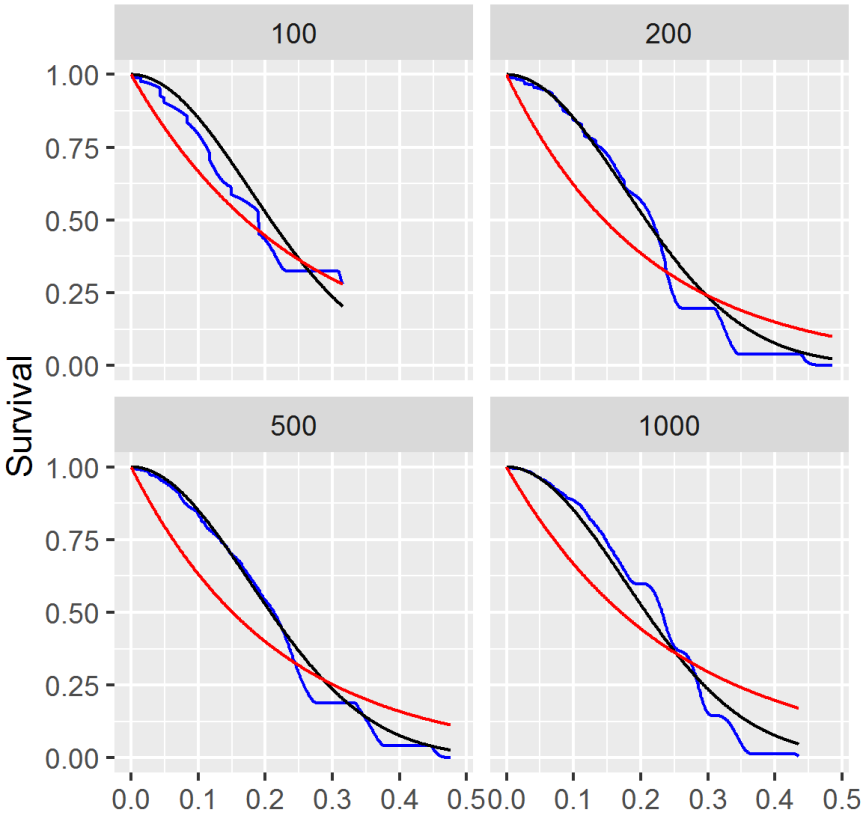

FIG 3. *Second setting. The blue, red, and black curves represent the nonparametric, parametric, and true survival functions, respectively.*

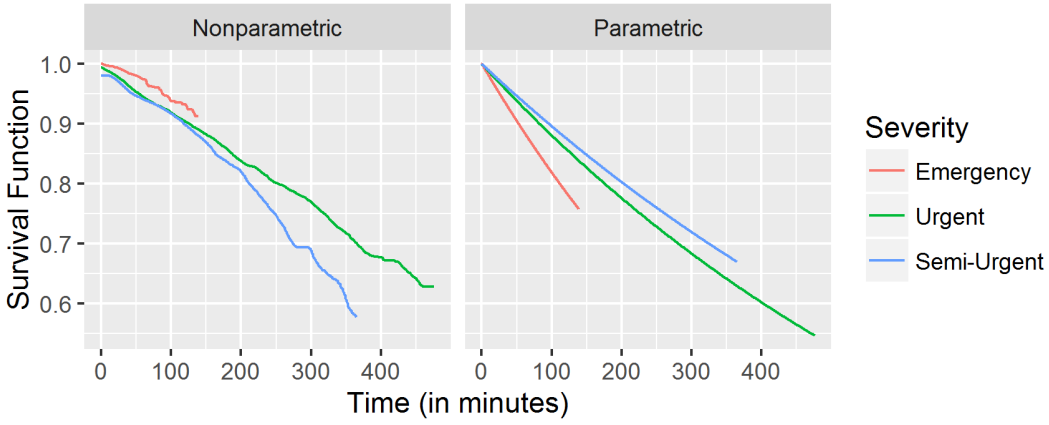

FIG 4. *Compression of the estimator for the survival function of the patience time at the three different levels of severity.*

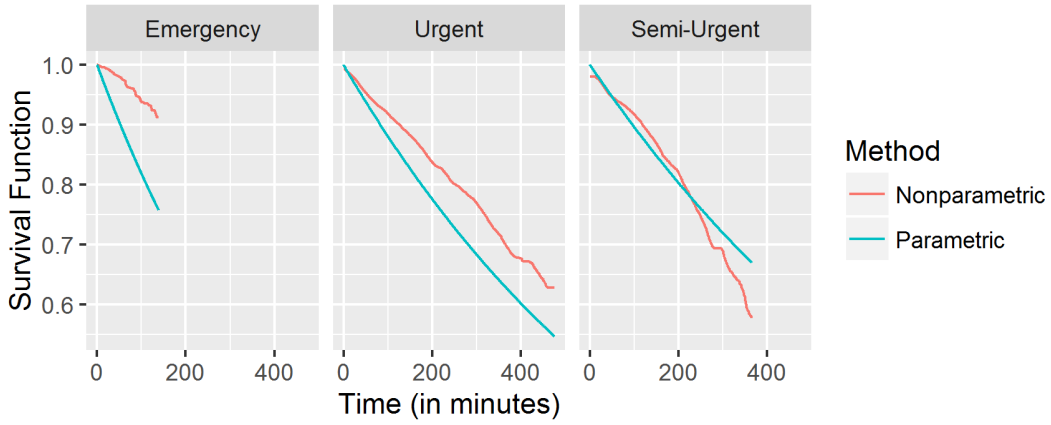

FIG 5. *Compression of the nonparametric and parametric estimators for the survival of the patience time by different levels of acuity.*

4. Discussion

The main question posed in this dissertation was how to overcome the difficulties and to analyze well complicated dataset structures. The estimation methods that were proposed include parametric and nonparametric approaches. The main question posed in this dissertation is how to analyze complicated dataset structures. The estimation methods included parametric and nonparametric estimators and are relevant to very wide settings.

Chapter 2 dealt with the challenge of estimating the test error. The test error is a functional of the data set and is defined as the probability of the classifier's misclassification. Due to its structure, it is difficult to construct a high quality point estimator for the test error. Instead of a point estimator, we proposed CIs for the test error. We proposed two construction methods—naive CIs and adaptive CIs. Each of these constructions was applied by both normal approximation and empirical bootstrap methods. Using simulation and analysis, we showed that the naive CIs are conservative and longer compared to the adaptive CIs. We also saw that there was no major difference between the CIs that were constructed by the normal approximation or the empirical bootstrap methods. The challenge of constructing these CIs has led to various theoretical perceptions about the convergence of empirical processes that are indexed by a class of RKHS functions. These theoretical perceptions can be used in future research. Future research directions include inference on more general functionals of the classifier, such as the probability of misclassification over a subset of the sample space.

In Chapter 3, we considered incomplete data that involves survival analysis, which has observed, right-censored, and left-censored data. Despite the fact that the data are not complete in this setting, we proposed parametric and nonparametric estimators. Using a simulation study, we showed that for a specified probabilistic structure, the parametric estimator holds and estimates well the patience time. When the model is misspecified, the nonparametric estimator behaves better. In the case study, we also observed that the nonparametric estimator performed better. Future research will address developing novel parametric and nonparametric estimators that can handle baseline covariates.

תקציר

בשנים האחרונות הרבה מסדי נתונים מצריכים ניתוח סטטיסטי מסובך. מסדי נתונים עם מבנה מורכב מתקבלים כאשר ההנחות על אפיון התפלגות הנתונים הן הנחות מאוד כלליות או כאשר חלק מן התצפיות אינן מכילות את המידע המדויק אותו אנו רוצים לנתח. במקרים כאלה המטרה היא לפתח תיאוריה סטטיסטית שתוכל להתגבר על הקשיים האלה ולהוביל לניתוח סטטיסטי אופטימלי.

דוגמה למודל שבו יש הנחות מאוד כלליות על אפיון התפלגות הנתונים הוא מודל של מיון נתונים סטטיסטי. במיון נתונים סטטיסטי המשתנה המסביר הוא רב ממדי ואילו המשתנה התלוי הוא בינארי. המטרה במידול כזה היא לייצר ממיין. הממיין הוא פונקציה שמאפשרת למיין תצפית מסבירה חדשה לאחד משני הערכים אותם מקבל המשתנה המוסבר. מידת הצלחת הממיין נמדדת בדרך כלל על ידי מבחן השגיאה שלו. מבחן השגיאה מוגדר כהסתברות למיון שגוי. לפיכך מבחן השגיאה הוא מדד חשוב המודד את טיב ההצלחה של הממיין. עם זאת, ניתוח סטטיסטי עבור מבחן השגיאה הוא מורכב שהרי מודל מיון נתונים סטטיסטים הוא כה כללי כך שאפילו קצב ההתכנסות של הממיין אינו ידוע. בפרק... אנו מציעים ניתוח סטטיסטי עבור מבחן השגיאה באמצעות בניית רווחי סמך. אנו בונים שני סוגים של רווחי סמך, נאיבי ואדפטיבי. יצירת רווח הסמך הנאיבי הייתה פשוטה יחסית ואילו יצירת רווח הסמך האדפטיבי מהווה אתגר קשה שמערב הבנה עמוקה בתיאורה של תהליכים אמפיריים. שני הסוגים של רווחי הסמך שאנו מציעים נבנים על סמך שתי שיטות ידועות בתהליכים אמפיריים-קירוב להתפלגות נורמלית ובוטסטרפ אמפירי.

סוג אחר של מודל עם מסד נתונים ובו מבנה מורכב הוא כאשר חלק מן התצפיות אינן מכילות את המידע המדויק אותו אנו רוצים לנתח. תופעה כזאת קורית במודל של ניתוח הישרדות. בפרק... אנו מתמודדים עם מודל הישרדות ספציפי. מודל זה מתייחס ללקוחות המחכים לטיפול בחדר מיון. התצפיות מחולקות לשלוש קטגוריות. חלק מן המטופלים מאבדים את סבלנותם תוך כדי המתנה בתור. מבין אלו שמאבדים את סבלנותם יש כאלה המכריזים למערכת כי הם נוטשים את התור. עבור מטופלים בקטגוריה זו יש זמן תיעוד מדויק של משך הזמן שעבר עד שפקעה סבלנותם. בקטגוריה השנייה יש מטופלים שהחליטו לנטוש את המערכת אך לא דיווחו על כך. כיוון שלא דיווחו, הם נקראים להיכנס לטיפול ומשך הזמן שעבר עד שנקראו להיכנס מתועד והוא מהווה חסם עליון למשך הזמן שעבר עד שפקעה סבלנותם. בקטגוריה השלישית יש את המטופלים שנכנסו לטיפול, משך הזמן שעבר עד שנקראו לטיפול מתועד והוא מהווה חסם תחתון למשך הזמן שעובר עד פקיעת הסבלנות שלהם. למרות שבמודל זה עבור חלק לא מבוטל מהמטופלים אין תיעוד מדויק של משך הזמן עד פקיעת הסבלנות, יש בפרק... תיאוריה לאמידה פרמטרית ואפרמטרית של התפלגות משך זמן עד פקיעת הסבלנות של המטופלים.

עבודה זו נעשתה תחת הדרכתם של

פרופסור יעקב ריטוב

ופרופסור יאיר גולדברג

אתגרים בניתוח סטטיסטי במסדי נתונים המצריכים ניתוח מורכב

חיבור לשם קבלת תואר דוקטור לפילוסופיה

מאת

יהונתן יפה נוף

הוגש לסנאט האוניברסיטה העברית

חשוון התש"פ