HHS Public Access

Author manuscript

Ann Stat. Author manuscript; available in PMC 2021 December 21.

Published in final edited form as:

Ann Stat. 2021 August ; 49(4): 2101–2128. doi:10.1214/20-aos2028.

BOOSTED NONPARAMETRIC HAZARDS WITH TIME-DEPENDENT COVARIATES

Donald K.K. Lee[1], **Ningyuan Chen**[2], **Hemant Ishwaran**[3]

[1]Goizueta Business School and Department of Biostatistics & Bioinformatics, Emory University,

[2]Rotman School of Management, University of Toronto,

[3]Division of Biostatistics, University of Miami,

Abstract

Given functional data from a survival process with time-dependent covariates, we derive a smooth convex representation for its nonparametric log-likelihood functional and obtain its functional gradient. From this we devise a generic gradient boosting procedure for estimating the hazard function nonparametrically. An illustrative implementation of the procedure using regression trees is described to show how to recover the unknown hazard. The generic estimator is consistent if the model is correctly specified; alternatively an oracle inequality can be demonstrated for tree-based models. To avoid overfitting, boosting employs several regularization devices. One of them is step-size restriction, but the rationale for this is somewhat mysterious from the viewpoint of consistency. Our work brings some clarity to this issue by revealing that step-size restriction is a mechanism for preventing the curvature of the risk from derailing convergence.

MSC2020 subject classifications:

Primary 62N02; secondary 62G05,90B22

Keywords and phrases;

survival analysis; gradient boosting; functional data; step-size shrinkage; regression trees; likelihood functional

1. Introduction.

Flexible hazard models involving time-dependent covariates are indispensable tools for studying systems that track covariates over time. In medicine, electronic health records systems make it possible to log patient vitals throughout the day, and these measurements can be used to build real-time warning systems for adverse outcomes such as cancer mortality [2]. In financial technology, lenders track obligors' behaviours over time to assess and revise default rate estimates. Such models are also used in many other fields of scientific inquiry since they form the building blocks for transitions within a Markovian state model.

donald.lee@emory.edu .

Indeed, this work was partly motivated by our study of patient transitions in emergency department queues and in organ transplant waitlist queues [20]. For example, allocation for a donor heart in the U.S. is defined in terms of coarse tiers [23], and transplant candidates are assigned to tiers based on their health status at the time of listing. However, a patient's condition may change rapidly while awaiting a heart, and this time-dependent information may be the most predictive of mortality and not the static covariates collected far in the past.

The main contribution of this paper is to introduce a fully nonparametric boosting procedure for hazard estimation with time-dependent covariates. We describe a generic gradient boosting procedure for boosting arbitrary weak base learners for this setting. Generally speaking, gradient boosting adopts the view of boosting as an iterative gradient descent algorithm for minimizing a loss functional over a target function space. Early work includes Breiman [6, 7, 8] and Mason et al. [21, 22]. A unified treatment was provided by Friedman [13], who coined the term "gradient boosting" which is now generally taken to be the modern interpretation of boosting.

Most of the existing boosting approaches for survival data focus on time-static covariates and involve boosting the Cox model. Examples include the popular R-packages mboost (Bühlmann and Hothorn [10]) and gbm (Ridgeway [26]) which apply gradient boosting to the Cox partial likelihood loss. Related work includes the penalized Cox partial likelihood approach of Binder and Schumacher [4]. Other important approaches, but not based on the Cox model, include L_2Boosting [11] with inverse probability of censoring weighting (IPCW) [16, 18], boosted transformation models of parametric families [15], and boosted accelerated failure time models [27].

While there are many boosting methods for dealing with time-static covariates, the literature is far more sparse for the case of time-dependent covariates. In fact, to our knowledge there is no general nonparametric approach for dealing with this setting. This is because in order to implement a fully nonparametric estimator, one has to contend with the issue of identifying the gradient, which turns out to be a non-trivial problem due to the functional nature of the data. This is unlike most standard applications of gradient boosting where the gradient can easily be identified and calculated.

1.1. Time-dependent covariate framework.

To explain why this is so challenging, we start by formally defining the survival problem with time-dependent covariates. Our description follows the framework of Aalen [1]. Let T denote the potentially unobserved failure time. Conditional on the history up to time t– the probability of failing at $T \in [t, t+dt)$ equals

$$\lambda(t, X(t))Y(t)dt\,. \tag{1}$$

Here $\lambda(t, x)$ denotes the unknown hazard function, $X(t) \in \mathfrak{X} \subseteq \mathbb{R}^p$ is a predictable covariate process, and $Y(t) \in \{0, 1\}$ is a predictable indicator of whether the subject is at risk at time t.[1] To simplify notation, without loss of generality we normalize the units of time so that $Y(t) = 0$ for $t > 1$.[2] In other words, the subject is not at risk after time $t = 1$, so we can restrict attention to the time interval (0, 1].

If failure is observed at $T \in (0, 1]$ then the indicator $\Delta = Y(T)$ equals 1, otherwise $\Delta = 0$ and we set $T = \infty$. Throughout we assume we observe n independent and identically distributed functional data samples $\{(X_i(\cdot), Y_i(\cdot), T_i)\}_{i=1}^n$. The evolution of observation i's failure status can then be thought of as a sequence of coin flips at time increments $t = 0, dt, 2dt, \cdots$, with the probability of "heads" at each time point given by (1). Therefore, observation i's contribution to the likelihood is

$$\{1 - \lambda(0, X_i(0))Y_i(0)dt\} \times \{1 - \lambda(dt, X_i(dt))Y_i(dt)dt\} \times \cdots$$
$$\times \lambda(T_i, X_i(T_i))^{\Delta_i} \xrightarrow[dt \downarrow 0]{} e^{-\int_0^1 Y_i(t)\lambda(t, X_i(t))dt} \lambda(T_i, X_i(T_i))^{\Delta_i},$$

where the limit can be understood as a product integral. Hence, if the log-hazard function is

$$F(t, x) = \log \lambda(t, x),$$

then the (scaled) negative log-likelihood functional is

$$R_n(F) = \frac{1}{n}\sum_{i=1}^n \int_0^1 Y_i(t)e^{F(t, X_i(t))}dt - \frac{1}{n}\sum_{i=1}^n \Delta_i F(T_i, X_i(T_i)), \tag{2}$$

which we shall refer to as the likelihood risk. The goal is to estimate the hazard function $\lambda(t, x) = e^{F(t, x)}$ nonparametrically by minimizing $R_n(F)$.

1.2. The likelihood does not have a gradient in generic function spaces.

As mentioned, our approach is to boost F using functional gradient descent. However the chief difficulty is that the canonical representation of the likelihood risk functional does not have a gradient. To see this, observe that the directional derivative of (2) equals

$$\frac{d}{d\theta}R_n(F + \theta f)\Big|_{\theta = 0} = \frac{1}{n}\sum_{i=1}^n \int_0^1 Y_i(t)e^{F(t, X_i(t))}f(t, X_i(t))dt$$
$$- \frac{1}{n}\sum_{i=1}^n \Delta_i f(T_i, X_i(T_i)), \tag{3}$$

which is the difference of two different inner products $\langle e^F, f\rangle_\dagger - \langle 1, f\rangle_\ddagger$ where

$$\langle g, f\rangle_\dagger = \frac{1}{n}\sum_{i=1}^n \int_0^1 Y_i(t)g(t, X_i(t))f(t, X_i(t))dt,$$

[1]The filtration of interest is $\sigma\{X(s), Y(s), I(T \le s) : s \le t\}$. If $X(t)$ is only observable when $Y(t) = 1$, we can set $X(t) = x^c \notin \mathcal{X}$ whenever $Y(t) = 0$.

[2]Since the data is always observed up to some finite time, there is no information loss from censoring at that point. For example, if T' is the failure time in minutes and the longest duration in the data is $\tau' = 60$ minutes, the failure time in hours, T, is at most $\tau = 1$ hour. The hazard function on the minute timescale, $\lambda_{T'}(t', X(t'))$, can be recovered from the hazard function on the hourly timescale, $\lambda_T(t, X(t))$, via $\lambda_{T'}(t', X(t')) = \frac{1}{\tau'}\lambda_T\left(\frac{t'}{\tau'}, X\left(\frac{t'}{\tau'}\right)\right)$.

$$\langle g, f\rangle_{\ddagger} = \frac{1}{n}\sum_{i=1}^{n} \Delta_i g(T_i, X_i(T_i)) f(T_i, X_i(T_i)).$$

Hence, (3) cannot be expressed as a single inner product of the form $\langle gF, f\rangle$ for some function $gF(t, x)$. Were it possible to do so, gF would then be the gradient function.

In simpler non-functional data settings like regression or classification, the loss can be written as $L(Y, \tilde{F}(x))$, where $\tilde{F}$ is the non-functional target statistical model and Y is the outcome, so the gradient is simply $\partial L(Y, \tilde{F}(x))/\partial \tilde{F}(x)$. The negative gradient is then approximated using a base learner $f \in \widetilde{\mathscr{F}}$ from a predefined class of functions $\widetilde{\mathscr{F}}$ (this being either parametric; for example linear learners, or nonparametric; for example tree learners). Typically, the optimal base learner $\hat{f}$ is chosen to minimize the L^2-approximation error and then scaled by a regularization parameter $0 < \nu \leq 1$ to obtain the updated estimate of $\tilde{F}$:

$$\tilde{F} \leftarrow \tilde{F} - \nu\hat{f}, \quad \hat{f} = \underset{f \in \widetilde{\mathscr{F}}}{\operatorname{argmin}} \left\| \frac{\partial L}{\partial \tilde{F}} - f \right\|_2.$$

Importantly, in the simpler non-functional setting the gradient does not depend on the space that $\tilde{F}$ belongs to. By contrast, a key insight of this paper is that the gradient of $R_n(F)$ can only be defined after carefully specifying an appropriate sample-dependent domain for $R_n(F)$. The likelihood risk can then be re-expressed as a smooth convex functional, and an analogous representation also exists for the population risk. These representations resolve the difficulty above, allow us to describe and implement a gradient boosting procedure, and are also crucial to establishing guarantees for our estimator.

1.3. Contributions of the paper.

A key discovery that unlocks the boosted hazard estimator is Proposition 1 of Section 2. It provides an integral representation for the likelihood risk from which several results follow, including, importantly, an explicit representation for the gradient. Proposition 1 relies on defining a suitable space of log-hazard functions defined on the time-covariate domain $[0, 1] \times \mathscr{X}$. Identifying this space is the key insight that allows us to rescue the likelihood approach and to derive the gradient needed to implement gradient boosting. Arriving at this framework is not conceptually trivial, and may explain the absence of boosted nonparametric hazard estimators until now.

Algorithm 1 of Section 2 describes our estimator. The algorithm minimizes the likelihood risk (2) over the defined space of log-hazard functions. In the special case of regression tree learners, expressions for the likelihood risk and its gradient are obtained from Proposition 1, which are then used to describe a tree-based implementation of our estimator in Section 4. In Section 5 we apply it to a high-dimensional dataset generated from a naturalistic simulation of patient service times in an emergency department.

Section 3 establishes the consistency of the procedure. We show that the hazard estimator is consistent if the space is correctly specified. In particular, if the space is the span of

regression trees, then the hazard estimator satisfies an oracle inequality and recovers λ up to some error tolerance (Propositions 3 and 4).

Another contribution of our work is to clarify the mechanisms used by gradient boosting to avoid overfitting. Gradient boosting typically applies two types of regularization to invoke slow learning: (i) A small step-size is used for the update; and (ii) The number of boosting iterations is capped. The number of iterations used in our algorithm is set using the framework of Zhang and Yu [31], whose work shows how stopping early ensures consistency. On the other hand, the role of step-size restriction is more mysterious. While [31] demonstrates that small step-sizes are needed to prove consistency, unrestricted greedy step-sizes are already small enough for classification problems [28] and also for commonly used regression losses (see the Appendix of [31]). We show in Section 3.3 that shrinkage acts as a counterweight to the curvature of the risk (see Lemma 2). Hence if the curvature is unbounded, as is the case for hazard regression, then the step-sizes may need to be explicitly controlled to ensure convergence. This important result adds to our understanding of statistical convergence in gradient boosting. As noted by Biau and Cadre [3] the literature for this is relatively sparse, which motivated them to propose another regularization mechanism that also prevents overfitting. Our work adds to this by considering the functional data setting.

Concluding remarks can be found in Section 6. Proofs not appearing in the body of the paper can be found in the Appendix.

2. The boosted hazard estimator.

In this section, we describe our boosted hazard estimator. To provide readers with concrete examples for the ideas introduced here, we will show how the quantities defined in this section specialize in the case of regression trees, which is one of a few possible ways to implement boosting.

We begin by defining in Section 2.1 an appropriate sample-dependent domain for the likelihood risk $R_n(F)$. As explained, this key insight allows us to re-express the likelihood risk and its population analogue as smooth convex functionals, thereby enabling us to compute their gradients in closed form in Propositions 1 and 2 of Section 2.2. Following this, the boosting algorithm is formally stated in Section 2.3.

2.1. Specifying a domain for $R_n(F)$.

We will make use of two identifiability conditions (A1) and (A2) to define the domain for $R_n(F)$. Condition (A1) below is the same as Condition 1(iv) of Huang and Stone [19].

Assumption (A1). The true hazard function $\lambda(t, x)$ is bounded between some interval $[\Lambda_L, \Lambda_U] \subset (0, \infty)$ on the time-covariate domain $[0, 1] \times \mathcal{X}$.

Recall that we defined $X(\cdot)$ and $Y(\cdot)$ to be predictable processes, and so it can be shown that the integrals and expectations appearing in this paper are all well defined. Denoting the indicator function as $I(\cdot)$, define the following population and empirical sub-probability measures on $[0, 1] \times \mathcal{X}$:

$$\mu(B) = \mathbb{E}\left(\int_0^1 Y(t) \cdot I[\{t, X(t)\} \in B]dt\right),$$

$$\mu_n(B) = \frac{1}{n}\sum_{i=1}^{n}\int_0^1 Y_i(t) \cdot I[\{t, X_i(t)\} \in B]dt,$$

and note that $\mathbb{E}\mu_n(B) = \mu(B)$ because the data is i.i.d. by assumption. Intuitively, μ_n measures the denseness of the observed sample time-covariate paths on $[0,1] \times \mathcal{X}$. For any integrable f,

$$\int f \, d\mu = \mathbb{E}\left(\int_0^1 Y(t) \cdot f(t, X(t))dt\right), \tag{4}$$

$$\int f \, d\mu_n = \frac{1}{n}\sum_{i=1}^{n}\int_0^1 Y_i(t) \cdot f(t, X_i(t))dt\,. \tag{5}$$

This allows us to define the following (random) norms and inner products

$$\|f\|_{\mu_n, 1} = \int |f| d\mu_n$$

$$\|f\|_{\mu_n, 2} = \left(\int f^2 d\mu_n\right)^{1/2}$$

$$\|f\|_\infty = \sup\{|f(t,x)| : (t,x) \in [0,1] \times \mathcal{X}\}$$

$$\langle f_1, f_2\rangle_{\mu_n} = \int f_1 f_2 d\mu_n,$$

and note that $\|\cdot\|_{\mu n, 1} \leq \|\cdot\|_{\mu n, 2} \leq \|\cdot\|_\infty$ because $\mu_n([0,1] \times \mathcal{X}) \leq 1$.

By careful design, μ_n allows us to specify a natural domain for $R_n(F)$. Let $\{\phi_j(t,x)\}_{j=1}^d$ be a set of bounded functions $[0,1] \times \mathcal{X} \mapsto [-1,1]$ that are linearly independent, in the sense that $\int_{[0,1]\times\mathcal{X}}\left(\sum_j c_j\phi_j\right)^2 dtdx = 0$ if and only if $c_1 = \cdots = c_d = 0$ (when some of the covariates are discrete-valued, dx should be interpreted as the product of a counting measure and the Lebesgue measure). The span of the functions is

$$\mathcal{F} = \left\{\sum_{j=1}^{d} c_j\phi_j : c_j \in \mathbb{R}\right\}.$$

For example, the span of all regression tree functions that can be defined on $[0, 1] \times \mathfrak{X}$, is $\mathcal{F} = \left\{ \sum_j c_j I_{B_j}(t, x) : c_j \in \mathbb{R} \right\}$,[3] which are linear combinations of indicator functions over disjoint time-covariate cubes indexed[4] by $j = (j_0, j_1, \cdots, j_p)$:

$$B_j = \left\{ (t, x) \in [0, 1] \times \mathfrak{X} : \begin{array}{c} t^{(j_0)} < t \le t^{(j_0+1)} \\ x^{(1, j_1)} < x^{(1)} \le x^{(1, j_1+1)} \\ \vdots \\ x^{(p, j_p)} < x^{(p)} \le x^{(p, j_p+1)} \end{array} \right\}. \tag{6}$$

Remark 1. The regions B_j are formed using all possible split points $\left\{ x^{(k, j_k)} \right\}_{j_k}$ for the k-th coordinate $x^{(k)}$, with the spacing determined by the precision of the measurements. For example, if weight is measured to the closest kilogram, then the set of all possible split points will be {0.5, 1.5, 2.5, ⋯} kilograms. Note that these split points are the finest possible for any realization of weight that is measured to the nearest kilogram. While abstract treatments of trees assume that there is a continuum of split points, in reality they fall on a discrete (but fine) grid that is pre-determined by the precision of the data.

When $\mathcal{F}$ is equipped with $\langle \cdot, \cdot \rangle_{\mu_n}$, we obtain the following sample-dependent subspace of $L^2(\mu_n)$, which is the appropriate domain for $R_n(F)$:

$$\left(\mathcal{F}, \langle \cdot, \cdot \rangle_{\mu_n} \right).$$

Note that the elements in $\left(\mathcal{F}, \langle \cdot, \cdot \rangle_{\mu_n} \right)$ are equivalence classes rather than actual functions that have well defined values at each (t, x). This is a problem because the likelihood risk (2) requires evaluating $F(t, x)$ at the points $(T_i, X_i(T_i))$ where $\Delta_i = 1$. We resolve this by fixing an orthonormal basis $\{\varphi_{nj}(t, x)\}_j$ for $\left(\mathcal{F}, \langle \cdot, \cdot \rangle_{\mu_n} \right)$, and represent each member of $\left(\mathcal{F}, \langle \cdot, \cdot \rangle_{\mu_n} \right)$ uniquely in the form $\sum_j c_j \varphi_{nj}(t, x)$. For example in the case of regression trees, applying the Gram-Schmidt procedure to $\left\{ \phi_j(t, x) = I_{B_j}(t, x) \right\}_j$ gives

$$\left\{ \varphi_{nj}(t, x) \right\}_j = \left\{ \frac{I_{B_j}(t, x)}{\mu_n(B_j)^{1/2}} : \mu_n(B_j) > 0 \right\},$$

which by design have disjoint support.

[3] It is clear that said span is contained in $\mathcal{F}$. For the converse, it suffices to show that $\mathcal{F}$ is also contained in the span of trees of some depth. This is easy to show for trees with $p + 1$ splits, because they can generate partitions of the form $(-\infty, t] \times (-\infty, x^{(1)}] \times \cdots \times (-\infty, x^{(p)}]$ in $[0, 1] \times \mathfrak{X}$ (Section 3 of [8]).

[4] With a slight abuse of notation, the index j is only considered multi-dimensional when describing the geometry of B_j, such as in (6). In all other situations j should be interpreted as a scalar index.

The second condition we impose is for $\{\phi_j\}_{j=1}^d$ to be linearly independent in $L^2(\mu)$, that is $\|\sum_j c_j\phi_j\|_{\mu,2}^2 = \sum_{ij} c_i\left(\int \phi_i\phi_j d\mu\right)c_j = 0$ if and only if $c_1 = \cdots = c_d = 0$. Since by construction $\{\phi_j\}_{j=1}^d$ are already linearly independent in $[0,1]\times\mathfrak{X}$, the condition intuitively requires the set of all possible time-covariate trajectories to be adequately dense in $[0,1]\times\mathfrak{X}$ to intersect a sufficient amount of the support of every ϕ_j. This is weaker than the identifiability conditions 1(ii)-1(iii) in [19] which require $X(t)$ to have a positive joint probability density on $[0,1]\times\mathfrak{X}$.

Assumption (A2). The Gram matrix $\Sigma_{ij} = \int \phi_i\phi_j d\mu$ is positive definite.

2.2. Integral representations for the likelihood risk.

Having deduced the appropriate domain for $R_n(F)$, we can now recast the risk as a smooth convex functional on $\left(\mathcal{F}, \langle\cdot,\cdot\rangle_{\mu_n}\right)$. Proposition 1 below provides closed form expressions for this and its gradient. We note that if the risk is actually of a certain simpler form, it might be possible to estimate its gradient empirically from our risk expression using [24].

Proposition 1. *For functions $F(t,x)$, $f(t,x)$ of the form $\sum_j c_j\varphi_{nj}(t,x)$, the likelihood risk* (2) *can be written as*

$$R_n(F) = \int \left(e^F - \lambda_n F\right)d\mu_n, \tag{7}$$

where $\lambda_n \in \left(\mathcal{F}, \langle\cdot,\cdot\rangle_{\mu_n}\right)$ is the function

$$\lambda_n(t,x) = \frac{1}{n}\sum_j \left\{\sum_{i=1}^n \Delta_i\varphi_{nj}(T_i, X_i(T_i))\right\}\varphi_{nj}(t,x).$$

Thus there exists $\rho \in (0,1)$ (depending on F and f) for which the Taylor representation

$$R_n(F+f) = R_n(F) + \langle g_F, f\rangle_{\mu_n} + \frac{1}{2}\int e^{F+\rho f} f^2 d\mu_n \tag{8}$$

holds, where the gradient

$$g_F(t,x) = \sum_j \left\langle e^F, \varphi_{nj}\right\rangle_{\mu_n}\varphi_{nj}(t,x) - \lambda_n(t,x) \tag{9}$$

of $R_n(F)$ is the projection of $e^F - \lambda_n$ onto $\left(\mathcal{F}, \langle\cdot,\cdot\rangle_{\mu_n}\right)$. Hence if $gF = 0$ then the infimum of $R_n(F)$ over the span of $\{\varphi_{nj}(t,x)\}_j$ is uniquely attained at F.

For regression trees the expressions (7) and (9) simplify further because $\mathcal{F}$ is closed under pointwise exponentiation, i.e. $e^F \in \mathcal{F}$ for $F \in \mathcal{F}$. This is because the B_j's are disjoint so $F = \sum_j c_j I_{B_j}$ and hence $e^F = \sum_j e^{c_j} I_{B_j}$. Thus

$$\lambda_n(t,x) = \sum_{j:\,\mu_n(B_j)>0} \frac{\text{Fail}_j}{n\mu_n(B_j)} I_{B_j}(t,x), \tag{10}$$

$$R_n(F) = \sum_{j:\,\mu_n(B_j)>0} \left(e^{c_j}\mu_n(B_j) - \frac{c_j \text{Fail}_j}{n} \right), \tag{11}$$

$$g_F(t,x) = \sum_{j:\,\mu_n(B_j)>0} \left(e^{c_j} - \frac{\text{Fail}_j}{n\mu_n(B_j)} \right) I_{B_j}(t,x), \tag{12}$$

where

$$\text{Fail}_j = \sum_i \Delta_i I\left[\{T_i, X_i(T_i)\} \in B_j\right]$$

is the number of observed failures in the time-covariate region B_j.

Proof of Proposition 1. Fix a realization of $\{(X_i(\cdot), Y_i(\cdot), T_i)\}_{i=1}^n$. Using (5) we can rewrite (2) as

$$R_n(F) = \int e^F d\mu_n - \frac{1}{n}\sum_{i=1}^n \Delta_i F(T_i, X_i(T_i)).$$

We can express F in terms of the basis $\{\varphi_{nk}\}_k$ as $F(t,x) = \sum_k c_k \varphi_{nk}(t,x)$. Hence

$$\begin{aligned}
\int \lambda_n F d\mu_n &= \int \frac{1}{n}\sum_j \left\{ \sum_{i=1}^n \Delta_i \varphi_{nj}(T_i, X_i(T_i)) \right\} \varphi_{nj}(t,x) F(t,x) d\mu_n \\
&= \frac{1}{n}\sum_j \left\{ \sum_{i=1}^n \Delta_i \varphi_{nj}(T_i, X_i(T_i)) \right\} \int \varphi_{nj}(t,x) F(t,x) d\mu_n \\
&= \frac{1}{n}\sum_j \left\{ \sum_{i=1}^n \Delta_i \varphi_{nj}(T_i, X_i(T_i)) \right\} \int \varphi_{nj}(t,x) \sum_k c_k \varphi_{nk}(t,x) d\mu_n \\
&= \frac{1}{n}\sum_j \left\{ \sum_{i=1}^n \Delta_i \varphi_{nj}(T_i, X_i(T_i)) \right\} c_j \\
&= \frac{1}{n}\sum_{i=1}^n \Delta_i \sum_j c_j \varphi_{nj}(T_i, X_i(T_i)) \\
&= \frac{1}{n}\sum_{i=1}^n \Delta_i F(T_i, X_i(T_i)),
\end{aligned}$$

where the fourth equality follows from the orthonormality of the basis. This completes the derivation of (7).

By an interchange argument we obtain

$$\frac{d}{d\theta}R_n(F+\theta f) = \int \left(e^{F+\theta f} - \lambda_n\right) f d\mu_n,$$

$$\frac{d^2}{d\theta^2}R_n(F+\theta f) = \int e^{F+\theta f} f^2 d\mu_n,$$

the latter being positive whenever $f \neq 0$; i.e., $R_n(F)$ is convex. The Taylor representation (8) then follows from noting that gF is the orthogonal projection of $e^F - \lambda_n \in L^2(\mu_n)$ onto $\left(\mathcal{F}, \langle \cdot, \cdot \rangle_{\mu_n}\right)$. □

The expectation of the likelihood risk also has an integral representation. A special case of the representation (13) below is proved in Proposition 3.2 of [19] for right-censored data only, under assumptions that do not allow for internal covariates. In the statement of the proposition below recall that Λ_L and Λ_U are defined in (A1). The constant $\alpha_{\mathcal{F}}$ is defined later in (28).

Proposition 2. *For* $F \in \mathcal{F} \cup \{\log \lambda\}$,

$$R(F) = \mathbb{E}\{R_n(F)\} = \int \left(e^F - \lambda F\right) d\mu\,. \tag{13}$$

Furthermore the restriction of $R(F)$ *to* $\mathcal{F}$ *is coercive:*

$$\frac{1}{2}R(F) \geq \frac{\Lambda_L}{\alpha_{\mathcal{F}}}\left\|F\right\|_\infty + \Lambda_U \min\{0, 1 - \log(2\Lambda_U)\}, \tag{14}$$

and it attains its minimum at a unique point $F^* \in \left(\mathcal{F}, \langle \cdot, \cdot \rangle_\mu\right)$. *If* $\mathcal{F}$ *contains the underlying log-hazard function then* $F^* = \log \lambda$.

Remark 2. Coerciveness (14) implies that any F with expected risk $R(F)$ less than $R(0) \leq 1 < 3$ is uniformly bounded:

$$\left\|F\right\|_\infty < \frac{\alpha_{\mathcal{F}}}{\Lambda_L}[3/2 + \Lambda_U \max\{0, \log(2\Lambda_U) - 1\}] \leq \alpha_{\mathcal{F}}\beta_\Lambda \tag{15}$$

where the constant

$$\beta_\Lambda = \frac{3/2 + \Lambda_U \max\{0, \log(2\Lambda_U) - 1\}}{\min\{1, \Lambda_L\}} \tag{16}$$

is by design no smaller than 1 in order to simplify subsequent analyses.

2.3. The boosting procedure.

In gradient boosting the key idea is to update an iterate in a direction that is approximately aligned to the negative gradient. To model this direction formally, we introduce the concept of an ε-gradient.

Definition 1. *Suppose* $gF \neq 0$. *We say that a unit vector* $g_F^\varepsilon \in \left(\mathcal{F}, \langle \cdot, \cdot \rangle_{\mu_n}\right)$ *is an* ε*-gradient at* F *if for some* $0 < \varepsilon \leq 1$,

$$\left\langle \frac{g_F}{\|g_F\|_{\mu_n, 2}}, g_F^\varepsilon \right\rangle_{\mu_n} \geq \varepsilon . \tag{17}$$

Call $-g_F^\varepsilon$ *a negative* ε*-gradient if* g_F^ε *is an* ε*-gradient.*

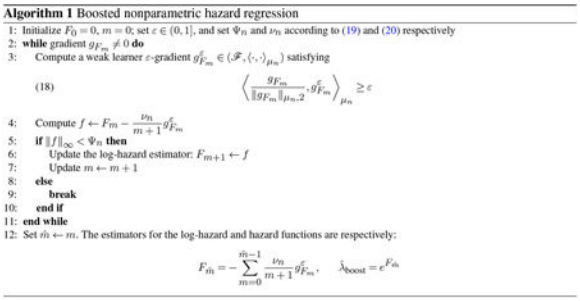
Our boosting procedure seeks approximations g_F^ε that satisfy (17) for some pre-specified alignment value ε. The larger ε is, the closer the alignment is between the negative gradient and the negative ε-gradient, and the greater the risk reduction. In particular, $-gF$ is the unique negative 1-gradient with maximal risk reduction. In practice, however, we find that using a smaller value of ε leads to simpler approximations that prevent overfitting in finite samples. This is consistent with other implementations of boosting: It is well known that the statistical performance of gradient descent generally improves when simpler base learners are used.

Algorithm 1 describes the proposed boosting procedure for estimating λ. For a given level of alignment ε, Line 3 finds an ε-gradient $g_{F_m}^\varepsilon$ at F_m satisfying (18) at the m-th iteration, and uses its negation for the boosting update in Line 4. If the ε-gradients are regression tree learners, as is the case with the implementation in Section 4, then the trees cannot be grown in the same way as the standard boosting algorithm in Friedman [13]. This is because the standard approach grows all regression trees to a fixed depth, which may or may not ensure ε-alignment at each boosting iteration.

To ensure ε-alignment, the depth of the trees are not fixed in the implementation in Section 4. Instead, at each boosting iteration a tree is grown to whatever depth is needed to satisfy (18). This can always be done because the alignment ε is non-decreasing in the number of tree splits, and with enough splits we can recover the gradient gF_m itself up to μ_n-almost everywhere.[5] As mentioned earlier, we recommend using small values of ε, which can be determined in practice using cross-validation. This differs from the standard approach where cross-validation is used to select a common tree depth to use for all boosting iterations.

In addition to the gradient alignment ε, Algorithm 1 makes use of two other regularization parameters, Ψ_n and ν_n. The first defines the early stopping criterion (how many boosting iterations to use), while the second controls the step-sizes of the boosting updates. These are two common regularization techniques used in boosting:

[5]Split the tree until each leaf node contains just one of the regions B_j in (6) with $\mu_n(B_j) > 0$. Then set the value of the node equal to the value of the gradient function (12) inside B_j.



Algorithm 1 Boosted nonparametric hazard regression

Initialize $F_0 = 0$, $m = 0$; set $\varepsilon \in (0, 1]$, and set Ψ_n and ν_n according to (19) and (20) respectively
while gradient $g_{F_m} \neq 0$ **do**
  Compute a weak learner ε-gradient $g^{\varepsilon}_{F_m} \in (\mathcal{F}, \langle \cdot, \cdot \rangle_{\mu_n})$ satisfying

(18) $$\left\langle \frac{g_{F_m}}{\|g_{F_m}\|_{\mu_n,2}}, g^{\varepsilon}_{F_m} \right\rangle_{\mu_n} \geq \varepsilon$$

  Compute $f \leftarrow F_m - \frac{\nu_n}{m+1} g^{\varepsilon}_{F_m}$
  if $\|f\|_\infty < \Psi_n$ **then**
    Update the log-hazard estimator: $F_{m+1} \leftarrow f$
    Update $m \leftarrow m + 1$
  else
    break
  end if
end while
Set $\hat{m} \leftarrow m$. The estimators for the log-hazard and hazard functions are respectively:

$$F_{\hat{m}} = -\sum_{m=0}^{\hat{m}-1} \frac{\nu_n}{m+1} g^{\varepsilon}_{F_m}, \qquad \hat{\lambda}_{\text{boost}} = e^{F_{\hat{m}}}$$

1. *Early stopping.* The number of boosting iterations $\hat{m}$ is controlled by stopping the algorithm before the uniform norm of the estimator $\|F_{\hat{m}}\|_\infty$ reaches or exceeds

$$\Psi_n = W\left(n^{1/4}\right) \to \infty, \tag{19}$$

where $W(y)$ is the branch of the Lambert function that returns the real root of the equation $ze^z = y$ for $y > 0$.

2. *Step-sizes.* The step-size $\nu_n \ll 1$ used in gradient boosting is typically held constant across iterations. While we can also do this in our procedure,[6] the role of step-size shrinkage becomes more salient if we use $\nu_n/(m+1)$ instead as the step-size for the m-th iteration in Algorithm 1. This step-size is controlled in two ways. First, it is made to decrease with each iteration according to the Robbins-Monro condition that the sum of the steps diverges while the sum of squared steps converges. Second, the shrinkage factor ν_n is selected to make the step-sizes decay with n at rate

$$\nu_n^2 e^{\Psi_n} < 1, \qquad \nu_n^2 e^{\Psi_n} \to 0. \tag{20}$$

This acts as a counterbalance to $R_n(F)$'s unbounded curvature:

$$\frac{d^2}{d\theta^2} R_n(F + \theta f)\Big|_{\theta = 0} = \int e^F f^2 d\mu_n, \tag{21}$$

which is upper bounded by e^{Ψ_n} when $\|F\|_\infty < \Psi_n$ and $\|f\|_{\mu_n,2} = 1$.

3. Consistency.

Under (A1) and (A2), guarantees for our hazard estimator $\hat{\lambda}_{\text{boost}}$ in Algorithm 1 can be derived for two scenarios of interest. The guarantees rely on the regularizations described in Section 2.3 to avoid overfitting. In the following development, recall from Proposition 2 that F^* is the unique minimizer of $R(F)$, so it satisfies the first order condition

$$\left\langle e^{F^*} - \lambda, F \right\rangle_\mu = 0 \tag{22}$$

[6] The term $\nu_n^2 e^{\Psi_n}$ in condition (20) would need to be replaced by $\hat{m}\nu_n^2 e^{\Psi_n}$ if a constant step-size is used.

for all $F \in \mathscr{F}$. Recall that the span of all trees is closed under pointwise exponentiation $\left(e^F \in \mathscr{F}\right)$, in which case (22) implies that $\lambda^* = e^{F^*}$ is the orthogonal projection of λ onto $\left(\mathscr{F}, \langle \cdot, \cdot \rangle_\mu\right)$.

1. *Consistency when $\mathscr{F}$ is correctly specified.* If the true log-hazard function $\log \lambda$ is in $\mathscr{F}$, then Proposition 2 asserts that $F^* = \log \lambda$. It will be shown in this case that $\hat{\lambda}_{\text{boost}}$ is consistent:

$$\left\|\hat{\lambda}_{\text{boost}} - \lambda\right\|_{\mu,2}^2 = o_p(1).$$

2. *Oracle inequality for regression trees.* If $\mathscr{F}$ is closed under pointwise exponentiation, it follows from (22) that λ^* is the best $L^2(\mu)$-approximation to λ among all candidate hazard estimators $\left\{e^F : F \in \mathscr{F}\right\}$. It can then be shown that $\hat{\lambda}_{\text{boost}}$ converges to this best approximation:

$$\left\|\hat{\lambda}_{\text{boost}} - \lambda\right\|_{\mu,2}^2 = \left\|\lambda^* - \lambda\right\|_{\mu,2}^2 + o_p(1).$$

   This oracle result is in the spirit of the type of guarantees available for tree-based boosting in the non-functional data setting. For example, if tree stumps are used for L_2-regression, then the regression function estimate will converge to the best approximation to the true regression function in the span of tree stumps [9]. Similar results also exist for boosted classifiers [5].

Propositions 3 and 4 below formalize these guarantees by providing bounds on the error terms above. While sharper bounds may exist, the chief purpose of this paper is to introduce our generic estimator for the first time and to provide guarantees that apply across different implementations. More refined convergence rates may exist for a specific implementation, just like the analysis in Bühlmann and Yu [11] for L_2Boosting when componentwise spline learners are specifically used. We leave this for future research.

En route to establishing the guarantees, Lemma 2 below clarifies the role played by step-size restriction in ensuring convergence of the estimator. As explained in the Introduction, explicit shrinkage is not necessary for classification and regression problems where the risk has bounded curvature. Lemma 2 suggests that it may, however, be needed when the risk has unbounded curvature, as is the case with $R_n(F)$. Seen in this light, shrinkage is really a mechanism for controlling the growth of the risk curvature.

3.1. Strategy for establishing guarantees.

The representations for $R_n(F)$ and its population analogue $R(F)$ from Section 2 are the key ingredients for formalizing the guarantees. We use them to first show that $F_{\hat{m}} \in \left(\mathscr{F}, \langle \cdot, \cdot \rangle_{\mu_n}\right)$ converges to $F^* \in \left(\mathscr{F}, \langle \cdot, \cdot \rangle_\mu\right)$: Applying Taylor's theorem to the representation for $R(F)$ in Proposition 2 yields

$$\|F_{\hat{m}} - F^*\|_{\mu,2}^2 \le 2\frac{R(F_{\hat{m}}) - R(F^*)}{\min_{t,x}\left(\lambda^* \wedge \hat{\lambda}_{\text{boost}}\right)}. \tag{23}$$

The problem is thus transformed into one of risk minimization $R(F_{\hat{m}}) \to R(F^*)$, for which [31] suggests analyzing separately the terms of the decomposition

$$\begin{aligned} 0 \le\ & R(F_{\hat{m}}) - R(F^*) \\ \le\ & |R_n(F_{\hat{m}}) - R(F_{\hat{m}})| \quad \text{(I)} \quad \text{COMPLEXITY ARGUMENT} \\ +\ & |R_n(F^*) - R(F^*)| \quad \text{(II)} \quad \text{STANDARD ARGUMENT} \\ +\ & \{R_n(F_{\hat{m}}) - R_n(F^*)\}. \quad \text{(III)} \quad \text{CURVATURE ARGUMENT} \end{aligned} \tag{24}$$

The authors argue that in boosting, the point of limiting the number of iterations $\hat{m}$ (enforced by lines 5–10 in Algorithm 1) is to prevent $F_{\hat{m}}$ from growing too fast, so that (I) converges to zero as $n \to \infty$. At the same time, $\hat{m}$ is allowed to grow with n in a controlled manner so that the empirical risk $R_n(F_{\hat{m}})$ in (III) is eventually minimized as $n \to \infty$. Lemmas 1 and 2 below show that our procedure achieves both goals. Lemma 1 makes use of complexity theory via empirical processes, while Lemma 2 deals with the curvature of the likelihood risk. The term (II) will be bounded using standard concentration results.

3.2. Bounding (I) using complexity.

To capture the effect of using a simple negative ε-gradient (17) as the descent direction, we bound (I) in terms of the complexity of[7]

$$\begin{aligned} &\mathcal{F}_\varepsilon = \mathcal{F}_{\varepsilon,\text{boost}} \cup \left\{F \in \mathcal{F} : \|F\|_\infty = 1\right\} \subseteq \mathcal{F}, \\ &\text{where } \mathcal{F}_{\varepsilon,\text{boost}} = \left\{F_m = -\sum_{k=0}^{m-1} \frac{\nu_n}{k+1} g^\varepsilon_{F_k} : m = 0, 1, \ldots\right\}. \end{aligned} \tag{25}$$

Depending on the choice of weak learners for the ε-gradients, $\mathcal{F}_\varepsilon$ may be much smaller than $\mathcal{F}$. For example, coordinate descent might only ever select a small subset of basis functions $\{\phi_j\}_j$ because of sparsity. As another example if $\lambda(t, x)$ is additively separable in time and also in each covariate, then regression trees might only ever select simple tree stumps (one tree split).

The measure of complexity we use below comes from empirical process theory. Define $\mathcal{F}_\varepsilon^\Psi = \{F \in \mathcal{F}_\varepsilon : \|F\|_\infty < \Psi\}$ for $\Psi > 0$ and suppose that Q is a sub-probability measure on $[0, 1] \times \mathfrak{X}$. Then the $L^2(Q)$-ball of radius $\delta > 0$ centred at some $F \in L_2(Q)$ is $\left\{F' \in \mathcal{F}_\varepsilon^\Psi : \|F' - F\|_{Q,2} < \delta\right\}$. The covering number $\mathcal{N}\left(\delta, \mathcal{F}_\varepsilon^\Psi, Q\right)$ is the minimum number of such balls needed to cover $\mathcal{F}_\varepsilon^\Psi$ (Definitions 2.1.5 and 2.2.3 of van der Vaart and Wellner [29]), so $\mathcal{N}\left(\delta, \mathcal{F}_\varepsilon^\Psi, Q\right) = 1$ for $\delta \ge \Psi$. A complexity measure for $\mathcal{F}_\varepsilon$ is

[7] For technical convenience, $\mathcal{F}_\varepsilon$ has been enlarged from $\mathcal{F}_{\varepsilon,\text{boost}}$ to include the unit ball.

$$J_{\mathcal{F}_\varepsilon} = \sup_{\Psi, Q}\left\{\int_0^1 \left\{\log \mathcal{N}\left(u\Psi, \mathcal{F}_\varepsilon^{\Psi}, Q\right)\right\}^{1/2} du\right\}, \tag{26}$$

where the supremum is taken over $\Psi > 0$ and over all non-zero sub-probability measures. As discussed, $J_{\mathcal{F}}$ is never greater than, and potentially much smaller than $J_{\mathcal{F}}$, the complexity of $\mathcal{F}$, which is fixed and finite.

Before stating Lemma 1, we note that the result also shows an empirical analogue to the norm equivalences

$$\left\|F\right\|_{\mu,1} \le \left\|F\right\|_{\mu,2} \le \left\|F\right\|_{\infty} \le \frac{\alpha_{\mathcal{F}}}{2}\left\|F\right\|_{\mu,1} \text{ for all } F \in \mathcal{F} \tag{27}$$

exists, where

$$\alpha_{\mathcal{F}} = 2 \sup_{F \in \mathcal{F}: \|F\|_\infty = 1}\left(\frac{\|F\|_\infty}{\|F\|_{\mu,1}}\right) = \frac{2}{\inf_{F \in \mathcal{F}: \|F\|_\infty = 1}\left\|F\right\|_{\mu,1}} > 1. \tag{28}$$

The factor of 2 serves to simplify the presentation, and can be replaced with anything greater than 1.

Lemma 1. *There exists a universal constant* κ *such that for any* $0 < \eta < 1$*, with probability at least*

$$1 - 4\exp\left\{-\left(\frac{\eta n^{1/4}}{\kappa \alpha_{\mathcal{F}} J_{\mathcal{F}_\varepsilon}}\right)^2\right\}$$

an empirical analogue to (27) *holds for all* $F \in \mathcal{F}$:

$$\|F\|_{\mu_n,1} \le \|F\|_{\mu_n,2} \le \|F\|_\infty \le \alpha_{\mathcal{F}}\|F\|_{\mu_n,1}, \tag{29}$$

and for all $F \in \mathcal{F}_\varepsilon^{\Psi_n}$,

$$\left|\{R_n(F) - R_n(0)\} - \{R(F) - R(0)\}\right| < \eta. \tag{30}$$

Remark 3. The equivalences (29) imply that $\dim\left(\mathcal{F}, \langle\cdot,\cdot\rangle_{\mu_n}\right)$ equals its upper bound $\dim\mathcal{F} = d$. That is, if $\|\sum_j c_j\phi_j\|_{\mu_n,2} = 0$, then $\|\sum_j c_j\phi_j\|_\infty = 0$, so $c_1 = \cdots = c_d = 0$ because $\{\phi_j\}_{j=1}^d$ are linearly independent on $[0,1] \times \mathfrak{X}$.

3.3. Bounding (III) using curvature.

We use the representation in Proposition 1 to study the minimization of the empirical risk $R_n(F)$ by boosting. Standard results for exact gradient descent like Theorem 2.1.15 of Nesterov [25] are in terms of the norm of the minimizer, which may not exist for $R_n(F)$.[8]

If coordinate descent is used instead, Section 4.1 of [31] can be applied to convex functions whose infimum may not be attainable, but its curvature is required to be uniformly bounded above. Since the second derivative of $R_n(F)$ is unbounded (21), Lemma 2 below provides two remedies: (i) Use the shrinkage decay (20) of ν_n to counterbalance the curvature; (ii) Use coercivity (15) to show that with increasing probability, $\{F_m\}_{m=0}^{\widehat{m}}$ are uniformly bounded, so the curvatures at those points are also uniformly bounded. Lemma 2 combines both to derive a result that is simpler than what can be achieved from either one alone. In doing so, the role played by step-size restriction becomes clear. The lemma relies in part on adapting the analysis in Lemma 4.1 of [31] for coordinate descent to the case for generic ε-gradients. The conditions required below will be shown to hold with high probability.

Lemma 2. *Suppose* (29) *holds and that*

$$|R_n(F^*) - R(F^*)| < 1, \qquad \sup_{F \in \mathcal{F}_S^* n} |R_n(F) - R(F)| < 1.$$

Then the largest gap between F^* *and* $\{F_m\}_{m=0}^{\widehat{m}}$,

$$\widehat{\gamma} = \max_{m \le \widehat{m}} \|F_m - F^*\|_\infty \vee 1, \tag{31}$$

is bounded by a constant no greater than $2\alpha_{\mathcal{F}}\beta_\Lambda$, *and for* $n \ge 55$,

$$R_n(F_{\widehat{m}}) - R_n(F^*) < 2e\beta_\Lambda \left(\frac{\log n}{4n^{1/4}}\right)^{\varepsilon/(\alpha_{\mathcal{F}}\widehat{\gamma})} + \nu_n^2 e^{\Psi_n}. \tag{32}$$

Remark 4. The last term in (32) suggests that the role of the step-size shrinkage ν_n is to keep the curvature of the risk in check, to prevent it from derailing convergence. Recall from (21) that e^{Ψ_n} describes the curvature of $R_n(F_m)$. Thus our result clarifies the role of step-size restriction in boosting functional data.

Remark 5. Regardless of whether the risk curvature is bounded or not, smaller step-sizes always improve the convergence bound. This can be seen from the parsimonious relationship between ν_n and (32). Fixing n, pushing the value of ν_n down towards zero yields the lower limit

$$2e\beta_\Lambda \left(\frac{\log n}{4n^{1/4}}\right)^{\varepsilon/(\alpha_{\mathcal{F}}\widehat{\gamma})}.$$

However, this limit is unattainable as ν_n must be positive in order to decrease the risk. This effect has been observed in practical applications of boosting. Friedman [13] noted

[8]The infimum of $R_n(F)$ is not always attainable: If f is non-positive and vanishes on the set $\{\{T_i, X_i(T_i)\} : \Delta_i = 1\}$, then $R_n(F + \theta f) = \int \left(e^{F + \theta f} - \lambda_n F\right) d\mu_n$ is decreasing in θ so f is a direction of recession. This is however not an issue for boosting because of early stopping.

improved performance for gradient boosting with the use of a small shrinkage factor ν. At the same time, it was also noted there was diminishing performance gain as ν became very small, and this came at the expense of an increased number of boosting iterations. This same phenomenon has also been observed for L_2Boosting [11] with componentwise linear learners. It is known that the solution path for L_2Boosting closely matches that of lasso as $\nu \to 0$. However, the algorithm exhibits cycling behaviour for small ν, which greatly increases the number of iterations and offsets the performance gain in trying to approximate the lasso (see Ehrlinger and Ishwaran [12]).

3.4. Formal statements of guarantees.

As a reminder, we have defined the following quantities:

$\hat{\lambda}_{\text{boost}} = e^{F\hat{m}}$, the boosted hazard estimator in Algorithm 1

$\lambda^* = e^{F^*}$, where F^* is the unique minimizer of $R(F)$ in Proposition 2

Λ_L, Λ_U = lower and upper bounds on $\lambda(t, x)$ as defined in (A1)

$\hat{\gamma}$ = maximum gap between F^* and $\{F_m\}_{m=0}^{\hat{m}}$ defined in (31)

κ = a universal constant

$\alpha_{\mathcal{F}}$ = constant defined in (28) β_Λ = constant defined in (16)

$J_{\mathcal{F}_\varepsilon}$ = complexity measure (26), bounded above by $J_{\mathcal{F}}$

To simplify the results, we will assume that $n \geq 55$ and also set the shrinkage to satisfy $\nu_n^2 e^{\Psi_n} = \log n/\left(64 n^{1/4}\right)$. Our first guarantee shows that our hazard estimator is consistent if the model is correctly specified.

Proposition 3. *(Consistency under correct model specification). Suppose $\mathcal{F}$ contains the true log-hazard function* $\log \lambda$. *Then with probability*

$$1 - 8 \exp\left\{-\left(\frac{\log n}{\kappa \alpha_{\mathcal{F}}\left(\Lambda_L^{-1} \vee \Lambda_U\right) J_{\mathcal{F}_\varepsilon}}\right)^2\right\}$$

we have that $\|\hat{\lambda}_{\text{boost}}\|_\infty$ *is bounded and*

$$\|\hat{\lambda}_{\text{boost}} - \lambda\|_{\mu,2}^2 < 13\beta_\Lambda \frac{\max_{t,x}\left(\lambda \vee \hat{\lambda}_{boost}\right)^2}{\min_{t,x}\left(\lambda \wedge \hat{\lambda}_{boost}\right)}\left(\frac{\log n}{4n^{1/4}}\right)^{\varepsilon/(\alpha_{\mathcal{F}}\hat{\gamma})}.$$

Thus $\hat{\lambda}_{\text{boost}}$ *is consistent.*

Via the tension between ε and $J_{\mathcal{F}_\varepsilon}$, Proposition 3 captures the trade-off in statistical performance between weak and strong learners in gradient boosting. The advantage of low complexity (weak learners) is reflected in the increased probability of the $L^2(\mu)$-bound

holding, with this probability being maximized when $J_{\mathcal{F}_\varepsilon} \to 0$, which generally occurs as $\varepsilon \to 0$. However, diametrically opposed to this, we find that the $L^2(\mu)$-bound is minimized by $\varepsilon \to 1$, which occurs with the use of stronger learners that are more aligned with the gradient. This same trade-off is also captured by our second guarantee which establishes an oracle inequality for tree learners.

Proposition 4. *(Oracle inequality for tree learners). Suppose* $e^F \in \mathcal{F}$ *for* $F \in \mathcal{F}$*. Then among* $\{e^F : F \in \mathcal{F}\}$, λ^* *is the best* $L^2(\mu)$*-approximation to* λ*, that is*

$$\lambda^* = \arg \min_{e^F : F \in \mathcal{F}} \|e^F - \lambda\|_{\mu,2}.$$

Moreover, $\hat{\lambda}_{\text{boost}}$ *converges to this best approximation* λ^**: With probability*

$$1 - 8 \exp\left\{-\left(\frac{\log n}{\kappa \alpha_{\mathcal{F}}\left(\Lambda_L^{-1} \vee \Lambda_U\right) J_{\mathcal{F}_\varepsilon}}\right)^2\right\}$$

we have that $\|\hat{\lambda}_{boost}\|_\infty$ *is bounded and*

$$\|\hat{\lambda}_{boost} - \lambda\|_{\mu,2}^2 < \rho_{\mathcal{F}}^2 + 13\beta_\Lambda \frac{\max_{t,x}\left(\Lambda_U \vee \hat{\lambda}_{boost}\right)^2}{\min_{t,x}\left(\Lambda_L \wedge \hat{\lambda}_{\text{boost}}\right)} \left(\frac{\log n}{4n^{1/4}}\right)^{\varepsilon/(\alpha_{\mathcal{F}}\hat{\gamma})},$$

where $\rho_{\mathcal{F}}^2 = \|\lambda^* - \lambda\|_{\mu,2}^2$ *is the smallest error one can achieve from using functions in* $\{e^F : F \in \mathcal{F}\}$ *to approximate* λ.

For tree learners, $\lambda^*(t, x)$ is constant over each region B_j in (6), and its value equals the local average of λ over B_j,

$$\lambda^*(t,x)|_{B_j} = \frac{1}{\mu(B_j)} \int_{B_j} \lambda d\mu.$$

Hence if the B_j's are small, λ^* should closely approximate λ (recall from Remark 1 that the size of the B_j's is fixed by the data). To estimate the approximation error $\rho_{\mathcal{F}}$ in terms of B_j, suppose that λ is sufficiently smooth, e.g. Hölder continuous $|\lambda(t,x) - \lambda(t',x')| \lesssim \|(t - t', x - x')\|^b$ for some $b > 0$. Then since $\inf_{B_j} \lambda \le \lambda^*|_{B_j} \le \sup_{B_j} \lambda$,

$$\rho_{\mathcal{F}} \le \|\lambda^* - \lambda\|_\infty \lesssim \max_j \left(\text{diam}\, B_j\right)^b.$$

4. A tree-based implementation.

Here we describe an implementation of Algorithm 1 using regression trees, whereby the ε-gradient $g^{\varepsilon}_{F_m}$ is obtained by growing a tree to satisfy (18) for a pre-specified ε.

To explain the tree growing process, first observe that the m-th step log-hazard estimator is an additive expansion of CART basis functions. Thus it can be written as

$$\begin{aligned} F_m(t,x) &= \sum_{b=0}^{m-1} \sum_{l=1}^{L_b} \gamma_{b,l} I_{A_{b,l}}(t,x) \\ &= \sum_j c_{m,j} I_{B_j}(t,x), \end{aligned} \tag{33}$$

where $A_{b,l}$ is the l-th leaf region of the b-th tree. Recall from Section 2.3 that each tree is grown until (18) is satisfied, so the number of leaf nodes L_b can vary from tree to tree. The leaf regions are typically large subsets of the time-covariate space $[0,1] \times \mathcal{X}$ adaptively determined by the tree growing process (to be discussed shortly). Since each leaf region can be further decomposed into the finer disjoint regions B_j in (6), $F_m(t, x)$ can be rewritten as (33). However, many of these regions will share the same coefficient value, so (33) can be written more compactly as

$$F_m(t,x) = \sum_j c_{m,j} I_{B'_{m,j}}(t,x),$$

where $B'_{m,j}$ is the union of contiguous regions whose coefficient equals $c_{m,j}$. This smooths the hazard estimator $\hat{\lambda}_{\text{boost}}(t,x)$ over $[0,1] \times \mathcal{X}$, thanks to the regularization imposed by limiting the number of trees (early stopping) and also by the use of weak tree learners. This is unlike the unconstrained hazard MLE $\lambda_n(t, x)$ defined in (10), which can take on a different value in each region B_j, making it prone to overfit the data.

To construct an ε-gradient $g^{\varepsilon}_{F_m}$ with ε-alignment to gF_m defined by (12),

$$g_{F_m}(t,x) = \sum_{j:\, \mu_n(B_j) > 0} \left(e^{c_{m,j}} - \frac{\text{Fail}_j}{n\mu_n(B_j)} \right) I_{B_j}(t,x),$$

the tree splits are adaptively chosen to reduce the $L^2(\mu_n)$-approximation error between $g^{\varepsilon}_{F_m}$ and gF_m. We implement tree splits for both time and covariates. Specifically, suppose we wish to split a leaf region $A \subseteq [0,1] \times \mathcal{X}$ into left and right daughter subregions A_1 and A_2, and assign values γ_1 and γ_2 to them. For example, a split on the k-th covariate could propose left and right daughters such as

$$A_1 = \left\{(t,x) \in A : x^{(k)} \le s\right\}, \quad A_2 = \left\{(t,x) \in A : x^{(k)} > s\right\}, \tag{34}$$

or a split on time t could propose regions

$$A_1 = \{(t,x) \in A : t \le s\}, \quad A_2 = \{(t,x) \in A : t > s\}. \tag{35}$$

Now note that gF_m is constant within each region B_j. We denote its value by $g_{F_m}\left(t_{B_j}, x_{B_j}\right)$ where $\left(t_{B_j}, x_{B_j}\right)$ is the centre of B_j. Hence the best split of A into A_1 and A_2 is the one that minimizes

$$\begin{aligned}
&\min_{\gamma_1} \int_{A_1} \left\{g_{F_m}(t,x) - \gamma_1\right\}^2 d\mu_n + \min_{\gamma_2} \int_{A_2} \left\{g_{F_m}(t,x) - \gamma_2\right\}^2 d\mu_n \\
&= \min_{\gamma_1} \sum_{j: B_j \subseteq A_1} \mu_n(B_j) \cdot \left\{g_{F_m}\left(t_{B_j}, x_{B_j}\right) - \gamma_1\right\}^2 + \min_{\gamma_2} \sum_{k: B_k \subseteq A_2} \mu_n(B_k) \cdot \left\{g_{F_m}\left(t_{B_k}, x_{B_k}\right) - \gamma_2\right\}^2 \\
&= \min_{\gamma_1} \sum_{\substack{j: z_j \in A_1, \\ w_j > 0}} w_j \cdot \left(\tilde{y}_j - \gamma_1\right)^2 + \min_{\gamma_2} \sum_{\substack{k: z_k \in A_2, \\ w_k > 0}} w_k \\
&\cdot \left(\tilde{y}_k - \gamma_2\right)^2,
\end{aligned} \tag{36}$$

where

$$\tilde{y}_j = g_{F_m}\left(t_{B_j}, x_{B_j}\right) = e^{c_{m,j}} - \frac{\text{Fail}_j}{n\mu_n\left(B_j\right)}$$

represents the j-th pseudo-response, $z_j = \left(t_{B_j}, x_{B_j}\right)$ its covariate and $w_j = \mu_n(B_j)$ its weight. Thus the splits use a weighted least squares criterion, which can be efficiently computed as usual.

We split the tree until (18) is satisfied, resulting in L_m leaf nodes ($L_m - 1$ splits). As discussed in Section 2.3, we can always find a deep enough tree that is an ε-gradient because with enough splits we can recover the gradient gF_m itself. Recall also that a small value of ε performs best in practice, and this can be chosen by cross-validating on a set of small-sized candidates: For each one we implement Algorithm 1, and we select the one that minimizes the cross-validated risk $R_n(F)$ defined in (11). By contrast, the standard boosting algorithm [13] uses cross-validation to select a common number of splits to use for all trees, which does not ensure that each tree is an ε-gradient.

Regarding the possible split points for the covariates (34), note that the k-th covariate $x^{(k)} = x^{(k)}(t)$ is a time series that is sampled periodically. This yields a set of unique values equal to the union of all of the sampled values for the n observations. In direct analogy to non-functional data boosting, we place candidate split points in-between the sorted values in this set. In other words, splits for covariates only occur at values corresponding to the observed data just as in non-functional boosting.

The resolution for the grid of candidate time splits (35) is set equal to the temporal resolution. For example, the covariate trajectories in the simulation in Section 5 are piecewise constant and may change every 0.002 days. Placing the candidate split points at 0.002, 0.004, ⋯ days simplifies the exact computation of $\mu_n(B_j)$ because every covariate

trajectory is constant between these points. Again, notice that the splits for time only occur at values informed by the observed data.

Putting it together, the setup above leverages our insight in (36) by transforming the survival functional data into the data values $\{w_j, \tilde{y}_j, z_j\}_{j:w_j>0}$, which enables the implementation to proceed like standard gradient boosting for non-functional data. Only the pseudo-response $\tilde{y}_j$ in $\{w_j, \tilde{y}_j, z_j\}$ needs to be updated at each boosting iteration, while the other two do not change. In terms of storage it costs $\mathcal{O}(np|\mathcal{T}|)$ to store $\{w_j, \tilde{y}_j, z_j\}_{j:w_j>0}$, where $|\mathcal{T}|$ is the cardinality of the set of candidate time splits.[9] Computationally, choosing a new tree split requires testing $\mathcal{O}(np|\mathcal{T}|)$ candidate splits.[10] The space and time complexities of the implementation are reasonable given that they are $\mathcal{O}(np)$ for non-functional data boosting: In the functional data setting, each sample can have up to $|\mathcal{T}|$ observations, so n functional data samples is akin to $\mathcal{O}(n|\mathcal{T}|)$ samples in a non-functional data setting.

5. Numerical experiment.

We now apply the boosting procedure of Section 4 to a high-dimensional dataset generated from a naturalistic simulation. This allows us to compare the performance of our estimator to existing boosting methods. The simulation is of patient service times in an emergency department (ED), and the hazard function of interest is patient service rate in the ED. The study of patient transitions in an ED queue is an important one in healthcare operations, because without a high resolution model of patient flow dynamics, the ED may be suboptimally utilized which would deny patients of timely critical care.

5.1. Service rate.

The service rate model used in the simulation is based upon a service time dataset from the ED of an academic hospital in the United States. The dataset contains information on 86,983 treatment encounters from 2014 to early 2015. Recorded for each encounter was: Age, gender, Emergency Severity Index (ESI)[11], time of day when treatment in the ED ward began, day of week of ED visit, and ward census. The last one represents the total number of occupied beds in the ED ward, which varies over the course of the patient's stay. Hence it is a time-dependent variable. Lastly, we also have the duration of the patient's stay (service time).

The service rate function is developed from the data in the following way. First, we apply our nonparametric estimator to the data to perform exploratory analysis. We find that:

[9] Each $\{w_j, \tilde{y}_j, z_j\}$ is of dimension $p+3$ and the number of time-covariate regions B_j with $w_j>0$ is at most $n(|\mathcal{T}|+1)$. To show the latter, observe that B_j will only have $w_j = \mu_n(B_j) > 0$ if it is traversed by at least one sample covariate trajectory. Then note that each of the n sample covariate trajectories can traverse at most $|\mathcal{T}|+1$ unique regions.

[10] A sample covariate trajectory can have at most $|\mathcal{T}|$ unique observed values for the k-th covariate $x^{(k)}$, so there are at most $n|\mathcal{T}|$ candidate splits for $x^{(k)}$. Thus there are $\mathcal{O}(np|\mathcal{T}|)$ candidate splits for p covariates. The number of candidate splits on time is obviously $|\mathcal{T}|$.

[11] Level 1 is the most severe (e.g., cardiac arrest) and level 5 is the least (e.g., rash). We removed level 1 patients from the dataset because they were treated in a separate trauma bay.

1. The key variables affecting the service rate (based on relative variable importance [13]) are ESI, age, and ward census. In addition, two of the most pronounced interaction terms identified by the tree splits are (AGE ≥ 34, ESI = 5) and (AGE ≥ 34, ESI ≤ 4).
2. Holding all the variables fixed, the shapes of the estimated service rate function resemble the hazard functions of log-normal distributions. This agrees with the queuing literature that find log-normality to be a reasonable parametric fit for service durations.

Guided by these findings, we specify the service rate $\lambda(t, X(t))$ for the simulation as a log-normal accelerated failure time (AFT) model, and estimate its parameters from data. This yields the service rate

$$\lambda(t, x) = \theta(x) \cdot \frac{\phi_l(\theta(x)t; m, \sigma)}{1 - \Phi_l(\theta(x)t; m, \sigma)}, \tag{37}$$

where $\phi_l(\cdot; m, \sigma)$ and $\Phi_l(\cdot; m, \sigma)$ are the PDF and CDF of the log-normal distribution with log-mean $m = -1.8$ and log-standard deviation $\sigma = 0.74$. The function $\theta(x)$ captures the dependence of the service rate on the covariates:

$$\begin{aligned}\log \theta(X(t)) = & -0.0071 \cdot \text{AGE} + 0.022 \cdot \text{ESI} - \min\left\{a \cdot \frac{\text{CENSUS}_t}{70}, 2\right\} + 0.10 \cdot I \\ & (\text{AGE} \geq 34, \text{ESI} = 5) - 0.10 \cdot I(\text{AGE} \geq 34, \text{ESI} \leq 4) + 0 \cdot \text{NUISANCE}_1 + \cdots \\ & + 0 \cdot \text{NUISANCE}_{43}.\end{aligned} \tag{38}$$

The specification for $\theta(X(t))$ above is a slight modification of the original estimate, with the free parameter a allowing us to study the effect of time-dependent covariates on hazard estimation. When $a = 0$, the service rate does not depend on time-varying covariates, but as a increases, the dependency becomes more and more significant. In the data, the ward census never exceeds 70, so we set the capacity of the simulated ED to 70 as well. The min operator caps the impact that census can have on the simulated service rate as a grows. The irrelevant covariates $\text{NUISANCE}_1, \cdots, \text{NUISANCE}_{43}$ are added to the data in order to assess how boosting performs in high dimensions. We explicitly include them in (38) to remind ourselves that the simulated data is high-dimensional. Forty of the irrelevant variables are generated synthetically as described in the next subsection, while the rest are variables from the original dataset not used in the simulation.

5.2. Simulation model.

Using (37) and (38), we simulate a naturalistic dataset of 10,000 patient visit histories. The value of a will be varied from 0 to 3 in order to study the impact of time-dependent covariates on hazard estimation. Each patient is associated with a 46-dimensional covariate vector consisting of:

- The time-varying ward census. The initial value is sampled from its marginal empirical distribution in the original dataset. To simulate its trajectory over a patient's stay, for every timestep advance of 0.002 days (≈3 minutes), a Bernoulli(0.02) random variable is generated. If it is one, then the census is

incremented by a normal random variable with zero mean and standard deviation 10. The result is truncated if it lies outside the range [1, 70], the upper end being the capacity of the ED.

- The other five time-static covariates in the original dataset. These are sampled from their marginal empirical distributions in the original dataset. Two of the variables (age and ESI) influence the service rate, while the other three are irrelevant.
- An additional forty time-static covariates that do not affect the service rate (irrelevant covariates). Their values are drawn uniformly from [0, 1].

We also generate independent censoring times (rounded to the nearest 0.002 days) for each visit from an exponential distribution. For each simulation, the rate of the exponential distribution is set to achieve an approximate target of 25% censoring.

5.3. Comparison benchmarks.

When the covariates are static in time, a few software packages are available for performing hazard estimation with tree ensembles. Given that the data is simulated from a log-normal hazard, we compare our nonparametric method to two correctly specified parametric estimators:

1. The blackboost estimator in the R package mboost [10] provides a tree boosting procedure for fitting the log-normal hazard function. In order to apply this to the simulated data, we make ward census a time-static covariate by fixing it at its initial value.
2. Transformation forests [17] in the R package trtf can also fit log-normal hazards. Moreover, it allows for left-truncated and right-censored data. Since the ward census variable is simulated to be piecewise constant over time, we can treat each segment as a left-truncated and right-censored observation. Thus for this simulation, transformation forests are able to handle time-dependent covariates with time-static effects. This falls in between the static covariate/static effect blackboost estimator and our fully nonparametric one.

Since the service rate model used in the simulations is in fact log-normal, the benchmark methods above enjoy a significant advantage over our nonparametric one, which is not privy to the true distribution. In fact, when $a = 0$ the log-normal hazard (37) depends only on time-static covariates, so the benchmarks should outperform our nonparametric estimator. However, as a grows, we would expect a reversal in relative performance.

To compare the performances of the estimators, we use Monte Carlo integration to evaluate the relative mean squared error

$$\%\text{MSE} = \frac{\mathbb{E}_X\left[\int_0^1 \{\lambda(t,X) - \hat{\lambda}(t,X)\}^2 dt\right]}{\mathbb{E}_X\left[\int_0^1 \lambda(t,X)^2 dt\right]}.$$

The Monte Carlo integrations are conducted using an independent test set of 10,000 uncensored patient visit histories. For the test set, ward census is held fixed over time at the initial value, and we use the grid {0, 0.02, 0.04, ⋯, 1} for the time integral. The nominator above is then estimated by the average of $\left\{\lambda(t,x)-\hat{\lambda}(t,x)\right\}^2$ evaluated at the 51×10,000 points of (t, x). The denominator is estimated in the same manner.

5.4. Results.

For the implementation of our estimator in Section 4, the value of ε and the number of trees $\hat{m}$ are jointly determined using ten-fold cross validation. The candidate values we tried for ε are {0.003, 0.004, 0.005, 0.006, 0.007}, and we limit $\hat{m}$ to no more than 1,000 trees. A wider range of values can be of course be explored for better performance (at the cost of more computations). As comparison, we run an ad-hoc version of our algorithm in which all trees use the same number of splits, as is the case in standard boosting [13]. This approach does not explicitly ensure that the trees will be ε-gradients for a pre-specified ε. The number of splits and the number of trees used in the ad-hoc method are jointly determined using ten-fold cross-validation.

In order to speed up convergence at the m-th iteration for both approaches, instead of using the step-size $\nu_n/(m+1)$ of Algorithm 1, we performed line-search within the interval $(0, \nu_n/(m+1)]$. While Lemma 2 shows that a smaller shrinkage ν_n is always better, this comes at the expense of a larger $\hat{m}$ and hence computation time. For simplicity we set $\nu_n = 1$ for all the experiments here.

For fitting the blackboost estimator, we use the default setting of nu = 0.1 for the step-size taken at each iteration. The other hyperparameters, mstop (the number of trees) and maxdepth (maximum depth of trees), are chosen to directly minimize the relative MSE on the test set. This of course gives the blackboost estimator an unfair advantage over our estimator, which is on top of the fact that it is based on the same distribution as the true model. Transformation forest (using also the true distribution) is fit using code kindly provided by Professor T. Hothorn.[12]

Variable selection.—The relative importances of variables [13] for our estimator are given in Table 1 for all four cases $a = 0, 1, 2, 3$. The four factors that influence the service rate (38) are explicitly listed, while the irrelevant covariates are grouped together in the last column. When $a = 0$, the service rate does not depend on census, and we see that the importance of census and the other irrelevant covariates are at least an order of magnitude smaller than the relevant ones. As a increases, census becomes more and more important as correctly reflected in the table. Across all the cases the importance of the relevant covariates are at least an order of magnitude larger than the others, suggesting that our estimator is able to pick out the influential covariates and largely avoid the irrelevant ones.

[12] In the code 100 trees are used in the forest, which takes about 700 megabytes to store the fitted object when applied to our simulated data.

Presence of time-dependent covariates.—Table 2 presents the relative MSEs for the estimators as the service rate function (38) becomes increasingly dependent on the time-varying census variable. When $a = 0$ the service rate depends only on time-static covariates, so as expected, the parametric log-normal benchmarks perform the best when applied to data simulated from a log-normal AFT model.

However, as a increases, the service rate becomes increasingly dependent on census. The corresponding performances of both benchmarks deteriorate dramatically, and is handily outperformed by the proposed estimator. We note that the inclusion of just one time-dependent covariate is enough to degrade the performances of the benchmarks, despite the fact that they have the exact same parametric form as the true model.

Finally we find comparable performance among the ad-hoc boosted estimator and our proposed one, although a slight edge goes to the latter especially in the more difficult simulations with larger a. The results here demonstrate that there is a place in the survival boosting literature for fully nonparametric methods like this one that can flexibly handle time-dependent covariates.

6. Discussion.

Our estimator can also potentially be used to evaluate the goodness-of-fit of simpler parametric hazard models. Since our approach is likelihood-based, future work might examine whether model selection frameworks like those in [30] can be extended to cover likelihood functionals. For this, [10] provides some guidance for determining the effective degrees of freedom for the boosting estimator. The ideas in [32] may also be germane.

The implementation presented in Section 4 is one of many possible ways to implement our estimator. We defer the design of a more refined implementation to future research, along with open-source code.

Acknowledgements.

The review team provided many insightful comments that significantly improved the paper. We are grateful to Brian Clarke, Jack Hall, Sahand Negahban, and Hongyu Zhao for helpful discussions. Special thanks to Trevor Hastie for early formative discussions. The dataset used in Section 5 was kindly provided by Dr. Kito Lord. NC was supported by HKUST grant R9382 and NSERC Discovery Grant RGPIN-2020-04038. HI was supported by National Institutes of Health grants R01 GM125072 and R01 HL141892. DKKL was supported by a hyperplane.

APPENDIX: PROOFS

Proof of Proposition 2.

Proof. Writing

$$R(F) = \mathbb{E}\left(\int_0^1 Y(t) \cdot e^{F(t, X(t))} dt - \Delta F(T, X(T))\right),$$

we can apply (4) to establish the first part of the integral in (13) when $F \in \mathscr{F} \cup \{\log \lambda\}$. To complete the representation, it suffices to show that the point process

$$M(B) = \Delta \cdot I[\{T, X(T)\} \in B]$$

has mean $\int_B \lambda d\mu$, and then apply Campbell's formula. To this end, write $N(t) = I(T \le t)$ and consider the filtration $\sigma\{X(s),\ Y(s), N(s) : s \le t\}$. Then $N(t)$ has the Doob-Meyer form $dN(t) = \lambda(t, X(t))\, Y(t) dt + dM(t)$ where $M(t)$ is a martingale. Hence

$$\begin{aligned}
&\mathbb{E}\left\{M(B)\right\} = \mathbb{E}\left(\int_0^1 I[\left\{t, X(t)\right\} \in B] dN(t)\right) \\
&= \mathbb{E}\left(\int_0^1 Y(t) \cdot I[\left\{t, X(t)\right\} \in B] \cdot \lambda(t, X(t)) dt\right) + \mathbb{E}\left(\int_0^1 I[\left\{t, X(t)\right\} \in B] dM(t)\right) \\
&= \int_B \lambda d\mu + \mathbb{E}\left(\int_0^1 I[\left\{t, X(t)\right\} \in B] dM(t)\right),
\end{aligned}$$

where the last equality follows from (4). Since $I[\{t, X(t)\} \in B]$ is predictable because $X(t)$ is, the desired result follows if the stochastic integral $\int_0^1 I[\left\{t, X(t)\right\} \in B] dM(t)$ is a martingale. By Section 2 of Aalen [1], this is true if $M(t)$ is square-integrable. In fact, $M(t) = N(t) - \int_0^t \lambda(t, X(t)) dt$ is bounded because $\lambda(t, x)$ is bounded above by (A1). This establishes (13).

Now note that for a positive constant Λ the function $e^y - \Lambda y$ is bounded below by both $-\Lambda y$ and $\Lambda y + 2\Lambda\{1 - \log 2\Lambda\}$, hence $e^y - \Lambda y \ge \Lambda |y| + 2\Lambda \min\{0, 1 - \log 2\Lambda\}$. Since $\Lambda \min\{0, 1 - \log 2\Lambda\}$ is non-increasing in Λ, (A1) implies that

$$\begin{aligned}
&e^{F(t,x)} - \lambda(t,x) F(t,x) \ge \min\left\{e^{F(t,x)} - \Lambda_L F(t,x), e^{F(t,x)} - \Lambda_U F(t,x)\right\} \\
&\ge \Lambda_L \left|F(t,x)\right| + 2\Lambda_U \min\{0, 1 - \log(2\Lambda_U)\}.
\end{aligned}$$

Integrating both sides and using the norm equivalence relation (27) shows that

$$\begin{aligned}
R(F) &\ge \Lambda_L \left\|F\right\|_{\mu,1} + 2\Lambda_U \min\{0, 1 - \log(2\Lambda_U)\} \\
&\ge \frac{2\Lambda_L}{\alpha_{\mathcal{F}}} \left\|F\right\|_{\infty} + 2\Lambda_U \min\{0, 1 - \log(2\Lambda_U)\} \\
&\ge \frac{2\Lambda_L}{\alpha_{\mathcal{F}}} \left\|F\right\|_{\mu,2} + 2\Lambda_U \min\{0, 1 - \log(2\Lambda_U)\}.
\end{aligned}$$

The lower bound (14) then follows from the second inequality. The last inequality shows that $R(F)$ is coercive on $(\mathcal{F}, \langle \cdot, \cdot \rangle_\mu)$. Moreover the same argument used to derive (8) shows that $R(F)$ is smooth and convex on $(\mathcal{F}, \langle \cdot, \cdot \rangle_\mu)$. Therefore a unique minimizer F^* of $R(F)$ exists in $(\mathcal{F}, \langle \cdot, \cdot \rangle_\mu)$. Since (A2) implies there is a bijection between the equivalent classes of $(\mathcal{F}, \langle \cdot, \cdot \rangle_\mu)$ and the functions in $\mathcal{F}$, F^* is also the unique minimizer of $R(F)$ in $\mathcal{F}$. Finally, since $e^{F(t,x)} - \lambda(t,x) F(t,x)$ is pointwise bounded below by $\lambda(t,x)\{1 - \log \lambda(t,x)\}$, $R(F) \ge \int (\lambda - \lambda \log \lambda) d\mu = R(\log \lambda)$ for all $F \in \mathcal{F}$. □

Proof of Lemma 1.

Proof. By a pointwise-measurable argument (Example 2.3.4 of [29]) it can be shown that all suprema quantities appearing below are sufficiently well behaved, so outer integration is not required. Define the Orlicz norm $\|X\|_{\Phi} = \inf\{C > 0 : \mathbb{E}\Phi(|X|/C) \leq 1\}$ where $\Phi(x) = e^{x^2} - 1$. Suppose the following holds:

$$\left\| \sup_{F \in \mathscr{F}_\varepsilon^{\Psi_n}} \left| \{R_n(F) - R_n(0)\} - \{R(F) - R(0)\} \right| \right\|_{\Phi} \leq \kappa' J_{\mathscr{F}_\varepsilon} / n^{1/4}, \tag{39}$$

$$\left\| \sup_{G \in \mathscr{F}_\varepsilon : \|G\|_\infty \leq 1} \left| \|G\|_{\mu_n, 1} - \|G\|_{\mu, 1} \right| \right\|_{\Phi} \leq \kappa'' J_{\mathscr{F}_\varepsilon} / n^{1/2}, \tag{40}$$

where $J_{\mathscr{F}_\varepsilon}$ is the complexity measure (26), and κ', κ'' are universal constants. Then by Markov's inequality, (30) holds with probability at least $1 - 2\exp\left[-\left\{\eta n^{1/4}/\left(\kappa' J_{\mathscr{F}_\varepsilon}\right)\right\}^2\right]$, and

$$\sup_{G \in \mathscr{F}_\varepsilon : \|G\|_\infty \leq 1} \left\{ \|G\|_{\mu, 1} - \|G\|_{\mu_n, 1} \right\} < 1/\alpha_{\mathscr{F}} \tag{41}$$

holds with probability at least $1 - 2\exp\left[-\left\{n^{1/2}/\left(\alpha_{\mathscr{F}} \kappa'' J_{\mathscr{F}_\varepsilon}\right)\right\}^2\right]$. Since $\alpha_{\mathscr{F}} > 1$ and $\eta < 1$, (30) and (41) jointly hold with probability at least $1 - 4\exp\left[-\left\{\eta n^{174}/\left(\kappa \alpha_{\mathscr{F}} J_{\mathscr{F}_\varepsilon}\right)\right\}^2\right]$. The lemma then follows if (41) implies (29). Indeed, for any non-zero $F \in \mathscr{F}$, its normalization $G = F/\|F\|_\infty$ is in $\mathscr{F}_\varepsilon$ by construction (25). Then (41) implies that

$$\frac{\|F\|_\infty}{\|F\|_{\mu_n, 1}} = 1/\left\|G\right\|_{\mu_n, 1} \leq \alpha_{\mathscr{F}}$$

because

$$1/\alpha_{\mathscr{F}} > \|G\|_{\mu, 1} - \|G\|_{\mu_n, 1} \geq 2/\alpha_{\mathscr{F}} - \|G\|_{\mu_n, 1},$$

where the last inequality follows from the definition of $\alpha_{\mathscr{F}}$ (28).

Thus it remains to establish (39) and (40), which can be done by applying the symmetrization and maximal inequality results in Sections 2.2 and 2.3.2 of [29]. Write $R_n(F) = (1/n)\sum_{i=1}^n l_i(F)$ where $l_i(F) = \int_0^1 Y_i(t) e^{F(t, X_i(t))} dt - \Delta_i F(T_i, X_i(T_i))$ are independent copies of the loss

$$l(F) = \int_0^1 Y(t) \cdot e^{F(t, X(t))} dt - \Delta \cdot F(T, X(T)), \tag{42}$$

which is a stochastic process indexed by $F \in \mathscr{F}$. As was shown in Proposition 2, $\mathbb{E}\{l(F)\} = R(F)$. Let $\zeta_1, \cdots, \zeta_N$ be independent Rademacher random variables that are

independent of $Z = \{(X_i(\cdot), Y_i(\cdot), T_i)\}_{i=1}^n$. It follows from the symmetrization Lemma 2.3.6 of [29] for stochastic processes that the left hand side of (39) is bounded by twice the Orlicz norm of

$$\begin{aligned}
&\sup_{F \in \mathscr{F}_\varepsilon^{\Psi_n}} \left| \frac{1}{n} \sum_{i=1}^n \zeta_i \{ l_i(F) - l_i(0) \} \right| \\
&\leq \frac{1}{n} \sup_{F \in \mathscr{F}_\varepsilon^{\Psi_n}} \left| \sum_{i=1}^n \zeta_i \int_0^1 Y_i(t) \left\{ e^{F(t, X_i(t))} - 1 \right\} dt \right| \\
&+ \frac{1}{n} \sup_{F \in \mathscr{F}_E^{\Psi_n}} \left| \sum_{i=1}^n \zeta_i \Delta_i F(T_i, X_i(T_i)) \right|.
\end{aligned} \tag{43}$$

Now hold Z fixed so that only $\zeta_1, \cdots, \zeta_n$ are stochastic, in which case the sum in the second line of (43) becomes a separable subgaussian process. Since the Orlicz norm of $\sum_{i=1}^n \zeta_i a_i$ is bounded by $\left(6 \sum_{i=1}^n a_i^2\right)^{1/2}$ for any constant a_i, we obtain the following the Lipschitz property for any $F_1, F_2 \in \mathscr{F}_\varepsilon^{\Psi_n}$:

$$\begin{aligned}
&\left\| \sum_{i=1}^n \zeta_i \int_0^1 Y_i(t) \left\{ e^{F_1(t, X_i(t))} - e^{F_2(t, X_i(t))} \right\} dt \right\|_{\Phi, \zeta \mid Z}^2 \\
&\leq 6 \sum_{i=1}^n \left[\int_0^1 Y_i(t) \left\{ e^{F_1(t, X_i(t))} - e^{F_2(t, X_i(t))} \right\} dt \right]^2 \\
&\leq 6 e^{2\Psi_n} \sum_{i=1}^n \left(\int_0^1 Y_i(t) \cdot |F_1(t, X_i(t)) - F_2(t, X_i(t))| dt \right)^2 \\
&\leq 6 e^{2\Psi_n} \sum_{i=1}^n \int_0^1 Y_i(t) \{ F_1(t, X_i(t)) - F_2(t, X_i(t)) \}^2 dt \\
&= 6ne^{2\Psi_n} \| F_1 - F_2 \|_{\mu_n, 2}^2,
\end{aligned}$$

where the second inequality follows from $|e^x - e^y| \leq e^{\max(x, y)} |x - y|$ and the last from the Cauchy-Schwarz inequality. Putting the Lipschitz constant $(6n)^{1/2} e^{\Psi_n}$ obtained above into Theorem 2.2.4 of [29] yields the following maximal inequality: There is a universal constant κ' such that

$$\begin{aligned}
&\left\| \sup_{F \in \mathscr{F}_\varepsilon^{\Psi_n}} \left| \sum_{i=1}^n \zeta_i \int_0^1 Y_i(t) \left\{ e^{F(t, X_i(t))} - 1 \right\} dt \right| \right\|_{\Phi, \zeta \mid Z} \\
&\leq \kappa' n^{1/2} e^{\Psi_n} \int_0^{\Psi_n} \left\{ \log \mathscr{N}\left(u, \mathscr{F}_\varepsilon^{\Psi_n}, \mu_n \right) \right\}^{1/2} du \\
&\leq \kappa' n^{1/2} e^{\Psi_n} \Psi_n J_{\mathscr{F}_\varepsilon},
\end{aligned}$$

where the last line follows from (26). Likewise the conditional Orlicz norm for the supremum of $\left|\sum_{i=1}^{n}\zeta_i \Delta_i F(T_i, X_i(T_i))\right|$ is bounded by $\kappa' J_{\mathscr{F}_\varepsilon} n^{1/2}\Psi_n$. Since neither bounds depend on Z, plugging back into (43) establishes (39):

$$\begin{aligned}&\left\|\sup_{F\in\mathscr{F}_\varepsilon^{\Psi_n}}\left|\{R_n(F)-R_n(0)\}-\{R(F)-R(0)\}\right|\right\|_{\Phi}\\&\le 2\kappa' J_{\mathscr{F}_\varepsilon}\frac{\Psi_n e^{\Psi_n}}{n^{1/2}}\left\{1+e^{-\Psi_n}\right\}\\&\le 4\kappa'\frac{J_{\mathscr{F}_\varepsilon}}{n^{1/4}},\end{aligned}$$

where $\Psi_n e^{\Psi_n} = n^{1/4}$ by (19). On noting that

$$\left\|G\right\|_{\mu_n,1}=\frac{1}{n}\sum_{i=1}^{n}\int_0^1 Y_i(t)\left|G(t,X_i(t))\right|dt,\qquad \left\|G\right\|_{\mu,1}=\mathbb{E}\left\{\int_0^1 Y(t)\left|G(t,X(t))\right|dt\right\},$$

(40) can be established using the same approach. □

Proof of Lemma 2.

Proof. For $m<\widehat{m}$, applying (8) to $R_n(F_{m+1})=R_n\left(F_m-\frac{\nu_n}{m+1}g^{\varepsilon}_{F_m}\right)$ yields

$$\begin{aligned}R_n(F_{m+1})&=R_n(F_m)-\frac{\nu_n}{m+1}\left\langle g_{F_m},g^{\varepsilon}_{F_m}\right\rangle_{\mu_n}\\&+\frac{\nu_n^2}{2(m+1)^2}\int\left(g^{\varepsilon}_{F_m}\right)^2\exp\{F_m-\rho(F_{m+1}-F_m)\}d\mu_n\\&<R_n(F_m)-\frac{\varepsilon\nu_n}{m+1}\|g_{F_m}\|_{\mu_n,2}+\frac{\nu_n^2 e^{\Psi_n}}{2(m+1)^2},\end{aligned}\tag{44}$$

where the bound for the second term is due to (18) and the bound for the integral follows from $\int\left(g^{\varepsilon}_{F_m}\right)^2 d\mu_n=1$ (Definition 1 of an ε-gradient) and $\|F_m\|_\infty, \|F_{m+1}\|_\infty<\Psi_n$ for $m<\widehat{m}$ (lines 5–6 of Algorithm 1). Hence for $m\le\widehat{m}$, (44) implies that

$$R_n(F_m)<R_n(0)+\sum_{m=0}^{\infty}\frac{\nu_n^2 e^{\Psi_n}}{2(m+1)^2}<R_n(0)+1\le 2$$

because $\nu_n^2 e^{\Psi_n}<1$ under (20). Since $\max_{m\le\widehat{m}}\|F_m\|_\infty<\Psi_n$, and using our assumption $\sup_{F\in\mathscr{F}_\varepsilon^{\Psi_n}}|R_n(F)-R(F)|<1$ in the statement of the lemma, we have

$$R(F_m)\le R_n(F_m)+|R_n(F_m)-R(F_m)|<3.$$

Clearly the minimizer F^* also satisfies $R(F^*)\le R(0)<3$. Thus coercivity (15) implies that

$$\|F_m\|_\infty, \|F^*\|_\infty < \alpha_{\mathscr{F}} \beta_\Lambda,$$

so the gap $\hat{\gamma}$ defined in (31) is bounded as claimed.

It remains to establish (32), for which we need only consider the case $R_n(F_{\hat{m}}) - R_n(F^*) > 0$. The termination criterion $gF_m = 0$ in Algorithm 1 is never triggered under this scenario, because by Proposition 1 this would imply that $F_{\hat{m}}$ minimizes $R_n(F)$ over the span of $\{\varphi_{nj}(t, x)\}_j$, which also contains F^* (Remark 3). Thus either $\hat{m} = \infty$, or the termination criterion $\left\| F_{\hat{m}} - \frac{\nu_n}{\hat{m}+1} \hat{g}^\varepsilon_{F_{\hat{m}}} \right\|_\infty \geq \Psi_n$ in line 5 of Algorithm 1 is met. In the latter case

$$\begin{aligned} \Psi_n &\leq \left\| F_{\hat{m}} - \frac{\nu_n}{\hat{m}+1} g^\varepsilon_{F_m} \right\|_\infty \leq \alpha_{\mathscr{F}} \left\| F_{\hat{m}} - \frac{\nu_n}{\hat{m}+1} g^\varepsilon_{F_m} \right\|_{\mu_n, 2} \\ &\leq \alpha_{\mathscr{F}} \left(\sum_{m=0}^{\hat{m}-1} \frac{\nu_n}{m+1} + 1 \right) \end{aligned} \tag{45}$$

where the inequalities follow from (29) and from $\left\| g^\varepsilon_{F_m} \right\|_{\mu_n, 2} = 1$. Since the sum is diverging, the inequality also holds for $\hat{m}$ sufficiently large (e.g. $\hat{m} = \infty$).

Given that F^* lies in the span of $\{\varphi_{nj}(t, x)\}_j$, the Taylor expansion (8) is valid for $R_n(F^*)$. Since the remainder term in the expansion is non-negative, we have

$$\begin{aligned} R_n(F^*) &= R_n(F_m + F^* - F_m) \\ &\geq R_n(F_m) + \left\langle g_{F_m}, F^* - F_m \right\rangle_{\mu_n}. \end{aligned}$$

Furthermore for $m \leq \hat{m}$,

$$\begin{aligned} \left\langle g_{F_m}, F_m - F^* \right\rangle_{\mu_n} &\leq \|F_m - F^*\|_{\mu_n, 2} \cdot \|g_{F_m}\|_{\mu_n, 2} \\ &\leq \|F_m - F^*\|_\infty \cdot \|g_{F_m}\|_{\mu_n, 2} \\ &\leq \hat{\gamma} \cdot \|g_{F_m}\|_{\mu_n, 2}. \end{aligned}$$

Putting both into (44) gives

$$\begin{aligned} R_n(F_{m+1}) &< R_n(F_m) + \frac{\varepsilon \nu_n}{\hat{\gamma}(m+1)} \left\langle g_{F_m}, F^* - F_m \right\rangle_{\mu_n} + \frac{\nu_n^2 e^{\Psi_n}}{2(m+1)^2} \\ &\leq R_n(F_m) + \frac{\varepsilon \nu_n}{\hat{\gamma}(m+1)} \{R_n(F^*) - R_n(F_m)\} + \frac{\nu_n^2 e^{\Psi_n}}{2(m+1)^2}. \end{aligned}$$

Subtracting $R_n(F^*)$ from both sides above and denoting $\delta_m = R_n(F_m) - R_n(F^*)$, we obtain

$$\delta_{m+1} < \left(1 - \frac{\varepsilon \nu_n}{\hat{\gamma}(m+1)}\right) \delta_m + \frac{\nu_n^2 e^{\Psi_n}}{2(m+1)^2}.$$

Since the term inside the first parenthesis is between 0 and 1, solving the recurrence yields

$$\begin{aligned}\delta_{\widehat{m}} &< \delta_0 \prod_{m=0}^{\widehat{m}-1}\left(1-\frac{\varepsilon \nu_n}{\widehat{\gamma}(m+1)}\right)+\nu_n^2 e^{\Psi_n}\sum_{m=0}^{\infty}\frac{1}{2(m+1)^2}\\ &\le \max\{0,\delta_0\}\exp\left(-\frac{\varepsilon}{\widehat{\gamma}}\sum_{m=0}^{\widehat{m}-1}\frac{\nu_n}{m+1}\right)+\nu_n^2 e^{\Psi_n}\\ &\le e\max\{0,\delta_0\}\exp\left(-\frac{\varepsilon}{\alpha_{\mathscr{F}}\widehat{\gamma}}\Psi_n\right)+\nu_n^2 e^{\Psi_n},\end{aligned}$$

where in the second inequality we used the fact that $0 \le 1 + y \le e^y$ for $|y| < 1$, and the last line follows from (45).

The Lambert function (19) in $\Psi_n = W(n^{1/4})$ is asymptotically $\log y - \log\log y$, and in fact by Theorem 2.1 of [14], $W(y) \ge \log y - \log\log y$ for $y \ge e$. Since by assumption $n \ge 55 > e^4$, the above becomes

$$\delta_{\widehat{m}} < e\max\{0,\delta_0\}\left(\frac{\log n}{4n^{1/4}}\right)^{\varepsilon/(\alpha_{\mathscr{F}}\widehat{\gamma})}+\nu_n^2 e^{\Psi_n}.$$

The last step is to control δ_0, which is bounded by $1 - R_n(F^*)$ because $R_n(F_0) = R_n(0) \le 1$. Then under the hypothesis $|R_n(F^*) - R(F^*)| < 1$, we have

$$\delta_0 \le 1 - R(F^*) + 1 < 2 - R(F^*).$$

Since (14) implies $R(F^*) \ge 2\Lambda_U \min\{0, 1 - \log(2\Lambda_U)\}$,

$$\delta_0 < 2 - R(F^*) \le 2 + 2\Lambda_U \max\{0, \log(2\Lambda_U) - 1\} < 2\beta_\Lambda.$$

□

Proof of Proposition 3.

Proof. Let $\delta = \log n/(4n^{1/4})$ which is less than one for $n \ge 55 > e^4$. Since $\alpha_{\mathscr{F}}, \widehat{\gamma} \ge 1$ it follows that

$$\delta < \left(\frac{\log n}{4n^{1/4}}\right)^{\varepsilon/(\alpha_{\mathscr{F}}\widehat{\gamma})}. \tag{46}$$

Now define the following probability sets

$$S_1 = \left\{\sup_{F \in \mathscr{F}_\varepsilon^{\Psi_n}} |\{R_n(F) - R_n(0)\} - \{R(F) - R(0)\}| < \delta/3\right\}$$

$$S_2 = \{|R_n(0) - R(0)| < \delta/3\}$$

$$S_3 = \{|R_n(F^*) - R(F^*)| < \delta/3\}$$

$$S_4 = \{(29) \text{ holds }\},$$

and fix a sample realization from $\cap_{k=1}^{4} S_k$. Then the conditions required in Lemma 2 are satisfied with $\sup_{F \in \mathcal{F}_\varepsilon^{\Psi_n}} |R_n(F) - R(F)| < 2\delta/3$, so $\hat{\gamma}$ (and hence $\|\hat{\lambda}_{\text{boost}}\|_\infty$) is bounded and (32) holds. Since Algorithm 1 ensures that $\|F_{\hat{m}}\|_\infty < \Psi_n$, we have $F_{\hat{m}} \in \mathcal{F}_\varepsilon^{\Psi_n}$ and therefore it also follows that $|R_n(F_{\hat{m}}) - R(F_{\hat{m}})| < 2\delta/3$. Combining (23) and (24) gives

$$\begin{aligned}
\|F_{\hat{m}} - F^*\|_{\mu,2}^2 &\le \frac{2}{\min_{t,x}\left(\lambda^* \wedge \hat{\lambda}_{\text{boost}}\right)}\left(\frac{2\delta}{3} + \frac{\delta}{3} + \{R_n(F_{\hat{m}}) - R_n(F^*)\}\right) \\
&< \frac{2}{\min_{t,x}\left(\lambda^* \wedge \hat{\lambda}_{\text{boost}}\right)}\left(\delta + 2e\beta_\Lambda\left(\frac{\log n}{4n^{1/4}}\right)^{\varepsilon/(\alpha_{\mathcal{F}}\hat{\gamma})} + \frac{1}{16}\cdot\frac{\log n}{4n^{1/4}}\right) \\
&< \frac{13\beta_\Lambda}{\min_{t,x}\left(\lambda^* \wedge \hat{\lambda}_{\text{boost}}\right)}\left(\frac{\log n}{4n^{1/4}}\right)^{\varepsilon/(\alpha_{\mathcal{F}}\hat{\gamma})},
\end{aligned}$$

where the second inequality follows from (32) and $v_n^2 e^{\Psi_n} = \log n/\left(64n^{1/4}\right)$, and the last from (46). Now, using the inequality $|e^x - e^y| \le \max(e^x, e^y)|x - y|$ yields

$$\|\hat{\lambda}_{\text{boost}} - \lambda^*\|_{\mu,2}^2 < 13\beta_\Lambda \frac{\max_{t,x}\left(\lambda^* \vee \hat{\lambda}_{\text{boost}}\right)^2}{\min_{t,x}\left(\lambda^* \wedge \hat{\lambda}_{\text{boost}}\right)}\left(\frac{\log n}{4n^{1/4}}\right)^{\varepsilon/(\alpha_{\mathcal{F}}\hat{\gamma})},$$

and the stated bound follows from $F^* = \log \lambda$ since $\mathcal{F}$ is correctly specified (Proposition 2).

The next task is to lower bound $\mathbb{P}\left(\cap_{k=1}^{4} S_k\right)$. It follows from Lemma 1 that

$$\mathbb{P}(S_1 \cap S_4) \ge 1 - 4\exp\left\{-\left(\frac{\log n}{12\kappa\alpha_{\mathcal{F}} J_{\mathcal{F}_\varepsilon}}\right)^2\right\}.$$

Bounds on $\mathbb{P}(S_2)$ and $\mathbb{P}(S_3)$ can be obtained using Hoeffding's inequality. Note from (2) that $R_n(0) = \sum_{i=1}^n \int_0^1 Y_i(t)dt/n$ and $R_n(F^*) = \sum_{i=1}^n l_i(F^*)/n$ for the loss $l(\cdot)$ defined in (42). Since $0 \le \int_0^1 Y_i(t)dt \le 1$ and $-\|F^*\|_\infty < l(F^*) \le \|e^{F^*}\|_\infty + \|F^*\|_\infty$,

$$\mathbb{P}(S_2) \ge 1 - 2\exp\left\{-2n^{1/2}\left(\frac{\log n}{12}\right)^2\right\}, \mathbb{P}(S_3) \ge 1 - 2\exp\left\{-2n^{1/2}\left(\frac{\log n}{36e^{\|F^*\|_\infty}}\right)^2\right\}.$$

By increasing the value of κ and/or replacing $J_{\mathcal{F}_\varepsilon}$ with $\max\left(1, J_{\mathcal{F}_\varepsilon}\right)$ if necessary, we can combine the inequalities to get a crude but compact bound:

$$\mathbb{P}\left\{ \cap_{k=1}^{4} S_k \right\} \geq 1 - 8\exp\left\{ -\left(\frac{\log n}{\kappa \alpha_{\mathcal{F}} J_{\mathcal{F}_\varepsilon} e^{\|F^*\|_\infty}} \right)^2 \right\}. \tag{47}$$

Finally, since $\|F^*\|_\infty = \|\log \lambda\|_\infty < \max\{|\log \Lambda_L|, |\log \Lambda_U|\}$, we can replace $e^{\|F^*\|_\infty}$ in the probability bound above by $\Lambda_L^{-1} \vee \Lambda_U$. □

Proof of Proposition 4.

Proof. It follows from (22) that λ^* is the orthogonal projection of λ onto $(\mathcal{F}, \langle \cdot, \cdot \rangle_\mu)$. Hence

$$\begin{aligned}
&\|\hat{\lambda}_{\text{boost}} - \lambda\|_{\mu,2}^2 = \|e^{F^*} - \lambda\|_{\mu,2}^2 + \|e^{F_m} - e^{F^*}\|_{\mu,2}^2 \\
&= \min_{F \in \mathcal{F}} \|e^F - \lambda\|_{\mu,2}^2 + \|e^{F_m} - e^{F^*}\|_{\mu,2}^2 \\
&\leq \min_{F \in \mathcal{F}} \|e^F - \lambda\|_{\mu,2}^2 + \max_{t,x} \left(\lambda^* \vee \hat{\lambda}_{\text{boost}}\right)^2 \|F_{\hat{m}} - F^*\|_{\mu,2}^2,
\end{aligned}$$

where the inequality follows from $|e^x - e^y| \leq \max(e^x, e^y)|x - y|$. Bounding the last term in the same way as Proposition 3 completes the proof. To replace $e^{\|F^*\|_\infty}$ in (47) by $\Lambda_L^{-1} \vee \Lambda_U$, it suffices to show that $\Lambda_L \leq \lambda^*(t, x) \leq \Lambda_U$. Since the value of λ^* over one of its piecewise constant regions B is $\int_B \lambda d\mu / \mu(B)$, the desired bound follows from (A1). We can also replace $\max_{t,x}\left(\lambda^* \vee \hat{\lambda}_{\text{boost}}\right)$ and $\min_{t,x}\left(\lambda^* \wedge \hat{\lambda}_{\text{boost}}\right)$ with $\max_{t,x}\left(\Lambda_U \vee \hat{\lambda}_{\text{boost}}\right)$ and $\min_{t,x}\left(\Lambda_L \wedge \hat{\lambda}_{\text{boost}}\right)$ respectively. □

REFERENCES

[1]. Aalen OO (1978). Nonparametric inference for a family of counting processes. Annals of Statistics 6 701–726.

[2]. Adelson K, Lee DKK, Velji S, Ma J, Lipka SK, Rimar J, Longley P, Vega T, Perez-Irizarry J, Pinker E and Lilenbaum R (2018). Development of Imminent Mortality Predictor for Advanced Cancer (IMPAC), a tool to predict short-term mortality in hospitalized patients with advanced cancer. Journal of Oncology Practice 14 e168–e175. [PubMed: 29206553]

[3]. Biau G and Cadre B (2017). Optimization by gradient boosting. arXiv preprint arXiv:1707.05023

[4]. Binder H and Schumacher M (2008). Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. BMC Bioinformatics 9 14. [PubMed: 18186927]

[5]. Blanchard G, Lugosi G and Vayatis N (2003). On the rate of convergence of regularized boosting classifiers. Journal of Machine Learning Research 4 861–894.

[6]. Breiman L (1997). Arcing the edge. U.C. Berkeley Dept. of Statistics Technical Report 486

[7]. Breiman L (1999). Prediction games and arcing algorithms. Neural Computation 11 1493–1517. [PubMed: 10490934]

[8]. Breiman L (2004). Population theory for boosting ensembles. Annals of Statistics 32 1–11.

[9]. Bühlmann P (2002). Consistency for L_2boosting and matching pursuit with trees and tree-type basis functions. In Research report 109. Seminar für Statistik, Eidgenössische Technische Hochschule (ETH).

[10]. Bühlmann P and Hothorn T (2007). Boosting algorithms: Regularization, prediction and model fitting. Statistical Science 22 477–505.

[11]. Bühlmann P and Yu B (2003). Boosting with the L_2 loss: Regression and classification. Journal of the American Statistical Association 98 324–339.

[12]. Ehrlinger J and Ishwaran H (2012). Characterizing L_2Boosting. Annals of Statistics 40 1074–1101.

[13]. Friedman JH (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics 29 1189–1232.

[14]. Hoorfar A and HASSANI M (2008). Inequalities on the Lambert W function and hyperpower function. Journal of Inequalities in Pure and Applied Mathematics 9.

[15]. Hothorn T (2019). Transformation boosting machines. Statistics and Computing 1–12.

[16]. Hothorn T, Bühlmann P, Dudoit S, Molinaro A and van der Laan MJ (2006). Survival ensembles. Biostatistics 7 355–373. [PubMed: 16344280]

[17]. Hothorn T and Zeileis A (2017). Transformation forests. arXiv preprint arXiv:1701.02110

[18]. Huang J, Ma S and Xie H (2006). Regularized estimation in the accelerated failure time model with high-dimensional covariates. Biometrics 62 813–820. [PubMed: 16984324]

[19]. Huang JZ and Stone CJ (1998). The L_2 rate of convergence for event history regression with time-dependent covariates. Scandinavian Journal of Statistics 25 603–620.

[20]. Lowsky DJ, Ding Y, Lee DKK, McCulloch CE, Ross LF, Thistlethwaite JR and Zenios SA (2013). A K-nearest neighbors survival probability prediction method. Statistics in Medicine 32 2062–2069. [PubMed: 23653217]

[21]. Mason L, Baxter J, Bartlett PL and Frean MR (1999). Functional gradient techniques for combining hypotheses. In Advances in Neural Information Processing Systems 221–246.

[22]. Mason L, Baxter J, Bartlett PL and Frean MR (2000). Boosting algorithms as gradient descent. In Advances in Neural Information Processing Systems 512–518.

[23]. Meyer DM, Rogers JG, Edwards LB, Callahan ER, Webber SA, Johnson MR, Vega JD, Zucker MJ and Cleveland JC Jr (2015). The future direction of the adult heart allocation system in the United States. American Journal of Transplantation 15 44–54. [PubMed: 25534445]

[24]. Müller HG and Yao F (2010). Additive modelling of functional gradients. Biometrika 97 791–805.

[25]. Nesterov Y (2004). Introductory lectures on convex optimization: A basic course. Springer.

[26]. Ridgeway G (1999). The state of boosting. Computing Science and Statistics 31 172–181.

[27]. Schmid M and Hothorn T (2008). Flexible boosting of accelerated failure time models. BMC Bioinformatics 9 269. [PubMed: 18538026]

[28]. Telgarsky M (2013). Margins, shrinkage, and boosting. In Proceedings of the 30th International Conference on International Conference on Machine Learning-Volume 28 II–307.

[29]. van der Vaart AW and Wellner JA (1996). Weak convergence and empirical processes with applications to statistics. Springer NY.

[30]. Vuong QH (1989). Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica 57 307–333.

[31]. Zhang T and Yu B (2005). Boosting with early stopping: convergence and consistency. Annals of Statistics 33 1538–1579.

[32]. Zou H, Hastie TJ and Tibshirani R (2007). On the ‘degrees of freedom’ of the lasso. Annals of Statistics 35 2173–2192.

Table 1

Relative importances of variables in the boosted nonparametric estimator. The numbers are scaled so that the largest value in each row is 1.

a	Time	Age	ESI	Census	All other variables
0	1	0.21	0.025	0.0011	<0.0010
1	1	0.22	0.013	0.46	<0.0003
2	0.34	0.064	0.0020	1	0
3	0.11	0.011	<0.0001	1	0

Table 2

Comparative performances (%MSE) as the service rate (38) becomes increasingly dependent on the time-varying ward census variable (by increasing a).

a	**blackboost (set to true log-normal distribution)**	***Transformation forest* (set to true log-normal distribution)**	***Boosted hazards (ε fixed for all iterations)***	***Ad-hoc (# splits fixed for all iterations)***
0	5.0%	5.0%	7.8%	7.1%
1	17%	6.1%	4.5%	8.1%
2	46%	9.7%	5.4%	7.0%
3	67%	18%	7.2%	7.4%