A Statistical Learning Take on the Concordance Index for Survival Analysis

Alex Nowak-Vila
Owkin Inc.

Kevin Elgui
Owkin Inc.

Geneviève Robin
Owkin Inc.

Abstract

The introduction of machine learning (ML) techniques to the field of survival analysis has increased the flexibility of modeling approaches, and ML based models have become state-of-the-art. These models optimize their own cost functions, and their performance is often evaluated using the concordance index (C-index). From a statistical learning perspective, it is therefore an important problem to analyze the relationship between the optimizers of the C-index and those of the ML cost functions. We address this issue by providing C-index Fisher-consistency results and excess risk bounds for several of the commonly used cost functions in survival analysis. We identify conditions under which they are consistent, under the form of three nested families of survival models. We also study the general case where no model assumption is made and present a new, off-the-shelf method that is shown to be consistent with the C-index, although computationally expensive at inference. Finally, we perform limited numerical experiments with simulated data to illustrate our theoretical findings.

1 INTRODUCTION

Survival analysis (Gross et al., 1981; Kalbfleisch and Prentice, 2002), the field of statistics concerned with modeling time-to-event data, is central to healthcare applications to predict time from diagnosis to death or risk of disease recurrence. Rather than directly modeling time-to-event, many survival models predict risk of event occurrence (Haider et al., 2020). Many definitions of risk can be found in the literature; the most classic are the expected time-to-event, the probability of an event occurring after a given time, or the multiplicative factor in the

hazard rate under the proportional hazards (PH) assumption. Importantly, survival data are often *right-censored*, and only a lower bound on the time-to-event is observed; it usually corresponds to the time at which patients leave the study. Most classical survival models have therefore been extended to the censored case (see, e.g. Klein and Moeschberger (2011) Chapter 3 for a review of the different types of censoring and Chapter 4 for survival estimation in the censored case).

Machine learning models are increasingly used in survival analysis and have shown state-of-the-art results in various application areas (Zhu et al., 2016, 2017; Yousefi et al., 2017; Katzman et al., 2018; Ching et al., 2018; Kvamme et al., 2019; Barnwal et al., 2020; Steingrimsson and Morrison, 2020; Cottin et al., 2022; Schutte et al., 2022). Evaluating this jungle of risk models is therefore an important issue to make it comprehensive for practitioners (Park et al., 2021). Among existing metrics in survival analysis, the concordance index (C-index) is probably the most commonly used (Harrell et al., 1996). It can be viewed as an extension of the Area Under the ROC Curve (AUC) for continuous outcomes and assesses the ability of a risk prediction method to correctly rank individuals according to their risk scores. More specifically, it is defined as the probability that pairs of predicted risks are ranked in the same order as the corresponding observed time-to-events.

From a statistical learning perspective, the question arises whether the C-index can be directly optimized, i.e., used as an objective function to be maximized in an ML approach. Unfortunately, the C-index is a non-concave, discontinuous loss with respect to the parameters of the risk model; consequently, gradient-based methods cannot be used directly to maximise it. In practice, models are often learned by minimizing a smooth surrogate loss on the training data and then evaluated using the C-index. Examples of training losses include the negative log-likelihood of survival models such as Cox (1972) Proportional Hazards (PH) or Accelerated Failure Time (AFT) models (Wei, 1992), loss functions defined as the expectation of an error measure between the time to event and the risk predictor (Steingrimsson and Morrison, 2020), or smooth approximations of the negated C-index (Chen et al., 2013).

Despite the widespread use of the C-index as an evaluation

measure, the relationship between optimizers of these training losses and those of the C-index is not well understood. In particular, it is not known under what conditions *Fisher consistency* (Fisher, 1922), also known as classification-calibration (Bartlett et al., 2006), holds, i.e., minimizers of the training loss correspond to optimizers of the C-index. If this property holds, we can safely say that the ML model converges to the optimal C-index as sample size grows to infinity—if the model is expressive enough.

The aim of this paper is to answer this very question. We study the consistency properties of classical cost functions in survival analysis with respect to the C-index, and provide associated excess risk bounds. We analyze in particular the properties of Maximum Likelihood Estimation (MLE), conditional average risk estimation, and smooth C-index maximization. We identify conditions under which these methods are consistent, under the form of three nested families of survival models. In addition, we study the more general case where no model assumption is made. In this case, we present a new, off-the-shelf convex method that is shown to be consistent with the C-index, although computationally expensive at inference. Finally, we perform limited numerical experiments with simulated data to illustrate our theoretical findings. In all cases, we discuss how censoring can be incorporated in our results. Note that, most of the theoretical results can be applied beyond survival analysis to any continuous ranking task in the sense of Clémençon and Achab (2018). Specifically, the following contributions are provided:

- The properties of commonly used risk estimation procedures in survival analysis are analyzed in terms of Fisher-consistency and C-index excess risk bounds.

- Conditions under which these procedures are Fisher-consistent with respect to the C-index are derived in the form of three nested families of survival models corresponding to increasingly stringent model assumptions. For each family, we characterize the maximizers of the C-index, and provide important examples.

- We discuss a novel, off-the-shelf convex estimation method which, although computationally expensive at inference, proves consistent without any modeling assumption.

- Limited experiments are conducted with simulated data to illustrate our theoretical findings.

Related Work

This work is in line with an extensive literature on the statistical efficiency of minimizing surrogate losses of non-convex and discontinuous evaluation metrics such as the 0-1 loss. Bartlett et al. (2006) derives upper bounds on the excess risk of convex surrogates for binary classification, while Agarwal (2014) provides similar results for surrogates of bipartite ranking losses. Cortes and Mohri (2003) provides a statistical analysis of the relationship between AUC and error rate minimization, and Gao and Zhou (2015) identifies sufficient conditions for consistency of pairwise surrogate losses with the AUC. From an optimization point of view, Calders and Jaroszewicz (2007) suggests maximizing the AUC directly using polynomial approximations.

While the C-index is widely used in survival analysis and several papers have investigated its properties from a practical point of view (Longato et al., 2020; Park et al., 2021), there are comparatively few statistical learning results evaluating its relationship to commonly used cost functions used in survival analysis. Steck et al. (2007) provide lower bounds on the C-index that can be directly optimized and examine their relationship to Cox's proportional hazards, showing that this popular model approximately maximizes the C-index. The authors do not examine consistency. Chen et al. (2013) develop a gradient-boosting procedure to optimize smooth surrogates of the C-index; however, the statistical consistency of such surrogates remains to be analyzed.

Since the C-index is fundamentally a ranking measure, our work has similarities with the extensive literature on ranking algorithms and their statistical properties (Clémençon et al., 2008; Duchi et al., 2010; Chapelle et al., 2011; Rajkumar and Agarwal, 2014; Yuan et al., 2016; He et al., 2018; Ai et al., 2019; Wu et al., 2021; Werner, 2021). As far as we are aware, none of these papers analyze ranking algorithms in terms of consistency with the C-index. Clémençon and Achab (2018) examines the optimizers of the C-index, but they focus only on the case where the conditional cumulative distribution functions do not cross one another; we consider much less restrictive assumptions.

2 SETTING AND BACKGROUND

2.1 Survival Analysis

Consider the classical survival analysis framework to model time-to-events and their relationship to individual covariates. The time-to-event, denoted T, is assumed continuous and takes values in $\mathbb{R}_+$; individual covariates are denoted by X and take values in $\mathcal{X} \subset \mathbb{R}^d$. Let $\mathrm{P} = \mathrm{Prob}(\mathbb{R}_+ \times \mathcal{X})$ be the space of joint probability densities μ of non-negative time-to-events and covariates; $\mu(t, x)$ is also referred to as the *survival model*. The density of events conditional on covariates x is denoted $\mu(t|x)$, and the conditional survival function

$$S(t|x) = \mathbb{P}\left\{T > t | X = x\right\} = \int_t^{+\infty} \mu(t|x)dt. \quad (1)$$

We also consider the right-censored setting where the time-to-event T is not directly observed but rather a lower bound $U = C \wedge T$ where C is a continuous, nonnegative random variable corresponding to the censoring time. The binary random variable $\Delta = 1(C \geq T)$ specifying whether the lower bounds corresponds to the time-to-event or to the censoring time is also observed. Throughout this paper, the censoring is assumed independent of the covariates, i.e. $C \perp X$, and the censoring curve is defined by $G(t) = \mathbb{P}\{C > t\}$.

2.2 Concordance Index

Consider a scalar valued function

$$f : x \in \mathbb{R}^d \mapsto f(x) \in \mathbb{R};$$

f may for instance come from an inference procedure assessing the risk of occurrence of events, depending on covariates x. In such settings, the higher $f(x)$, the smaller the time-to-event. Many quantities may be used to define the risk; for instance, the conditional expectation of the time-to-event $\mathbb{E}\{T|X = x\}$, the probability of an event occurring beyond a certain time t_0, $\mathbb{P}\{T > t_0|X = x\}$, or a statistic specific to a survival model such as the multiplicative factor of the baseline hazard under the PH assumption.

The C-index is defined as the probability of having a pairwise concordant order between the predicted risks and the observed time-to-events (Harrell et al., 1982). It is usually presented as the following conditional probability:

$$C(f) = \mathbb{P}\{f(X) < f(X') \mid T > T'\}. \quad (2)$$

The C-index depends on the joint distribution of (T, X), so that two survival models μ and μ' yield two different definitions of the C-index C_μ and $C_{\mu'}$. Whenever the model μ is clear from context, we drop the subscript for ease of notation. Since the C-index measures the quality of the ranking induced by f rather than the risk values themselves, it is defined up to monotone transformation of the risk.

For any pair of real random variables Y and Z, the *statistical preference* order (Taplin, 1997) denoted by $\succeq$, is defined as

$$Y \succeq Z \iff \mathbb{P}\{Y > Z\} \geq \frac{1}{2}. \quad (3)$$

A risk function f defining a global ordering maximally preserving the statistical preference (3) in expectation for all pairs of conditional random variables $T|X = x$ and $T|X = x'$ is optimal with respect to the C-index. This follows directly from the fact that:

$$\begin{aligned} C(f) &\propto \mathbb{P}\{f(X) < f(X'), T > T'\} \\ &= \mathbb{E}\, 1(f(X) < f(X'))1(T > T') \\ &= \mathbb{E}_{X,X'}\, \mathbb{P}\{T > T'|X, X'\}1(f(X) < f(X')). \end{aligned} \quad (4)$$

Definition 2.1 presents the case where there exists a global ordering respecting *all* pairwise comparisons under (3).

Definition 2.1 (Optimal risk ordering). *An optimal risk ordering is a function $f^\star$ satisfying*

$$f^\star(x) \leq f^\star(x') \Rightarrow \mathbb{P}\{T > T'|x, x'\} \geq \frac{1}{2}, \quad (5)$$

for all pairs x, x'. Note that if condition (5) is satisfied then it follows directly that $f^\star$ is an optimizer of the C-index and it only depends on the conditional density of events $\mu(t|x)$.

We show in Sec. 3 that an optimal risk ordering does not exists in general.

C-index estimation with right-censored data

In the survival analysis setting introduced in Sec. 2.1, time-to-events T are not observed but rather a lower bound $U = C \wedge T$ [1] and an event indicator $\Delta = 1(C \geq T)$. In this case the question arises of how to consistently estimate the C-index of a risk model f. Using the Inverse Censored Probability Weighting (IPCW) strategy of Robins and Finkelstein (2000), one can see that

$$\begin{aligned} \mathbb{E}\{1(T > T')\} &= \mathbb{E}\left\{\frac{1(T > T')1(C \wedge C' \geq T')}{G(T')^2}\right\} \\ &= \mathbb{E}\left\{\frac{\Delta' 1(U > U')}{G(U')^2}\right\}, \end{aligned}$$

where G is the censoring curve defined at the end of Sec. 2.1. In particular, this leads to the following expression for the C-index as an expectation over (X, U, Δ):

$$C(f) = \mathbb{E}\left\{\frac{\Delta' 1(U > U')1(f(X) < f(X'))}{G(U')^2}\right\}$$

The C-index can be consistently estimated from data using the empirical average instead of the expectation; this estimator is known as Uno's C-index (Uno et al., 2011).

2.3 Fisher Consistency

As previously discussed, the C-index (2) cannot be maximized using gradient descent. Instead, the function f is learned by minimizing a *smooth* risk $\mathcal{R}(f)$. Fisher consistency is a property guaranteeing that the minimizers of the smooth risk are also maximizers of the C-index. Definition 2.2 formalizes the notion of Fisher consistency over a family of distributions.

Definition 2.2 (Fisher Consistency). *The risk $\mathcal{R}$ is said to be Fisher consistent to the C-index under a distribution family $Q \subseteq P$ if*

$$\mathcal{R}_\mu(f^\star) = \min_f \mathcal{R}_\mu(f) \implies C_\mu(f^\star) = \max_f C_\mu(f),$$

for all distribution $\mu \in Q$, where $\mathcal{R}_\mu, C_\mu$ are the smooth risk and C-index computed over the joint distribution μ.

---

[1] $a \wedge b = \min(a, b)$.

In this paper, pairs of smooth risks and families of distributions are examined for which Fisher consistency holds for the C-index.

Consistency for AUC with binary labels The C-index is a continuous outcome version of the well-known AUC used in binary output ranking problems, also known as bipartite ranking. AUC is defined as the pairwise probability of concordant order between the risk and the binary labels:

$$\text{AUC}(f) = \mathbb{P}\{f(X) > f(X')|Y = 1, Y' = 0\}, \quad (6)$$

where $Y, Y' \in \{0, 1\}$ are binary. The loss is non-continuous and cannot be optimized directly by gradient descent. In this case, however, the set of optimizers of (6) can be easily characterized as monotone transformations of the conditional distribution (Agarwal, 2014). This can be seen as the function $f(x) = \mathbb{P}\{Y = 1|X = x\}$ satisfies

$$\mathbb{P}\{Y > Y'|x, x'\} \geq \frac{1}{2} \iff f(x) \geq f(x'),$$

which directly follows from the identity $f(x)(1 - f(x')) = \mathbb{P}\{Y > Y'|x, x'\}$. Hence, any smooth risk $\mathcal{R}$ whose minimizer is a monotone transformation of the conditional distribution is consistent to the AUC. This includes least squares, logistic regression, and more generally any proper loss function (Agarwal, 2014).

3 Maximizers of C-index

The problem of maximizing the C-index associated to a survival model μ writes

$$C_\mu^\star = \max_f C_\mu(f). \quad (7)$$

The aim of this section is to classify the set of possible survival models μ—the joint density of (T, X)—in terms of properties of the associated maximizers of (7). We introduce four families of survival models, denoted by $A \subsetneq B \subsetneq C$ and $D := C^c = P \setminus C$, defined informally as follows:

- A: Survival curves $S(t|x)$ do not cross. Existing work studying consistency with respect to C-index, such as Clémençon et al. (2013); Clémençon and Achab (2018) are limited to this family of models.

- B: $-\mathbb{E}\{T|X = x\}$ is an optimal risk ordering (5). We prove in Sec. 4 that a large family of smooth risk functions is consistent in this family.

- C: There exists an optimal risk ordering satisfying (5). We prove in Sec. 4 that MLE is Fisher-consistent for several examples of models in family C.

- D: There is no optimal risk ordering satisfying (5). We introduce in Sec. 4 an off-the-shelf estimation method which proves consistent in this setting, although being computationally expensive at inference.

In the following, the four families of survival models are described alongside examples and theoretical results characterizing the associated oracle C-index maximizers.

A. Conditional Survival Curves do not Cross

Family A is defined as the set of survival models whose conditional survival curves uniformly bound one another. In other words, models $\mu \in A$ satisfy assumption A.

Assumption A. *For all* $(x, x') \in \mathcal{X}^2$, $t \mapsto S(t|x) - S(t|x')$ *has constant sign.*

Under Assumption A, Thm. 3.1 shows that the negative conditional expectation $-\text{CE}(x) = \mathbb{E}\{T \mid X = x\}$ satisfies Condition (5); the proof can be found in the Appendix.

Theorem 3.1. *If* $\mu \in A$, *the negative conditional expectation is an optimal risk ordering for the C-index satisfying Condition* (5), *thus* $C_\mu(-\text{CE}) = C_\mu^\star$.

The two most commonly used survival models, namely Cox PH and AFT, satisfy Assumption A and therefore Thm. 3.1 applies, as discussed below.

Proportional Hazards model. The hazard function is defined as $h(t) = S'(t)(1 - S(t))^{-1}$, where S' denotes the notetime derivative of the survival curve. In the PH model, the conditional hazard $h(t|x)$ factorizes as $h(t|x) = h_0(t)e^{f(x)}$, where h_0 is the (non-negative) baseline hazard and f is a function of the covariates (Cox, 1972). This yields

$$S(t|x) = S_0(t)^{e^{f(x)}}, \quad S_0(t) = e^{-\int_0^t h_0(\tau)d\tau},$$

where S_0 is the baseline survival curve. It directly follows that survival curves do not cross at any point in time, therefore Assumption A is satisfied and Thm. 3.1 applies. In this example $f(x)$ also defines an optimal ranking; derivations are provided in the appendix to support this claim.

Accelerated Failure Times model. The AFT model assumes the following form for time-to-events:

$$\log T = f(X) + \varepsilon, \quad (8)$$

where ε is an independent random variable. The survival curve is parametrized as $S(t|x) = S_0(te^{-f(x)})$, where S_0 is the survival curve of e^ε. Note that the survival curves do not cross each other as they are defined as scaling by $e^{f(x)}$. In this example also, Assumption A is satisfied and Thm. 3.1 applies. As for the PH model, note that $f(x)$ also defines an optimal ranking; derivations are provided in the appendix to support this claim.

B. Conditional Expectation is an Optimal Ordering

Family B is defined as the set of survival models for which the negative conditional expectation is an optimal risk ordering. In other words, models $\mu \in B$ satisfy the following condition.

Assumption B. *The negative conditional expectation* $-\operatorname{CE}(x) = -\mathbb{E}\{T|X = x\}$ *is an optimal risk ordering satisfying* (5).

Note that Thm. 3.1 proved the inclusion A $\subset$ B. Now, the strict inclusion, A $\subsetneq$ B is proved by providing an example of survival models such that $\mu \in$ B $\setminus$ A. Consider the extended AFT model presented in the previous section by adding *symmetric heteroscedastic noise*.

Definition 3.2 (AFT-H). *In the AFT-H model, the time-to-event has the form*

$$\log T = f(x) + \sigma(x)\varepsilon, \qquad (9)$$

where $\sigma : \mathcal{X} \to \mathbb{R}_+$ *is a positive-valued function satisfying* $f(x) \leq f(x') \implies \sigma(x) \leq \sigma(x')$ *and* ε *is a centered Gaussian random variable.*

Prop. 3.3 shows that AFT-H satisfies assumption B.

Proposition 3.3 (AFT-H satisfies B). *Assume that μ is in AFT-H. Then, the negative conditional expectation is an optimal risk ordering, thus $C_\mu(-\operatorname{CE}) = C_\mu^\star$.*

This result is a reformulation of Corollary 2 by Lebedev (2019); we provide the original statement in the appendix. Note that under AFT-H the conditional survival curves take the following form

$$S(t|x) = S_\varepsilon\Big(\frac{\log t - f(x)}{\sigma(x)}\Big),$$

where $S_\varepsilon(X) = \mathbb{P}\{X > \varepsilon\}$. Fixing $f(x)$ and varying $\sigma(x)$ we can clearly see how the survival curves cross so that AFT-H *does not* always satisfy assumption A.

C. There exists an Optimal Risk Ordering

Family C is defined as the set of survival models admitting an optimal risk ordering satisfying (5).

Assumption C. *There exists an optimal ordering $f_\mu^\star$ for survival model μ, satisfying* (5).

Under assumption C, a closed form for the maximum C-index attained at $f^\star$ can be derived by combining Prop. 3.4 below to the expression of pairwise conditional probabilities for specific models.

Proposition 3.4. *Assume that μ satisfies assumption C. Then, the optimal C-index takes the following form:*

$$C_\mu^\star = C_\mu(f_\mu^\star) = \mathbb{E}_{X,X'}\,\varphi(\mathbb{P}\{T > T'|X,X'\}),$$

where $\varphi(a) = \max(a, 1-a)$ *and $f^\star$ satisfies* (5).

Analogously to the previous cases, a family of distributions satisfying assumption C is introduced, using exponential family models. Then, an example of model $\mu \in$ C $\setminus$ B is provided.

Definition 3.5 (Exponential family survival model). *For* $\theta : \mathcal{X} \to \mathbb{R}$, $\beta : \mathbb{R}_+ \to \mathbb{R}_+$, $\tau : \mathbb{R}_+ \to \mathbb{R}$, $\eta : \mathbb{R} \to \mathbb{R}$, *and* $A : \mathbb{R} \to \mathbb{R}$ *such that, for all $x \in \mathcal{X}$, the conditional density of a curved exponential family model is given by*

$$\mu(t|x) = \beta(t)\exp\left[\eta \circ \theta(x)\tau(t) - A \circ \theta(x)\right], \quad (10)$$

with associated parameter $\theta(x)$. For instance, $\theta(x) = \theta^\top x$ in a generalized linear model.

The scalar exponential family from definition 3.5 covers many of the survival curves classically used in survival analysis, e.g., the exponential, chi-squared, Laplace and normal distributions. Under the exponential family model, the scalar parameterization $\theta(x)$ satisfies the optimal risk ordering condition, as shown in the following proposition proved in the Appendix. Thus, maximization of the C-index can be achieved by estimating the parameters $\theta(x)$ of the model.

Proposition 3.6. *Under Assumption C, with θ continuous, β positive, τ non-decreasing and η continuously differentiable and non-decreasing, $\theta(x)$ is an optimal risk ordering for the C-index, thus $C_\mu(\theta) = C_\mu^\star$.*

Weibull with varying shape parameter Consider the model defined by Weibull conditional survival curves with varying shape parameter

$$S(t|x) = e^{-t^{f(x)}}. \qquad (11)$$

The following Prop. 3.7 proves the strict inclusion B $\subsetneq$ C.

Proposition 3.7. *The above Weibull model* (11) *satisfies assumption C but not assumption B.*

The proof is based on the result by Lebedev (2019) showing that f gives an optimal risk ordering. Assumption B is not satisfied as the expectation of a Weibull is not monotone on the shape parameter. Indeed, the expectation is given by $\Gamma(1 + \frac{1}{f(x)})$, where Γ denotes the Gamma function. This function has a minimum between 1.46 and 1.47, decreasing first and increasing for larger values. Thus, it gives a different ranking than the optimal risk ordering f.

D. There is no Optimal Risk Ordering

Family D contains survival models for which there does not exist an optimal risk ordering satisfying condition (5). The following example illustrates this phenomenon. Let $\{T_i\}_{1 \leq i \leq n}$ be random variables corresponding to time-to-event of individuals $1 \leq i \leq n$, and consider the following assumption.

Assumption D. *There exists $m \geq 3$, a subset of indices $\mathcal{I} \subset \{1,\ldots,n\}$, $|\mathcal{I}| = m$, and an ordering $i_1 < i_2 < \ldots < i_m$ such that, denoting $i_{m+1} = i_1$,*

$$\min_{k \in \{1,\ldots,m\}} \left(\mathbb{P}\{T_{i_k} < T_{i_{k+1}}\}\right) > \frac{1}{2}.$$
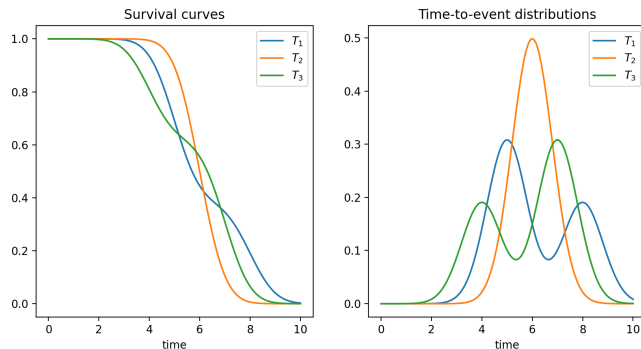
Figure 1: Survival curves and time-to-event distributions of a cycle of length three. In this case $P(T_1 < T_2) = P(T_2 < T_3) \approx 0.52$ and $(T_3 < T_1) \approx 0.55$. Hence, it is a cycle as $\min(P(T_1 < T_2), P(T_2 < T_3), P(T_3 < T_1)) > \frac{1}{2}$.

Assumption D implies the existence of a cyclic sequence with respect to the statistical preference order (3), which implies there is no ranking function satisfying the optimal risk ordering (5). In Fig. 1, we illustrate this phenomenon with a cyclic sequence made of a uni-modal and two multi-modal time-to-event distributions.

In the previous sections, we showed that for $\mu \in A, B, C$, the maximizer of the C-index in fact only depends on the *conditional density* $\mu(t|x)$; however, if $\mu \in D$, the maximizer may also depend on the marginal covariate distribution $\mu(x)$. This is an important characteristic of cyclic sequences, since the optimizer of the C-index may change under distributional shifts of the marginal population of patients. This phenomenon is shown in the following result, proved in the Appendix.

Proposition 3.8. *Under Assumption D, the maximizer of the C-index depends on the marginal distribution of the patients covariates $\mu(x)$.*

In particular, this means that the optimal relative order between patients may change if new patients are added to the original patient cohort. This phenomena does not happen in the binary setting where the ranking measure is the AUC as its optimizer is the conditional expectation, as discussed in Sec. 2.3.

4 CONSISTENCY AND EXCESS RISK BOUNDS

We now study several estimation procedures classically used in survival analysis, all based on Fisher-consistent, smooth cost functions. In particular, we discuss under which families of survival models introduced in the previous section Fisher-consistency holds. Although existing work was limited to family A (e.g. Clémençon et al. (2013); Clémençon and Achab (2018)), we prove that all the considered methods are consistent in family $B \supsetneq A$. We start

by introducing the considered method and their extensions to the censoring case. Then, we provide excess risk bounds on the associated C-index suboptimality.

4.1 Estimation Procedures

Estimating the conditional expectation (A, B). Without a specific survival model but under assumption B, one can use any cost function minimized by a monotone transformation of the conditional expectation. The following Thm. 4.1 provides a family of risks based on Fenchel-Young losses (Blondel et al., 2020) satisfying this property.

Theorem 4.1. *Let $\Omega : \mathcal{C} \to \mathbb{R}$ be a twice-differentiable strongly convex function [2] defined in a closed domain $\mathcal{C} \supseteq \mathbb{R}_+$ such that $\lim_{u \to \infty} \nabla \Omega(u) = +\infty$. Define the cost function $S(v,t) = \Omega^*(v) - vt$ where Ω^* is the Fenchel conjugate of Ω (Rockafellar, 1997) [3]. Then, the following risk*

$$\mathcal{R}(f) = \mathbb{E}_{(X,U,\Delta)} \; \frac{\Delta S(f(X), U)}{G(U)}, \qquad (12)$$

is convex, smooth, and its minimizer is a monotone transformation of the conditional expectation. Thus, under assumption B it is Fisher consistent to the C-index.

When $\mathcal{C} = \mathbb{R}$ and $\Omega(u) = u^2/2$ this corresponds precisely to ICPW (Robins and Finkelstein, 2000) on least squares. However, Thm. 4.1 provides a larger family of consistent smooth risks by choosing Ω and its domain $\mathcal{C}$ using the construction of Fenchel-Young losses.

Note that the estimator minimizing (12) is not efficient in the presence of censoring as only samples corresponding to events $\Delta = 1$ contribute to the cost function. To alleviate this issue, Steingrimsson and Morrison (2020) uses semi-parametric efficiency theory for missing data (Tsiatis, 2006; Robins et al., 1994), and develops an augmented estimator with smallest asymptotic variance among all unbiased estimators of $\mathbb{E} \; S(f(X), T)$.

Maximum Likelihood Estimators (A, B, C). Assume the conditional survival model $\mu_f(t|x)$ lies in a family of distributions so that one of the parameters $f(x)$ gives the optimal risk ranking (5), thus belonging to class C. We can learn the optimal parameter under censoring by minimizing the following MLE loss (Kalbfleisch and Prentice, 2002):

$$\mathcal{R}(f) = -\mathbb{E}_{(X,U,\Delta)} \; \Delta \log \mu_f(U|X) + (1-\Delta) \log S_f(U|X).$$

The derivation of this loss can be found in the Appendix. This loss can be used for all models presented in the previous section, namely PH, AFT, AFT-H, Weibull and the exponential family. Whenever the model is identifiable, the

---

[2] A one-dimensional strongly convex function is one for which the Hessian is uniformly lower bounded $\nabla_u^2 \Omega \geq C > 0$ for all u in the domain.

[3] The Fenchel conjugate Ω^* of Ω is defined for all $v \in \mathbb{R}$ as $\Omega^*(v) = \sup_{u \in \mathcal{C}} \; vu - \Omega(u)$
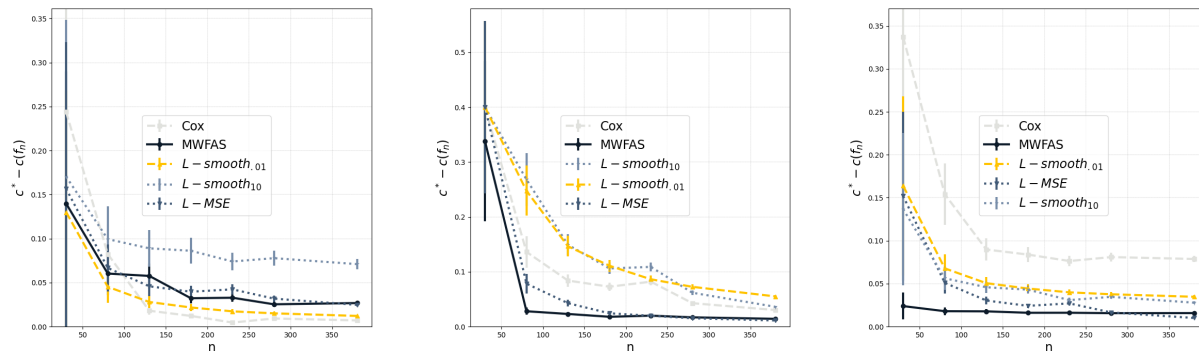
Figure 2: C-index empirical excess risks of the different methods of ranking for various training sizes n. The left, center and right figures correspond to data generated from models satisfying assumptions A, B and C, respectively.

MLE is consistent to the true parameter and thus Fisher consistent to the C-index. An important advantage of this loss is that the censored samples have an explicit role and provide signal during the learning procedure. Note that, penalized versions of (4.1) can also be used to obtain explicit finite sample risk bounds (see Prop. 4.4).

Smooth C-index (A, B, C, D). In family D, one cannot assume the existence of an optimal risk ordering. In this case, an alternative method is smooth C-index maximization, which is based on smoothing the indicator function defining the C-index under censoring as presented in Sec. 2.2, leading to a non-convex smooth loss (Mayr et al., 2016). The non-convexity of this approach may cause convergence problems and a scaling parameter may be tuned to guarantee proper convergence.

Estimating pairwise probabilities (A, B, C, D). The smooth C-index has two-main problems: its non concavity and the fact its optimizer is not robust to marginal distribution shifts due to Prop. 3.8. We propose a novel methodology whereby instead of learning a risk function f specifying the ranking on the training cohort, we (1) learn the pairwise conditional probabilities $\mathbb{P}\{T > T'|x, x'\}$ on the training data with an estimator $h(x, x')$ and then (2) we construct the ranking that better satisfies the relative order constraints in expectation over the finite validation cohort $x_1, \ldots, x_n$ by solving:

$$\min_{\sigma \in \mathbb{S}} \sum_{i,j=1}^{n} \gamma_{ji} \mathbb{1}(\sigma(i) < \sigma(j)), \quad (13)$$

where $\mathbb{S}$ is the set of permutations of size n and $\gamma_{ij} = |2h(x_i, x_j) - 1| \mathbb{1}(h(x_i, x_j) > 1/2)$. The derivation of this methodology follows easily from the C-index expression (4) and it is consistent by construction. The combinatorial problem (13) is known as the Minimum Weight Feedback Arc Set (MWFAS) problem (Duchi et al., 2010; Karp, 1972); it is known to be NP-Hard. However, multi-

ple approximations of this problem exist (Even et al., 1998; Demetrescu and Finocchi, 2003).

This method addresses the two problems of the smooth C-index. First, the pairwise conditional probabilities can be learned using convex estimation methods, such as logistic regression on the binary problem $\tilde{Y} = \mathrm{sign}(T - T')$ and $\tilde{X} = (X, X')$. Second, the ranking estimator is robust to marginal distributional shifts as the inference algorithm (13) is computed on the validation cohort. It is interesting to note that the computational bottleneck of the smooth C-index coming from its non concavity has now been transposed into the computational bottleneck of the combinatorial inference problem.

4.2 Excess Risk Bounds

The question is now to obtain excess risk bounds on the C-index when there exists an optimal risk ranking $f^\star$ satisfying (5). The following Thm. 4.2 bounds the excess risk of the C-index.

Theorem 4.2 (Excess risk bounds)**.** *Let $f^\star$ be an optimal risk ranking satisfying (5). Let $L > 0$ a positive constant satisfying*

$$|2\mathbb{P}\{T > T'|x, x'\} - 1| \leq L|f^\star(x) - f^\star(x')|, \quad (14)$$

for all pairs x, x'. Then, the excess risk of the C-index can be bounded as

$$C(f^\star) - C(f) \leq 2L\,\mathbb{E}_X\,|f(X) - f^\star(X)|.$$

The proof of this result can be found in the Appendix, and is based on a reduction of the problem to a binary classification problem with input (X, X') and output $Z = \mathrm{sign}(T - T')$, similar to Agarwal (2014) for the bipartite ranking setting. The following Prop. 4.3 shows that condition (14) is satisfied for most of the models presented in Sec. 3. The proof is in the Appendix.

Proposition 4.3. *Condition* (14) *is satisfied by:*

(i) *The PH model with $L = 1$.*

(ii) *The AFT model* (8) *with L the Lipschitz constant of the cumulative distribution function $F_{\varepsilon - \varepsilon'}$ of the symmetric random variable $\varepsilon - \varepsilon'$.*

(iii) *The AFT-H model* (9) *with L the same as AFT scaled by a factor a where $\sigma(x) \geq 1/a$ for all x.*

(iv) *The exponential family* (10).

We now provide two examples of application of Prop. 4.3.

Example of Lasso estimator on Cox PH model Prop. 4.3 shows that Thm. 4.2 applies in particular to the PH model. For this specific case, we illustrate our theoretical findings with an application to the Lasso estimator for the Cox PH model. This estimator is analyzed in Huang et al. (2013), where finite sample bounds on the ℓ_1-penalized negative log-likelihood estimation and prediction errors are proven. More specifically, assuming a generalized linear model $f^\star(x) = \theta^{\star\top} x$, the Lasso estimator $\hat{\theta}_{\text{Lasso}}$ is shown to satisfy $\|\hat{\theta}_{\text{Lasso}} - \theta^\star\|_1 \lesssim \frac{\|\theta^\star\|_0 \log(d)}{n}$ in high probability for sufficiently large training datasets. Operator $\lesssim$ denotes inequality up to log and constant terms depending on the model, and $\|\theta^\star\|_0$ denotes the number of non-zero entries of $\theta^\star$. Combining this result with Thm. 4.2, and noticing that

$$\mathbb{E}_X \left| (\hat{\theta}_{\text{Lasso}} - \theta^\star)^\top X \right| \leq \|\hat{\theta}_{\text{Lasso}} - \theta^\star\|_1 \, \mathbb{E}_X \|X\|_\infty,$$

we obtain the following informal result.

Proposition 4.4 (informal). *With probability at least $1 - \varepsilon(n)$, $\varepsilon(n) \to 0$ as $n \to +\infty$,*

$$C(\theta^\star) - C(\hat{\theta}_{\text{Lasso}}) \lesssim \mathbb{E}_X \|X\|_\infty \cdot \frac{\|\theta^\star\|_0 \log(d)}{n}.$$

The quantity $\mathbb{E}_X \|X\|_\infty$ depends on the covariates model and corresponds to "the size" of the input space.

Finite sample bounds for family B. Under assumption B and using the smooth cost function (12) we can obtain excess risk bounds on the C-index in terms of the excess of the smooth risk $\mathcal{R}(f)$.

Theorem 4.5. *Let $\mathcal{R}$ be the risk defined in* (12)*. Under assumption B, the following inequality holds:*

$$C^\star - C(f) \leq 4L\gamma \sqrt{\mathcal{R}(f) - \mathcal{R}^\star},$$

where $\Omega''(u) \geq 1/\gamma^2$ for all u in the domain $\mathcal{C}$ [4].

Combining the above Thm. 4.5 with finite sample bounds on the excess risk such as the ones obtained by Ausset et al. (2019) one can translate them to the C-index.

---

[4]Recall that Ω is convex thus $\Omega''(u) \geq 0$

5 EXPERIMENTS

In this section we perform experiments to validate our theoretical findings [5]. More specifically, we assess empirically the consistency of different estimation methods with respect to the C-index, under simulation regimes corresponding to families A, B and C.

Data Generation and Evaluation Procedure We simulate survival data using the three different regimes A, B and C presented in Sec. 3. We first simulate the training covariates $x_1, \ldots, x_n \in \mathbb{R}^d$ as well as a unit vector $\beta \in \mathbb{R}^d$ with $d = 10$ and parameterize the optimal ranking linearly $f(x) = \beta^\top x$. We then simulate the corresponding time-to-events $t_1, \ldots, t_n$ as realizations of the distribution of $T|X$, parameterized by β and depending on the selected regime of simulation. Four different models are fit using the the cost functions studied in Sec. 4. First, a linear model (**L-MSE**) optimizing a mean square error loss function to regress the conditional expectation of T given x; second, a linear Cox model (**Cox**) optimizing a log-likelihood; third, a linear model (**L-smooth$_\sigma$**) optimizing a smooth C-index (with smoothing parameter $\sigma \in \{.01, 10\}$ that accounts for the smoothness of the approximation of the indicator function), and finally a pairwise model (**MWFAS**) that predicts pairwise probabilities using XGBoost, and from which the ranking is obtained by solving the MWFAS combinatorial problem presented in Sec. 4.1 with a fast approximation algorithm. We compute the obtained C-index from a test dataset of fixed size ($n_{\text{test}} = 3000$) using the same distribution as the training dataset.

Results The results are provided in Fig. 2. For the three generation regimes, we observe that the proposed method **MWFAS** yields the best performance and converges to the optimal C-index when n is sufficiently large. It has similar performance to **Cox** in the regime A where **Cox** is well specified. When **Cox** is not well-specified it does not converge to the optimal ranking. Lastly, note that the value of the smoothing parameter used by **L-smooth$_\sigma$** can harm the performance of the resulting model and it is important to choose it properly to improve convergence guarantees.

Acknowledgements

References

Agarwal, S. (2014). Surrogate regret bounds for bipartite ranking via strongly proper losses. *The Journal of Machine Learning Research*, 15(1):1653–1674.

Ai, Q., Wang, X., Bruch, S., Golbandi, N., Bendersky, M.,

---

[5]Code can be found in https://github.com/owkin/owkin-metric

and Najork, M. (2019). Learning groupwise multivariate scoring functions using deep neural networks. In *Proceedings of the 2019 ACM SIGIR international conference on theory of information retrieval*, pages 85–92.

Ausset, G., Clémençon, S., and Portier, F. (2019). Empirical risk minimization under random censorship: Theory and practice. *arXiv preprint arXiv:1906.01908*.

Barnwal, A., Cho, H., and Hocking, T. D. (2020). Survival regression with accelerated failure time model in xgboost. *arXiv preprint arXiv:2006.04920*.

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156.

Blondel, M., Martins, A. F., and Niculae, V. (2020). Learning with fenchel-young losses. *Journal of Machine Learning Research*, 21(35):1–69.

Calders, T. and Jaroszewicz, S. (2007). Efficient auc optimization for classification. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 42–53. Springer.

Chapelle, O., Chang, Y., and Liu, T.-Y. (2011). Future directions in learning to rank. In *Proceedings of the Learning to Rank Challenge*, pages 91–100. PMLR.

Chen, Y., Jia, Z., Mercola, D., and Xie, X. (2013). A gradient boosting algorithm for survival analysis via direct optimization of concordance index. *Computational and mathematical methods in medicine*, 2013.

Ching, T., Zhu, X., and Garmire, L. X. (2018). Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data. *PLOS Computational Biology*, 14(4):e1006076.

Clémençon, S. and Achab, M. (2018). Ranking data with continuous labels through oriented recursive partitions. *arXiv preprint arXiv:1801.05772*.

Clémençon, S., Lugosi, G., and Vayatis, N. (2008). Ranking and empirical minimization of u-statistics. *The Annals of Statistics*, 36(2):844–874.

Clémençon, S., Robbiano, S., and Vayatis, N. (2013). Ranking data with ordinal labels: optimality and pairwise aggregation. *Machine Learning*, 91(1):67–104.

Cortes, C. and Mohri, M. (2003). Auc optimization vs. error rate minimization. *Advances in neural information processing systems*, 16:313–320.

Cottin, A., Pecuchet, N., Zulian, M., Guilloux, A., and Katsahian, S. (2022). Idnetwork: A deep illness-death network based on multi-state event history process for disease prognostication. *Statistics in Medicine*, 41(9):1573–1598.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

Demetrescu, C. and Finocchi, I. (2003). Combinatorial algorithms for feedback problems in directed graphs. *Information Processing Letters*, 86(3):129–136.

Duchi, J. C., Mackey, L. W., and Jordan, M. I. (2010). On the consistency of ranking algorithms. In *ICML*, pages 327–334.

Even, G., Schieber, B., Sudan, M., et al. (1998). Approximating minimum feedback sets and multicuts in directed graphs. *Algorithmica*, 20(2):151–174.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society A*, 222:309–368.

Gao, W. and Zhou, Z.-H. (2015). On the consistency of auc pairwise optimization. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, page 939–945. AAAI Press.

Gross, A. J., Elandt-Johnson, R. C., and Johnson, N. L. (1981). Survival models and data analysis. *BioScience*.

Haider, H., Hoehn, B., Davis, S., and Greiner, R. (2020). Effective ways to build and evaluate individual survival distributions. *J. Mach. Learn. Res.*, 21:85–1.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. (1982). Evaluating the yield of medical tests. *Jama*, 247(18):2543–2546.

Harrell, F. E., Lee, K. L., and Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15 4:361–87.

He, Y., Liu, J., and Ning, X. (2018). Drug selection via joint push and learning to rank. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 17(1):110–123.

Huang, J., Sun, T., Ying, Z., Yu, Y., and Zhang, C.-H. (2013). Oracle inequalities for the lasso in the Cox model. *The Annals of Statistics*, 41(3).

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, 2nd edition.

Karp, R. M. (1972). Reducibility among combinatorial problems. In *Complexity of computer computations*, pages 85–103. Springer.

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., and Kluger, Y. (2018). DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1):24.

Klein, J. P. and Moeschberger, M. L. (2011). *Survival Analysis : Techniques for Censored and Truncated Data*. New York: Springer, second edition.

Kvamme, H., Borgan, Ø., and Scheel, I. (2019). Time-to-Event Prediction with Neural Networks and Cox Regression. *Journal of Machine Learning Research*, 20(129):1–30.

Lebedev, A. V. (2019). The nontransitivity problem for three continuous random variables. *Automation and Remote Control*, 80(6):1058–1068.

Longato, E., Vettoretti, M., and Di Camillo, B. (2020). A practical perspective on the concordance index for the evaluation and selection of prognostic time-to-event models. *Journal of Biomedical Informatics*, 108:103496.

Mayr, A., Hofner, B., and Schmid, M. (2016). Boosting the discriminatory power of sparse survival models via optimization of the concordance index and stability selection. *BMC bioinformatics*, 17(1):1–12.

Park, S. Y., Park, J. E., Kim, H., and Park, S. H. (2021). Review of statistical methods for evaluating the performance of survival or other time-to-event prediction models (from conventional to deep learning approaches). *Korean Journal of Radiology*, 22:1697 – 1707.

Rajkumar, A. and Agarwal, S. (2014). A statistical convergence perspective of algorithms for rank aggregation from pairwise data. In *International conference on machine learning*, pages 118–126. PMLR.

Robins, J. and Finkelstein, D. (2000). Correcting for noncompliance and dependent censoring in an aids clinical trial with inverse probability of censoring weighted (ipcw) log-rank tests. *Biometrics*, 56(3):779—788.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866.

Rockafellar, R. T. (1997). Convex analysis. princeton landmarks in mathematics.

Schutte, K., Brulport, F., Harguem-Zayani, S., Schiratti, J.-B., Ghermi, R., Jehanno, P., Jaeger, A., Alamri, T., Naccache, R., Haddag-Miliani, L., Orsi, T., Lamarque, J.-P., Hoferer, I., Lawrance, L., Benatsou, B., Bousaid, I., Azoulay, M., Verdon, A., Bidault, F., Balleyguier, C., Aubert, V., Bendjebbar, E., Maussion, C., Loiseau, N., Schmauch, B., Sefta, M., Wainrib, G., Clozel, T., Ammari, S., and Lassau, N. (2022). An artificial intelligence model predicts the survival of solid tumour patients from imaging and clinical data. *European Journal of Cancer*, 174:90–98.

Steck, H., Krishnapuram, B., Dehing-Oberije, C., Lambin, P., and Raykar, V. C. (2007). On ranking in survival analysis: Bounds on the concordance index. *Advances in neural information processing systems*, 20.

Steingrimsson, J. A. and Morrison, S. (2020). Deep learning for survival outcomes. *Statistics in medicine*, 39(17):2339–2349.

Taplin, R. H. (1997). The statistical analysis of preference data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(4):493–512.

Tsiatis, A. A. (2006). Semiparametric theory and missing data.

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., and Wei, L.-J. (2011). On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, 30(10):1105–1117.

Wei, L. J. (1992). The accelerated failure time model: A useful alternative to the cox regression model in survival analysis. *Statistics in Medicine*, 11(14-15):1871–1879.

Werner, T. (2021). A review on instance ranking problems in statistical learning. *Machine Learning*, pages 1–49.

Wu, X., Zeng, W., Lin, F., and Zhou, X. (2021). Neurank: learning to rank with neural networks for drug–target interaction prediction. *BMC bioinformatics*, 22(1):1–17.

Yousefi, S., Amrollahi, F., Amgad, M., Dong, C., Lewis, J. E., Song, C., Gutman, D. A., Halani, S. H., Velazquez Vega, J. E., Brat, D. J., and Cooper, L. A. D. (2017). Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. *Scientific Reports*, 7(1):11707.

Yuan, Q., Gao, J., Wu, D., Zhang, S., Mamitsuka, H., and Zhu, S. (2016). Druge-rank: improving drug–target interaction prediction of new candidate drugs or targets by ensemble learning to rank. *Bioinformatics*, 32(12):i18–i27.

Zhu, X., Yao, J., and Huang, J. (2016). Deep convolutional neural network for survival analysis with pathological images. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 544–547, Shenzhen, China. IEEE.

Zhu, X., Yao, J., Zhu, F., and Huang, J. (2017). WSISA: Making Survival Prediction from Whole Slide Histopathological Images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6855–6863, Honolulu, HI. IEEE.

A Statistical Learning Take on the Concordance Index for Survival Analysis: Supplementary Materials

Outline. The supplementary material is organized as follows. In Section 1 we provide the proofs of the results from Section 3 of the main paper and in Section 2 we provide the proofs of the results from Section 4 of the main paper

1 PROOF OF RESULTS FROM SECTION 3

Let's first recall the definition of an optimal risk ordering and the assumption A, B, C and D presented in Section 3 of the paper.

Definition 1.1 (Optimal risk ordering). *An optimal risk ordering is a function $f^\star$ satisfying*

$$f^\star(x) \leq f^\star(x') \Rightarrow \mathbb{P}\{T > T'|x, x'\} \geq \frac{1}{2}, \tag{1}$$

for all pairs x, x'. Note that if condition (1) is satisfied then it follows directly that $f^\star$ is an optimizer of the C-index and it only depends on the conditional density of events $\mu(t|x)$.

Assumption A. *For all $(x, x') \in \mathcal{X}^2$, $t \mapsto S(t|x) - S(t|x')$ has constant sign.*

Assumption B. *The negative conditional expectation $- \mathrm{CE}(x) = - \mathbb{E}\{T|X = x\}$ is an optimal risk ordering satisfying (1).*

Assumption C. *There exists an optimal ordering $f_\mu^\star$ for survival model μ, satisfying (1).*

Assumption D. *There is not an optimal ordering $f_\mu^\star$ for survival model μ, satisfying (1).*

1.1 Proof of Theorem 3.1

We start with the following Lemma 1.2 from Levy and Robinson (2006).

Lemma 1.2 (Levy and Robinson (2006)). *Let T, T' be two random variables with survival functions S_T and $S_{T'}$, respectively. We have that $S_T(t) \geq S_{T'}(t)$ for all $t \geq 0$ if and only if $\mathbb{E}_{T'}\varphi(T) \geq \mathbb{E}_{T'}\varphi(T')$ for any non-decreasing function $\varphi : \mathbb{R} \to \mathbb{R}$.*

Equipped with Lemma 1.2, we can state the following proposition.

Proposition 1.3. *Let T, T' be two independent continuous random variables.*

$$S_T(t) \geq S_{T'}(t) \ \text{for all} \ t \geq 0 \implies \mathbb{P}\{T > T'\} \geq \frac{1}{2}.$$

Proof. Since the random variables are continuous, $\mathbb{P}\{T \geq T'\} \geq \frac{1}{2}$ is equivalent to $\mathbb{P}\{T \geq T'\} \geq \mathbb{P}\{T' \geq T\}$. In addition,

$$\begin{aligned}
\mathbb{P}\{T \geq T'\} &= \mathbb{E}\, 1(T \geq T') \\
&= \mathbb{E}_{T'}\, \mathbb{E}_T\{1(T \geq T')|T'\} \\
&= \mathbb{E}_{T'}\, \mathbb{E}_T\{1(T \geq T')\} = \mathbb{E}_{T'} S_T(T'),
\end{aligned}$$

where we used the independence of T, T'. Thus, $\mathbb{P}\{T \geq T'\} \geq \mathbb{P}\{T' \geq T\}$ also equivalent to

$$\mathbb{E}_{T'} S_T(T') \geq \mathbb{E}_T S_{T'}(T). \tag{2}$$

On the other hand, since the survival curves are uniformly bounded, it holds that $\mathbb{E}_{T'}\, S_T(T') \geq \mathbb{E}_{T'}\, S_{T'}(T')$. Applying Lemma 1.2 to $-S_{T'}$ which is non-decreasing, we obtain that $\mathbb{E}_{T'}\, S_{T'}(T') \geq \mathbb{E}_T S_{T'}(T)$, which concludes the proof. $\square$

Recall the statement of Theorem 3.1 below.

Theorem 3.1. *If μ satisfies assumption A, the negative conditional expectation is an optimal risk ordering for the C-index satisfying Condition (1), thus $C_\mu(- \mathrm{CE}) = C_\mu^\star$.*

Proof. Note that, under Assumption A, for any x, x', $- \mathbb{E}\{T|X = x\} \leq - \mathbb{E}\{T'|X = x'\}$ if and only if $S(t|x) \geq S(t|x')$ for any $t \geq 0$. Using that $\{T|X = x\}$ and $\{T|X = x'\}$ are independent continuous random variables and applying Prop. 1.3 yields the desired result. $\square$

1.2 Proof of Proposition 3.3

The AFT-H model has the following form:

$$\log T = f(x) + \sigma(x)\varepsilon, \tag{3}$$

where ε is a standard Gaussian random variable. Recall the statement of Proposition 3.3.

Proposition 3.3 (AFT-H satisfies B). *Assume μ is in AFT-H and that if $f(x) < f(x')$ then $\sigma(x) < \sigma(x')$. Then, the negative conditional expectation is an optimal risk ordering, thus $C_\mu(-\,\mathrm{CE}) = C_\mu^\star$.*

Proof. We will first show that the negative conditional expectation $-\mathbb{E}\{\log T \mid X = x\}$ is an optimal risk ordering satisfying (1). Following the proof of Corollary 2 by Lebedev (2019), we show that if T, T' are defined as $\log T = a + b\varepsilon, \log T' = a' + b'\varepsilon'$ where $\varepsilon, \varepsilon'$ are symmetric around zero and independent, then $\mathbb{P}\{T > T'\} \geq \frac{1}{2}$ if and only if $a \geq a'$. If we denote by F_{12} the c.d.f of $b\varepsilon - b'\varepsilon'$, which is symmetric around zero too, then

$$\mathbb{P}\{T > T'\} = \mathbb{P}\{(a + b\varepsilon) - (a' + b'\varepsilon') < 0\} = F_{12}(a - a'),$$

where the right hand side is greater or equal than 1/2 if and only if $a \geq a'$. In particular, this means that $-\mathbb{E}\{\log T \mid X = x\} = -f(x)$ gives an optimal risk ordering. We conclude that $-\mathbb{E}\,T|X = x$ is also an optimal risk ordering by noting that if $f(x) < f(x')$, then $\log \mathbb{E}\,T|X = x = f(x) + \log \mathbb{E}\,e^{\sigma(x)\varepsilon} \leq f(x') + \log \mathbb{E}\,e^{\sigma(x')\varepsilon} \leq \log \mathbb{E}\,T'|X = x'$. This follows from the assumption $f(x) < f(x') \implies \sigma(x) < \sigma(x')$ and that the log moment generating function of a standard Gaussian random variable is monotone and takes the closed form $\log \mathbb{E}\,e^{t\varepsilon} = \frac{1}{2}t^2$. □

1.3 Proof of Proposition 3.4

Recall the statement of Proposition 3.4.

Proposition 3.4. *Assume μ satisfies assumption C. Then, the optimal C-index takes the following form:*

$$C_\mu^\star = C_\mu(f_\mu^\star) = \mathbb{E}_{X,X'}\,\varphi(\mathbb{P}\{T > T'|X, X'\}),$$

where $\varphi(a) = \max(a, 1 - a)$ and $f_\mu^\star$ satisfies (1).

Proof. We use the following expression of the C-index: $C(f) = \mathbb{E}_{X,X'}\,\mathbb{P}\{(f(X) - f(X'))(T - T') < 0\}$. From this, we obtain that

$$
\begin{aligned}
C(f) &= \mathbb{E}_{X,X'}\,\mathbb{P}\{T > T'|X, X'\}\,\mathbb{1}(r(X, X') < 0) + \mathbb{E}_{X,X'}\,\mathbb{P}\{T < T'|X, X'\}\,\mathbb{1}(r(X, X') > 0) \\
&= \mathbb{E}_{X,X'}\,\mathbb{P}\{T > T'|X, X'\}\,\mathbb{1}(r(X, X') < 0) + \mathbb{E}_{X,X'}(1 - \mathbb{P}\{T > T'|X, X'\})\mathbb{1}(r(X, X') > 0)
\end{aligned}
$$

where $r(x, x') = f(x) - f(x')$. Under assumption C, if $f^\star$ satisfies (1), then the largest term of the right hand side of the above equation is non-zero, leading to the result. □

1.4 Proof of Proposition 3.6

We first recall the definition of the exponential family we presented in the main paper.

Definition 1.4 (Exponential family survival model). *For $\theta : \mathcal{X} \to \mathbb{R}$, $\beta : \mathbb{R}_+ \to \mathbb{R}_+$, $\tau : \mathbb{R}_+ \to \mathbb{R}$, $\eta : \mathbb{R} \to \mathbb{R}$, and $A : \mathbb{R} \to \mathbb{R}$ such that, for all $x \in \mathcal{X}$, the conditional density is a curved exponential family model*

$$\mu(t|x) = \beta(t) \exp\left[\eta \circ \theta(x)\tau(t) - A \circ \theta(x)\right], \tag{4}$$

with associated parameter $\theta(x)$. For instance, $\theta(x) = \theta^\top x$ in a generalized linear model.

Recall also the statement of Proposition 3.6

Proposition 3.6. *Under Assumption C, with θ continuous, β positive, τ non-decreasing and η continuously differentiable and non-decreasing, $\theta(x)$ is an optimal risk ordering for the C-index, thus $C_\mu(\theta) = C_\mu^\star$.*

Proof. We need to prove that, for all $x, x' \in \mathcal{X}$,

$$\eta \circ \theta(x) \geq \theta(x') \Rightarrow \mathbb{P}\{T > T'|x, x'\} \geq \frac{1}{2}.$$

We start by showing that, under the assumptions of Proposition 3.6, the scalar exponential family (conditional) distribution $\mu(t|x) = \rho(t)e^{\eta \circ \theta(x)\tau(t) - A \circ \theta(x)}$ satisfies the Monotone Likelihood Ratio (MLR) property, i.e., for all $x, x' \in \mathcal{X}$ such that $\theta(x) \geq \theta(x')$ and for all $t_1 \geq t_0$,

$$\frac{\mu(t_1|x)}{\mu(t_1|x')} \geq \frac{\mu(t_0|x)}{\mu(t_0|x')}. \tag{5}$$

Indeed, using the canonical form

$$\mu(t|x) = \beta(t) \exp\left[\eta \circ \theta(x)\tau(t) - A \circ \theta(x)\right],$$

we have that, for all $x, x' \in \mathcal{X}$ such that $\theta(x) \geq \theta(x')$ and for all $t_1 \geq t_0$

$$\frac{\mu(t_1|x)}{\mu(t_1|x')} = \frac{e^{A \circ \theta(x')}}{e^{A \circ \theta(x)}} \exp\left\{\tau(t_1)[\eta \circ \theta(x) - \eta \circ \theta(x')]\right\} \geq \frac{e^{A \circ \theta(x')}}{e^{A \circ \theta(x)}} \exp\left\{\tau(t_0)[\eta \circ \theta(x) - \eta \circ \theta(x')]\right\} = \frac{\mu(t_0|x)}{\mu(t_0|x')},$$

where we have used that $\tau(t_1) \geq \tau(t_0)$ and $\eta \circ \theta(x) \geq \eta \circ \theta(x')$ by assumption to obtain the inequality of the third line.

We now prove that inequality (5) implies $\mathbb{P}\{T > T'|x, x'\} \geq \frac{1}{2}$. Indeed,

$$\mathbb{P}\{T > T'|x, x'\} = \int_0^{+\infty}\left[\int_{t_0}^{+\infty}\mu(t_1|x)\mu(t_0|x')dt_1\right]dt_0 \geq \int_0^{+\infty}\left[\int_{t_0}^{+\infty}\mu(t_1|x')\mu(t_0|x)dt_1\right]dt_0 = \mathbb{P}\{T' > T|x, x'\}$$

$$\square$$

1.5 Proof of Proposition 3.7

We consider the following Weibull model with varying shape parameter:

$$S(t|x) = e^{-t^{f(x)}}. \tag{6}$$

Recall the statement of Proposition 3.7

Proposition 3.7. *The Weibull model* (6) *satisfies assumption C but not assumption B.*

Proof. We have the following identity:

$$\mathbb{P}\{T < T'\} = \int_0^{+\infty}\exp^{-t^{\alpha_2}} d(1 - \exp^{-x^{\alpha_1}}) = \int_0^{+\infty}\exp^{-u^{\alpha_2/\alpha_1} - u} du = I(\alpha_2/\alpha_1),$$

where we have done the change of variables $u = x^{\alpha_1}$. Now we follow the proof of Proposition 1 by Lebedev (2019) and show that $I(\beta) \geq \frac{1}{2}$ if and only if $\beta \geq 1$. The integrand is increasing in β for $0 < \beta < 1$ and decreasing for $\beta > 1$. Let $\beta \geq 2$, then

$$I(\beta) > \int_0^1\exp^{-u^\beta - u} du \geq \int_0^1\exp^{-u^2 - u} du = \sqrt{\pi}e^{1/4}\left(\Phi\left(\frac{3\sqrt{2}}{2}\right) - \left(\frac{\sqrt{2}}{2}\right)\right) \approx 0.507 > 1/2,$$

where Φ is the c.d.f of a standard Gaussian random variable. Let $1 < \beta < 2$. Then the function $u^{\beta-1}, \beta > 0$, is concave and its graph lies below the tangent at $\beta = 1$:

$$u^{\beta-1} \leq (\beta - 1)u + (2 - \beta).$$

As a result,

$$u^\beta \leq (\beta - 1)u^2 + (2 - \beta)u,$$

which can be used for estimating the integral as:

$$I(\beta) \geq \int_0^\infty\exp^{-(\beta-1)u^2 - (3-\beta)u} du = \sqrt{\frac{\pi}{\beta - 1}}\exp^{\frac{(3-\beta)^2}{4(\beta-1)}}\left(1 - \Phi\left(\frac{3 - \beta}{\sqrt{2(\beta - 1)}}\right)\right).$$

Plotting this function (see Fig. 3 of Lebedev (2019)) we can clearly see how it exceeds 1/2. Moreover, this inequality turns into equality at the limits of the interval (Steinhaus and Trybula, 1959). We have thus proved that the shape parameter $f(x)$ is an optimal risk ordering under the Weibull model (6), thus assumption C is satisfied. On the other hand, Assumption B is not satisfied as the expectation of a Weibull is not monotone on the shape parameter. Indeed, the expectation is given by $\Gamma(1 + \frac{1}{f(x)})$, where Γ denotes the Gamma function. This function has a minimum between 1.46 and 1.47, decreasing first and increasing for larger values. Thus, it gives a different ranking than the optimal risk ordering f, and consequently it does not define an optimal risk ordering. Assumption B is consequently not satisfied. $\qquad\square$

1.6   Proof of Proposition 3.8

Recall the statement of Proposition 3.8.

Proposition 3.8. *Under Assumption D, the maximizer of the C-index depends on the marginal distribution of the patients covariates $\mu(x)$.*

Proof.   As shown in Section 2.1, the maximiser of the C-index can be written as the solution of the following problem:

$$\max_f \ \mathbb{E}_{X,X'}\, \mathbb{P}\left\{T > T'|X, X'\right\} 1(f(X) < f(X')).$$

Assume for simplicity that the marginal population has support in a finite set of elements $x_1, \ldots, x_n$. Then the optimal ranking of these n points is the solution of the following combinatorial problem over the set of permutations $\mathcal{S}_n$ of size n:

$$\max_{\sigma \in \mathcal{S}_n} \ \sum_{i,j=1}^n \gamma_{ij} 1(\sigma_i < \sigma_j),$$

where $\gamma_{ij} = p_i p_j \mathbb{P}\left\{T > T'|x_i, x_j\right\}$. This corresponds to the Minimum Weighted Feedback Arc Set (MWFAS) problem with weights γ_{ij}. Thus, its solution clearly depends on the factor $p_i p_j$ under no further assumption. $\qquad\square$

2 PROOF OF RESULTS FROM SECTION 4

2.1 Proof of Theorem 4.1

We first recall below the definitions of strongly convex functions and of the Fenchel-Young conjugate.

Definition 2.1 (Strongly convex). *A twice-differentiable strongly convex function Ω with parameter β defined in a convex domain $\mathcal{C} \subset \mathbb{R}$ is a function satisfying $\Omega''(u) \geq \beta$ for all u in the domain $\mathcal{C}$.*

Definition 2.2 (Fenchel-Young conjugate (Rockafellar, 1997)). *Let Ω a function with domain $\mathcal{C}$. The Fenchel-conjugate Ω^* of Ω is a convex function defined as*

$$\Omega^*(v) = \sup_{u \in \mathcal{C}} vu - \Omega(u).$$

If Ω is differentiable, then the domain of Ω^* corresponds to the image of the gradient Ω' union the subgradients at the boundary of the domain $\mathcal{C}$. If Ω is strongly convex, then Ω^* is continuously differentiable.

Recall now the statement of Theorem 4.1

Theorem 4.1. *Let $\Omega : \mathcal{C} \to \mathbb{R}$ be a twice-differentiable strongly convex function defined in a closed convex domain $\mathcal{C} \supseteq \mathbb{R}_+$ such that $\lim_{|u| \to \infty} \Omega'(u) = +\infty$ where the limit is taken also to the negative real numbers if $\mathcal{C}$ is unbounded in this direction. Define the cost function $S(v, t) = \Omega^*(v) - vt$ where Ω^* is the Fenchel conjugate of Ω (Rockafellar, 1997). Then, the following risk*

$$\mathcal{R}(f) = \mathbb{E}_{(X,U,\Delta)} \ \frac{\Delta S(f(X), U)}{G(U)}, \tag{7}$$

is convex, smooth, and its minimizer is a monotone transformation of the conditional expectation. Thus, under assumption B it is Fisher consistent to the C-index.

Proof. Let's first consider the case without censoring where we observe the pair of random variables (X, T). We want to prove that the minimiser f^* of the risk $\mathcal{R}(f) = \mathbb{E}_{(X,T)} \ S(f(X), T)$ where $S(v, t) = \Omega^*(v) - vt$ is a monotone transformation of the conditional expectation. The loss function S has the following properties:

- The loss S is a Fenchel-Young loss function (Blondel et al., 2020) with domain $\mathbb{R}$ (i.e., $S : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$), which follows from the property $\lim_{u \to +\infty} \nabla\Omega(u) = +\infty$ (and also $\lim_{u \to -\infty} \nabla\Omega(u) = +\infty$ if $\mathcal{C}$ is unbounded to the negative values). indeed for any v, there exists u such that $v \in \partial\Omega(u)$ if $u \in \mathcal{C}$, where $\partial\Omega(u)$ stands for the subgradient of Ω at point u, which equals just the gradient at the interior of the domain $\mathcal{C}$. Thus the domain of Ω^* is the real line (see definition 2.2).

- S is smooth as Ω is strongly convex (Rockafellar, 1997) (see also definition 2.2).

- Its minimizer f^* can be written as $f^*(x) = \Omega'(\mathrm{CE}(x))$, where $\mathrm{CE}(x) = \mathbb{E}\{T|X = x\}$ stands for the conditional expectation (Blondel et al., 2020). Note that the smooth transformation $\Omega' : \mathcal{C} \supseteq \mathbb{R}_+ \to \mathbb{R}$ is monotone due to the convexity of Ω.

When censoring is present, we can use Inverse Censoring Probability Weighting (ICPW) (Robins and Finkelstein, 2000) to construct an unbiased estimate of the risk

$$\mathcal{R}(f) = \mathbb{E}_{(X,T)} \ S(f(X), T) = \mathbb{E}_{(X,T)} \ \frac{\mathbb{E}\{1(T \leq C)|X\} S(f(X), T)}{\mathbb{E}\{1(T \leq C)|X\}}$$

$$= \mathbb{E}_X \ \mathbb{E}_{T|X} \ \mathbb{E}_{C|X} \ \frac{1(T \leq C) S(f(X), T)}{G(U)}$$

$$= \mathbb{E}_{(X,U,\Delta)} \ \frac{\Delta S(f(X), T)}{G(U)},$$

where $G(u) = \mathbb{P}\{C \geq u\}$ stands for the censoring survival function. $\qquad\square$

2.2 Proof of Theorem 4.2

Recall the statement of Theorem 4.2

Theorem 4.2 (Excess risk bounds)**.** *Let $f^\star$ be an optimal risk ranking satisfying* (1)*. Let $L > 0$ a positive constant satisfying*

$$|2\mathbb{P}\left\{T > T'|x, x'\right\} - 1| \leq L|f^\star(x) - f^\star(x')|, \tag{8}$$

for all pairs x, x'. Then, the excess risk of the C-index can be bounded as

$$C(f^\star) - C(f) \leq L \mathbb{E}_X |f(X) - f^\star(X)|.$$

The result will be based on a reduction to a binary classification problem using the following Lemma 2.3.

Lemma 2.3 (Theorem 2.2 of Devroye et al. (2013))**.** *Let $Y \in \{-1, 1\}$ a binary random variable, and consider the binary classification risk $\mathcal{E}(r) = \mathbb{E}_{Z,Y} \; 1(r(Z)Y < 0)$. Then, we have that*

$$\mathcal{E}(r) - \mathcal{E}^\star = \mathbb{E}_Z \; |2\mathbb{P}\left\{Y|Z\right\} - 1|1(r(Z)r^\star(Z) < 0),$$

for any measurable function $r : \mathcal{Z} \to \mathbb{R}$.

The proof now consists in lifting the C-index ranking problem in (X, T) into a binary classification problem to apply Lemma 2.3. Let $Z = (X, X')$ and $Y = \text{sign}(T - T') \in \{-1, 1\}$ and consider the binary classification risk $\mathcal{E}(r) = \mathbb{E}_{Z,Y} \; 1(r(Z)Y < 0)$ where $r : \mathcal{X} \times \mathcal{X} \to \{-1, 1\}$. Note that the risk $\mathcal{E}$ is equal to the C-index if we set

$$r(z) = r((x, x')) = 1(f(x) \geq f(x')).$$

This can be easily seen from

$$\begin{aligned} C(f) &= \mathbb{E} \; 1(f(X) > f(X'), T > T') \\ &= \frac{1}{2} \mathbb{E} \; 1((f(X) - f(X'))(T - T') < 0) \\ &= \frac{1}{2} \mathbb{E} \; 1(r(Z)Y < 0) = \frac{1}{2}\mathcal{E}(r). \end{aligned}$$

Moreover, as the maximizer $r^\star$ of the binary classification risk is known to be $r^\star(z) = 1(2\mathbb{P}\left\{Y = 1|z\right\} \geq 1)$ and $f^\star$ satisfies the optimal risk ranking condition (1), we also have $r^\star(z) = 1(f^\star(x) \geq f^\star(x'))$. Thus, applying Lemma 2.3 we obtain that

$$C^\star - C(f) = \frac{1}{2} \mathbb{E}_{X,X'}\{|2\mathbb{P}\left\{T > T'|x, x'\right\} - 1|r_{X,X'}(f, f^\star)\},$$

where $r_{x,x'}(f, f^\star) = 1((f^\star(x) - f^\star(x'))(f(x) - f(x')) < 0)$. Using now (8) we obtain

$$\begin{aligned} \{|2\mathbb{P}\left\{T > T'|x, x'\right\} &- 1|r_{x,x'}(f, f^\star)\} \\ &\leq L|f^\star(x) - f^\star(x')|r_{x,x'}(f, f^\star)\} \\ &\leq L|f^\star(x) - f(x) - (f^\star(x') - f(x'))| \\ &\leq L(|f^\star(x) - f(x)| + |f^\star(x') - f(x')|). \end{aligned}$$

where we have used that $ab \leq 0 \implies |a| \leq |a - b|$ for $a, b \in \mathbb{R}$. Finally, taking the expectation over the i.i.d. pairs X, X' on both sides of the inequality we obtain the desired result.

2.3 Proof of Proposition 4.3

Recall the statement of Proposition 4.3.

Proposition 4.3. *Condition* (8) *is satisfied by:*

 (i) *The PH model with $L = 1$.*

 (ii) *The AFT model with L the Lipschitz constant of the cumulative distribution function of the symmetric random variable $\varepsilon - \varepsilon'$.*

 (iii) *The AFT-H model* (3) *with L is the Lipschitz constant of the cumulative distribution function of ε scaled by a factor a where $\sigma(x) \geq 1/a$ for all x.*

 (iv) *The exponential family.*

Cox model (i). In the Cox PH model, condition (8) of 4.2 is satisfied with $L = 1$. This comes from the identity

$$\mathbb{P}\{T > T'|x, x'\} = \int_0^{+\infty} h(t|x)S(t|x)S(t|x')dt$$

$$= \int_0^{+\infty} h_0(t)e^{f(x)}e^{-\int_0^t h_0(\tau)d\tau(e^{f(x)}+e^{f(x')})}dt = \frac{e^{f(x)}}{e^{f(x)} + e^{f(x')}},$$

where at the last identity we have done the change of variables $u = -\int_0^t h_0(\tau)d\tau$. The final results follows from:

$$|2\mathbb{P}\{T > T'|x, x'\} - 1| = \frac{|e^{f(x)} - e^{f(x')}|}{e^{f(x)} + e^{f(x')}} = \frac{\max(e^{f(x)}, e^{f(x')})}{e^{f(x)} + e^{f(x')}}|f(x) - f(x')| \le |f(x) - f(x')|.$$

AFT model (ii). Under the AFT model we have the following identity:

$$\mathbb{P}\{T > T'|x, x'\} = \mathbb{P}\{\log T > \log T'|x, x'\} = \mathbb{P}\{f(x) - f(x') > \varepsilon - \varepsilon'\} = F_{\varepsilon-\varepsilon'}(f(x) - f(x')),$$

where $F_{\varepsilon-\varepsilon'}$ is the cumulative distribution function of the symmetric random variable $\varepsilon - \varepsilon'$. Thus, inequality (8) is satisfied with the Lipschitz constant of the $F_{\varepsilon-\varepsilon'}$.

AFT-H model (iii). Note that the symmetric random variable $\sigma(x)\varepsilon - \sigma(x')\varepsilon'$ follows the same law as $\sqrt{(\sigma(x)^2 + \sigma(x')^2)}\varepsilon$. Using this and performing the same computations as for the AFT model, we obtain the following identity:

$$|2\mathbb{P}\{T > T'|x, x'\} - 1| = F_\varepsilon\left(\frac{f(x) - f(x')}{\sqrt{\sigma(x)^2 + \sigma(x')^2}}\right).$$

If $\sigma(x) \ge a > 0$ for all x, then

$$F_\varepsilon\left(\frac{f(x) - f(x')}{\sqrt{\sigma(x)^2 + \sigma(x')^2}}\right) \le L\left|\frac{f(x) - f(x')}{\sqrt{\sigma(x)^2 + \sigma(x')^2}}\right| \le \frac{L}{\sqrt{2}a}|f(x) - f(x')|.$$

Exponential family model (iv). Assume the conditional time-to-event density writes, for $x \in \mathcal{X}$,

$$\mu(t|x) = \beta(t)\exp\left[\eta \circ \theta(x)\tau(t) - A \circ \theta(x)\right],$$

where θ and A are continuous functions, and η and τ are continuous, non-decreasing functions. Assume further that $\mathcal{X}$ is compact (this assumption is satisfied, e.g., if all covariates are bounded). Note that, by continuity of the application θ, $\theta(\mathcal{X})$ is also compact; without loss of generality, we assume $\Theta = \theta(\mathcal{X})$. Then, the following Proposition holds.

Proposition 2.4. *There exists $L > 0$ such that, for all $(x, x') \in \mathcal{X}^2$,*

$$|2\mathbb{P}\{T > T'|x, x'\} - 1| \le L|\theta(x) - \theta(x')|.$$

Proof. Consider the function $G : \Theta \to \mathbb{R}_+$ defined by

$$G_{\theta_0}(\theta) = \int_0^{+\infty} S(t; \theta_0)\mu(t; \theta)dt,$$

where, for all $\theta \in \Theta$,

$$\mu(t; \theta) = \beta(t)\exp\left[\eta(\theta)\tau(t) - A(\theta)\right]$$

and

$$S(t; \theta_0) = \int_t^{+\infty} \beta(t_1)\exp\left[\eta(\theta_0)\tau(t) - A(\theta_0)\right]dt_1.$$

First, $t \mapsto S(t; \theta_0)\mu(t; \theta)$ is integrable w.r.t. the variable t. Second, $\theta \mapsto S(t; \theta_0)\mu(t; \theta)$ is continuously differentiable w.r.t. the variable θ. Third, since Θ is compact and η, A are continuous, $\eta_- \le \eta(\theta) \le \eta_+$ and $A_- \le A(\theta) \le A_+$ for all for all $\theta \in \Theta$; we obtain that for all $t \in \mathbb{R}_+$ and for all $\theta \in \Theta$,

$$S(t; \theta_0)\mu(t; \theta) \le S(t; \theta_0)\beta(t)\exp\left[\max\{\eta_-\tau(t), \eta_+\tau(t)\} - A_-\right].$$

Thus, by the dominated convergence theorem, G_{θ_0} is differentiable w.r.t. θ, and

$$G'_{\theta_0}(\theta) = \int_0^{+\infty} S(t;\theta_0)\mu'(t;\theta)dt.$$

We now use similar arguments to show that G'_{θ_0} is continuous. First, $t \mapsto S(t;\theta_0)\mu'(t;\theta)$ is integrable w.r.t. the variable t. Second, $\theta \mapsto S(t;\theta_0)\mu'(t;\theta)$ is continuous w.r.t. the variable θ. Third, since Θ is compact and η, A are continuous, $\eta_- \leq \eta(\theta) \leq \eta_+$ and $A_- \leq A(\theta) \leq A_+$ for all for all $\theta \in \Theta$; we obtain that for all $t \in \mathbb{R}_+$ and for all $\theta \in \Theta$,

$$S(t;\theta_0)\mu(t;\theta) \leq S(t;\theta_0)\beta(t) \left(\max\{\eta_-\tau(t),\eta_+\tau(t)\} - A_-\right) \tag{9}$$
$$\times \exp\left(\max\{\eta_-\tau(t),\eta_+\tau(t)\} - A_-\right). \tag{10}$$

By the dominated convergence theorem, G'_{θ_0} is continuous w.r.t. θ. We obtain that G_{θ_0} is continuously differentiable w.r.t. θ on a bounded domain. It is thus Lipschitz continuous, and there exists a positive constant L_{θ_0} such that, for all $\theta, \tilde{\theta} \in \Theta$,

$$|G_{\theta_0}(\theta) - G_{\theta_0}(\tilde{\theta})| \leq L_{\theta_0}|\theta - \tilde{\theta}|,$$

where $|\cdot|$ denotes the Euclidean norm on Θ.

The final step consists in proving that the family of constants $\{L_{\theta_0}\}_{\theta_0 \in \Theta}$ are uniformly bounded. To do so, it suffices to show that the derivative $G'_{\theta_0}(\theta)$ is bounded uniformly for all θ_0 and for all θ. Using (9), we obtain that

$$G'_{\theta_0}(\theta) \leq \int_0^{+\infty} \{S(t;\theta_0)\beta(t) \left(\max\{\eta_-\tau(t),\eta_+\tau(t)\} - A_-\right)$$
$$\times \exp\left(\max\{\eta_-\tau(t),\eta_+\tau(t)\} - A_- -\right)\}dt.$$

Noticing that

$$S(t;\theta_0) \leq \int_t^{+\infty} \beta(t) \exp\left[\max\{\eta_-\tau(t_1),\eta_+\tau(t_1)\} - A_-\right] dt_1.$$

completes the proof. $\qquad\square$

2.4 Proof of Theorem 4.5

Recall the statement of Theorem 4.5

Theorem 4.5. *Let $\mathcal{R}$ be the risk defined in* (7)*. Under assumption B, the following inequality holds:*

$$C^\star - C(f) \leq 2L\gamma\sqrt{\mathcal{R}(f) - \mathcal{R}^\star},$$

where $\Omega''(u) \geq 1/\gamma^2$ for all u in the domain $\mathcal{C}$.

Proof. We can safely assume there is no censoring ($\mathcal{R}(f) = \mathbb{E}_{(X,T)}S(f(X),T)$); if there is censoring we can just use the ICPW technique described in the proof of Theorem 4.1 to write the expectation in terms of (X, U, Δ). We assume that (8) is satisfied by the conditional expectation CE as $|2\mathbb{P}\{T > T'|x,x'\} - 1| \leq L|\operatorname{CE}(x) - \operatorname{CE}(x')|$, so applying 4.2 we obtain

$$C(f^\star) - C(h) \leq L\,\mathbb{E}_X\,|h(X) - \operatorname{CE}(X)|.$$

The result is based on the excess risk bound proof by Nowak-Vila et al. (2019). If S is a Fenchel-Young loss Blondel et al. (2020) then we can write

$$\mathcal{R}(f) - \mathcal{R}^\star = \mathbb{E}_X\,D_\Omega(\operatorname{CE}(X), \nabla\Omega^*(f(X))),$$

where $D_\Omega(u,v) = \Omega(u) - \Omega(v) - \nabla\Omega(v)'(u-v)$ is the *Bregman divergence* of the convex function Ω at the pair of points u, v (Blondel et al., 2020). Moreover, if Ω is $1/\gamma^2$ strongly convex (i.e., $\Omega'' \geq 1/\gamma^2$), then

$$(u - v)^2 \leq 2\gamma^2 D_\Omega(u, v).$$

Then, the final result follows from Cauchy-Schwartz:

$$
\begin{aligned}
C(f^\star) - C(f) &= C(f^\star) - C(\nabla\Omega^*(f)) \\
&=\leq L\,\mathbb{E}_X \left|\nabla\Omega^*(f(X)) - \mathrm{CE}(X)\right| \\
&\leq L\sqrt{\mathbb{E}_X \left|\nabla\Omega^*(f(X)) - \mathrm{CE}(X)\right|^2} \\
&\leq \sqrt{2}L\gamma\sqrt{\mathbb{E}_X\, D_\Omega(\mathrm{CE}(X), \nabla\Omega^*(f(X)))} = \sqrt{2}L\gamma\sqrt{\mathcal{R}(f) - \mathcal{R}^\star}.
\end{aligned}
$$

$\square$

References

Blondel, M., Martins, A. F., and Niculae, V. (2020). Learning with fenchel-young losses. *Journal of Machine Learning Research*, 21(35):1–69.

Devroye, L., Györfi, L., and Lugosi, G. (2013). *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media.

Lebedev, A. V. (2019). The nontransitivity problem for three continuous random variables. *Automation and Remote Control*, 80(6):1058–1068.

Levy, H. and Robinson, M. (2006). *Stochastic dominance: Investment decision making under uncertainty*, volume 34. Springer.

Nowak-Vila, A., Bach, F., and Rudi, A. (2019). A general theory for structured prediction with smooth convex surrogates. *arXiv preprint arXiv:1902.01958*.

Robins, J. and Finkelstein, D. (2000). Correcting for noncompliance and dependent censoring in an aids clinical trial with inverse probability of censoring weighted (ipcw) log-rank tests. *Biometrics*, 56(3):779—788.

Rockafellar, R. T. (1997). Convex analysis. princeton landmarks in mathematics.

Steinhaus, H. and Trybula, S. (1959). On a paradox in applied probabilities. *Bull. Acad. Polon. Sci*, 7(67-69):108.